



A Ranking Approach to Summarising Twitter Home Timelines

By:

Dominic Paul Rout

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Engineering
Department of Computer Science

30th June, 2015

Abstract

The rise of social media services has changed the ways in which users can communicate and consume content online. Whilst online social networks allow for fast and convenient delivery of knowledge, users are prone to information overload when too much information is presented for them to read and process.

Automatic text summarisation is a tool to help mitigate information overload. In automatic text summarisation, short summaries are generated algorithmically from extended text, such as news articles or scientific papers. This thesis addresses the challenges in applying text summarisation to the Twitter social network. It also goes beyond text, exploiting additional information that is unique to social networks to create summaries which are personal to an intended reader.

Unlike previous work in tweet summarisation, the experiments here address the home timelines of readers, which contain the incoming posts from authors to whom they have explicitly subscribed.

A novel contribution is made in this work the form of a large gold standard (19,350 tweets), the majority of which will be shared with the research community. The gold standard is a collection of timelines that have been subjectively annotated by the readers to whom they belong, allowing fair evaluation of summaries which are not limited to tweets of general interest, but which are specific to the reader.

Where the home timeline is used by professional users for social media analysis, automatic text summarisation can be applied to give results which beat all baselines. In the general case, where no limitation is placed on the types of readers, personalisation features which exploit the relationship between author and reader and the reader's own previous posts, were shown to outperform both automatic text summarisation and all baselines.

Acknowledgements

I wish to thank my supervisors Dr Kalina Bontcheva and Dr Mark Hepple, who have provided invaluable guidance and motivation throughout the course of my PhD. I have benefited endlessly from their enormous knowledge and experience.

I am grateful for the advice, annotations and opportunities for collaboration offered by all of the members of the Natural Language Processing Research Group, including Daniel Preoțiuc-Pietro, Leon Derczynski, Diana Maynard, Genevieve Gorrill, Ian Roberts, Trevor Cohn, Isabelle Augenstein and Roland Roller. A great deal of help was offered in the form of moral support and study participation by Chris Wright, Alasdair Armstrong and Mathew Hall.

This work would not have been possible without the funding of English Physical Sciences Research Council and the European Commission, nor without the hundreds of tireless volunteer staff and students from The University of Sheffield.

The personal emotional support provided by Belinda Alexander and my mother Helen Richardson have helped me through the difficult but worthwhile journey of completing this thesis.

Contents

Abstract	ii
Acknowledgements	iii
1 Introduction	1
1.1 Ranking Home Timelines	3
1.2 Aims and Objectives	3
1.2.1 Evaluating Home Timeline Ranking	4
1.2.2 Information Retrieval and Summarisation Techniques	4
1.2.3 Social Media Monitoring	4
1.2.4 Personalised Summaries of Twitter Timelines	5
1.3 Thesis Structure	5
1.4 Previously Published Material	6
2 Background	8
2.1 Twitter	8
2.1.1 Status updates	9
2.1.2 Communication	10
2.1.3 Types of Tweet	11
2.1.4 Followers and Friends	13
2.1.5 User profiles	15
2.1.6 Discussion	15
2.2 Text Summarisation	16
2.2.1 Automatic Textual Summarisation	17
2.2.2 Frequency Based Summarisation	18
2.2.3 Context and Term Position	20
2.2.4 Machine Learning for Summarisation	20

2.2.5	Generating Summaries	22
2.2.6	Sentence Modification	22
2.2.7	Evaluation	23
2.3	Twitter Summarisation	24
2.3.1	Phrase Generation	24
2.3.2	URL Recommendation	25
2.3.3	Extractive Summarisation of Tweets	25
2.3.4	Tweet Recommendation	27
2.3.5	Search Ranking	28
2.3.6	Home Timelines	30
2.3.7	Evaluation	31
2.4	Social Media Summarisation	32
2.5	Discussion	34
3	Task Definition and Gold Standard Creation	37
3.1	Introduction	37
3.2	Task Definition	38
3.2.1	Retrieval Process	38
3.2.2	Tweet Filtering	39
3.2.3	Creating Summaries	41
3.3	Data Collection	41
3.3.1	Pilot Data	42
3.3.2	Redesigned Personal Timelines Data Set	46
3.3.3	Political Timelines	53
3.3.4	Validation	56
3.4	Reasons for Relevance	62
3.5	Retweet Limitations	63
3.6	Discussion	67
4	Methods and Representations	69
4.1	Introduction	69
4.2	Approach	70
4.3	Baseline Methods	71
4.3.1	Random	71
4.3.2	Tweet Meta Data Baselines	72
4.3.3	Text Baselines	72

4.4	Comparison to Related Work	72
4.5	Text Summarisation	73
4.5.1	Centroid	73
4.5.2	TextRank	74
4.5.3	Sum Term Frequency (TF)	75
4.5.4	BM25	76
4.5.5	Maximal Marginal Relevance	77
4.6	Representations	78
4.6.1	Stoplist	79
4.6.2	Inverse Document Frequency	79
4.6.3	Dimensionality Reduction	80
4.6.4	Overall Preprocessing	82
4.7	Evaluation Metrics	83
4.7.1	Mean Average Precision	84
4.7.2	ROUGE	85
4.7.3	Significance	86
4.8	Discussion	86
5	Ranking for Social Media Analysis	88
5.1	Introduction	88
5.2	Methodology	90
5.2.1	German Language Stoplist	90
5.2.2	German Language IDF	90
5.3	Baselines	91
5.3.1	Results	91
5.4	Ranking with Text	92
5.4.1	Results	92
5.4.2	Dimensionality Reduction	96
5.5	User Evaluation	99
5.5.1	Study Design	100
5.5.2	Evaluated approaches	102
5.5.3	Annotators	103
5.5.4	Agreement	103
5.5.5	Results	104
5.5.6	Revisiting automatic evaluation	105
5.6	Discussion	106

6	Ranking for General Twitter Users	108
6.1	Introduction	108
6.2	Methodology	109
6.2.1	English Language Stoplist	109
6.2.2	English Language IDF	109
6.2.3	Baselines	111
6.3	Text-based Ranking	112
6.3.1	Centroid	112
6.3.2	TextRank	113
6.3.3	Sum Term Frequency (TF)	115
6.3.4	BM25	116
6.3.5	Maximal Marginal Relevance	116
6.3.6	Dimensionality Reduction	118
6.3.7	Discussion	120
6.4	Ranking with Social Information	121
6.4.1	Features	122
6.4.2	Results	127
6.4.3	Discussion	129
6.5	Machine Learning	129
6.5.1	Machine Learning Methods	130
6.5.2	Learning objective	133
6.5.3	Features	134
6.5.4	Use of Features in Related Work	137
6.5.5	Feature Selection	137
6.5.6	Results	139
6.5.7	Discussion	142
6.6	Personalised Machine Learning Experiments	144
6.6.1	Data	145
6.6.2	Results	145
6.6.3	Discussion	147
6.7	Discussion	149
6.7.1	Validation on Unseen Data	149
6.7.2	Discussion	151

7 Conclusion	154
7.1 Contributions	156
7.1.1 Effective Evaluation	156
7.1.2 Social Media Monitoring	156
7.1.3 Personalised Ranking	157
7.2 Future Work	157
7.2.1 User Profiles	157
7.2.2 Machine Learning Models	158
7.2.3 Tweet Classification	159
7.2.4 Extractive Summaries	159

Chapter 1

Introduction

Online social networks have led to new ways in which individuals can consume information, interact socially and maintain friendships. Unlike in published media, the readers in such networks are also content producers, and are not generally professionals; social networks are a common type of user generated content (Moens et al., 2014). Users can connect, generally as way of subscribing to see one another's posts. Microblogging services, such as Twitter¹, are a specific kind of social network in which posts are limited in length. In Twitter, as in many other services, posts are organised chronologically rather than by author or subject.

The volume of content on these services is enormous; readers can choose the accounts to which they subscribe but are still faced with more information than they can read (Douglis, 2010). These users are said to suffer from information overload (Beaudoin, 2008). Hargittai et al. (2012) argue that overload occurs because of a lack of structure in information, because messages are time sensitive, or because there is a low signal-to-noise ratio in the information stream.

The problem is not addressed in social networks such as Twitter wherein content is organised by time rather than by theme or topic. In these networks, readers are not given an effective way to home in on the tweets that are the most personally interesting to them. There is evidence of information overload in the Twitter network, with a clear threshold in tweet volume at which users stop interacting effectively with content (Gomez-Rodriguez et al., 2014). Twitter authors post many different kinds of information (Naaman et al., 2010), and are not bound to post a single topic.

To address information overload, posts can be prioritised according to predictions

¹<http://twitter.com>

of their value to the reader. Services such as Facebook already highlight content that is otherwise liked by a large number of users, and prioritise paid advertisements. In Twitter, popular posts are served to users via daily or weekly emails.

For long form text, such as news articles or journal publications, automatic text summarisation is used to generate summaries which paraphrase the content or which help readers decide which documents merit further exploration (Mani, 2001). Information overload might be mitigated if this type of summarisation can also be applied to social media. Likewise, techniques from information retrieval which prioritise content according to relevance to a query (Manning et al., 2008) may also help users engage only with posts of interest.

Social media posts, and especially microblogs, differ from traditional text so much as to provide challenges for text summarisation and information retrieval methods. Posts are extremely short, limited to just 140 posts on the Twitter service, providing both limited context for textual comparison and a lack of structure to exploit. They exist as part of a social graph, with users connected to readers, and many posts are purely social in function (Naaman et al., 2010).

The focus of this work is in understanding how microblogs may be summarised, including the differences between summaries of extended text and summaries of microblogs. The ways in which these differences impact the utility of existing methods are discovered, and they are exploited in new ways to give stronger performance in a social media ranking task.

The specific microblogging platform under consideration is Twitter. This is an example of a popular microblogging platform on which users share short posts, through which they post status updates, share links or interact with one another. A description of Twitter itself will be presented in Section 2.1. Other services could have been considered, including Sina Weibo, a Chinese microblogging service, or Instagram, a photo sharing website in which users can follow one another without reciprocating. Twitter was chosen for the following reasons.

- It is very widely used.
- Twitter posts are very short, meaning they take little time to consume and are immediate.
- Twitter’s default view is organised by time, not by topic, meaning that tweets are transient.

- The service is highly connected, and one user can connect to another without a reciprocated relationship (Easley and Kleinberg, 2010).
- It is used in diverse ways (Naaman et al., 2010).

1.1 Ranking Home Timelines

The home timeline is the default collection of posts seen by ordinary users when they log into Twitter. The posts are ordered chronologically, and are drawn from authors to whom the user has subscribed.

This work experiments with the ranking, or prioritisation of social network posts from the home timeline according to their predicted interestingness to these ordinary Twitter users. Because tweets are drawn from the home timeline, containing the reader's own subscriptions, the ranking generated is personal in nature, and as with other collections of tweets, the ideal ranking is subjective (Alonso et al., 2013).

The evaluation in this work is carried out using a gold standard of tweets drawn from the timelines of actual users, filtered by topics of interest. There has been some previous work which summarised tweets from the home timeline (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013), which is not focussed around any particular topic, and other work which summarised tweets from search results (Ounis et al., 2011; Choi et al., 2012; Duan et al., 2010; Huang et al., 2011; Becker et al., 2011; Inouye and Kalita, 2011), which are not personal to the reader. In both types of tweets, different kinds of summarisation will be favoured - since personalisation is more important for home timelines, and topical relevance is more important for search results. As such, personalisation and topical relevance algorithms cannot be effectively compared, or combined. To our knowledge, this is the first work which allows fair comparison of the two, as reflected in the features and models used throughout this thesis.

1.2 Aims and Objectives

This work explores the task of Twitter home timeline summarisation through the following goals.

1.2.1 Evaluating Home Timeline Ranking

The summarisation of Twitter home timelines is viewed as a subjective ranking task. This work will develop a framework in which timeline rankings can be evaluated automatically. Given the personal nature of the home timeline (every user curates their own timeline), and the subjectivity of the task (Alonso et al., 2013), relevance judgements should be made by the target user.

Existing work in this domain pulls relevance judgements from the retweets of Twitter users (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013), assuming that posts which are retweeted and only posts which are retweeted are of relevance to the reader. Section 3.5 motivates the need for a user-reported ground truth for tweet relevance, given the nature of retweets.

1.2.2 Information Retrieval and Summarisation Techniques

A variety of techniques from both text summarisation and information retrieval will be evaluated using a new gold standard, as will be defined in Chapter 3. Tweets vary from extended prose not just in terms of message length, but in quality, purpose and their social nature. As such, a variety of existing techniques will be applied to tweets, to understand their performance in the new domain.

The methods selected for evaluation here are based on a combination of their popularity within informationretrieval and summarisation, and the technical aspect of whether it was feasible to apply them to tweets using current technology. For example, approaches which rely heavily on complex links between sentences, based on features other than term counts, such as coreference, are not considered because of the lack of tools for those tasks developed for tweets. Investigation of the following hypothesis will discover to what extent the chosen existing techniques can apply to this new problem domain, and in what role:

Hypothesis 1 - One or more of a selected subset of methods from information retrieval and summarisation can give effective ranking in the Twitter domain.

1.2.3 Social Media Monitoring

The reasons for which users in general find tweets interesting are discovered through an empirical process (Section 3.4), giving a broad, nonspecific use case for summarisation. However, a more specific case study is carried out wherein summaries are produced for the purpose of social media monitoring.

In this small-scale use case, the organisation surveyed uses social media exploration for the purpose of driving deeper qualitative and quantitative analysis surrounding Austrian political events and analysis. This work will develop and evaluate systems which are useful in this sort of social media exploration, and to understand this use case in depth.

1.2.4 Personalised Summaries of Twitter Timelines

Given that the home timeline is personal to the user that curated it, both the reader and author of tweets within that timeline are known. This is different to many summarisation settings, in which text is summarised for an unknown user, or in response to a specific query. This thesis will discover how the knowledge of the reader impacts the summarisation task.

This knowledge of the reader allows for a degree of personalisation of the rankings (Kapanipathi et al., 2011; Ren et al., 2013; Feng and Wang, 2013). Experiments will be carried out to determine the extent to which and ways in which summaries of Twitter home timelines should exploit knowledge of the relationship between author and reader and the reader's previous posts. These experiments will validate the following hypothesis:

Hypothesis 2 - Methods which rely on personalisation using information about the readers, authors and the network between them can be shown to give stronger performance than a number of textual baselines.

The textual baselines referred to in this hypothesis are the same information retrieval and summarisation methods as will be used to evaluate Hypothesis 1.

1.3 Thesis Structure

The remaining chapters of this thesis will be arranged as follows.

Chapter 2: Background An introduction to the Twitter platform is presented, including the features that make it unique amongst social network services. A brief taxonomy of text summarisation methods is given, along with an overview of how classical text summarisation differs from the summarisation of tweets. Multiple variants for the home timeline summarisation task are discussed, along with an in-depth discussion of previous related work.

Chapter 3: Task Definition and Gold Standard Creation This chapter specifies in detail the exact ranking task to be addressed in this thesis, and describes the process of creating a series of large gold standards to evaluate tweet rankers. The limitations of retweets as a ground truth for this work are shown empirically.

Chapter 4: Methods and Representations Defines the approach taken to summarising Twitter home timelines and discusses specific evaluation criteria which will be used to compare these summaries. This chapter also presents a number of existing ranking and summarisation algorithms and text representations, which rank on the sole basis of relevance to the topics contained in the tweets themselves. Several of these methods must be adapted or reinterpreted for Twitter timelines, and these modifications are also given here.

Chapter 5: Ranking for Social Media Analysis Beginning with a specific use case of social analysis of tweets surrounding political speeches and events, it is shown that automatic summarisation can assist the social media monitoring work of real users in a political analysis organisation.

Chapter 6: Ranking for General Twitter Users While the previous chapter discusses home timeline ranking for the benefit of a social analysis organisation, this chapter expands beyond this single use case to attempt to provide useful summaries for general users of Twitter. Personalisation and social features are used to supplement text based and tweet popularity based ranking. Machine learning models are trained to combine multiple kinds of features.

Chapter 7: Conclusion The key results from this thesis are revisited, and a summary is given of its main contributions in evaluation, social media monitoring and personalised tweet ranking. The conclusion looks towards future work, including ways to further enhance the work in this thesis and to broaden the nature of the problem area.

1.4 Previously Published Material

Portions of this thesis have been published in conference proceedings. The pilot study described in Chapter 3, and the limitations of that study were published in

the 2013 proceedings of the AAAI Symposium on Analyzing Microtext (Rout et al., 2013a).

The social media monitoring use case presented in Chapter 5, along with the gold standard data set used to enable the work, has been accepted to appear at HyperText in 2015 (Rout and Bontcheva, 2015).

A third publication, based on the results of Chapter 6, is under review at the time of writing.

The author has produced a series of other publications in related areas of social media analysis, including a review chapter on the summarisation of user generated content, and has contributed to the book “Mining User Generated Content” (Moens et al., 2014). The author’s work on the geolocation of Twitter users through homophily and the use of an automatic classification framework won the Ted Nelsen best newcomer award at HyperText in 2013 (Rout et al., 2013b).

Chapter 2

Background

This chapter presents a brief overview of Twitter as a service, the task of automatic text summarisation, and a discussion of existing work in the domain of Twitter home timeline summarisation and ranking.

Section 2.1 introduces Twitter, including the way the home timeline is created, the nature of posts to the service and the social connections formed between users. Twitter is used as a prominent example of a microblogging service, but the core ideas and experiments in the work are transferable to other networks provided they share certain traits, which will be enumerated in Section 2.1.6.

Automated text summarisation, discussed in Section 2.2 produces shorter summaries of long documents automatically. In this section, a broad taxonomy of text summarisation is given, including discussion of how ideas from text summarisation are not always easily transferable to social networks like Twitter.

Lastly, a review of the different kinds of summarisation approaches that have been applied to tweets will be presented in Section 2.3. The commonalities and differences between the summarisation of tweets and the summarisation of long-form text will be explored, and existing efforts within the domain of tweet summarisation will be compared with the work in this thesis.

2.1 Twitter

The social network Twitter is studied in this work as a widely used microblogging service. Twitter is multi-lingual and international (Hong et al., 2011), in the first quarter of 2015, there were on average 302 million users active on the service each

month¹.

In Twitter, there is no difference between celebrities and other users in terms of the type of account that they may hold, however celebrity users garner far more attention than others, and have different following behaviour to other, non-elite users (Wu et al., 2011). It is also unique because of the nature of the short messages, and the conventions that develop from there, though these conventions do not appear to be fixed and can vary between languages (Hong et al., 2011). Twitter warrants study as a relatively new and innovative medium for publishing and communication, though many of the properties of Twitter that are central to this work may be found in other social networks, as discussed in Section 2.1.6.

2.1.1 Status updates

The basic unit of textual content in Twitter is a tweet. Tweets must be at most 140 characters long, a limitation which was first imposed to match the 160 characters available in an SMS message, less 20 characters to encode a username (Boyd et al., 2010). As with SMS, several sequential tweets can be used to create longer messages, with the three characters “...” often (but not always) used to signal a run-on between messages.

Tweets are succinct; Go et al. (2009) found that the average length of a Tweet is just 14 words. Whilst very limited repetition or redundancy is possible in the 14 words of a single tweet, collections of tweets around a subject often repeat one another (Sharifi et al., 2010). Twitter is not a curated medium (like newswire or academic journals), but rather an informal one with space restrictions, so authors make use of abbreviations and write ungrammatically (Bontcheva et al., 2013a). Many such abbreviations are common, such as ‘LRT’ meaning ‘in reference to last retweet’, ‘tho’ as shorthand for ‘though’. Other unusual tweet conventions are listed and explained in Table 2.1, many of which are ungrammatical.

This unusual language makes tweets difficult to parse, leading to degraded performance of linguistic pipelines (Ritter et al., 2011). Syntactic features such as hashtags and mentions require the modification of tokenisers to be handled properly. Misspellings and word variants, which cause problems when terms are being matched exactly, have been dealt with in part of speech tagging through normalisation, substituting canonical forms of misspelled tweets (Derczynski et al., 2013), and by clustering term variants (Ritter et al., 2011).

¹<http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

Post	Explanation
Monster munch? Yum! RT @UserRemoved: roast beef monster munch are vile	Retweet with quotation
@UserRemoved Whatever. I'm gonna go play Portal :-)	@Mention of a user
On a lighter note... here is a phoenix http://fb.me/(removed)	URL
Never forget. #Paste	Hashtag for a topic
Starving... *eats*	Signifies an *Action*
This. So much this.	Enthusiastic agreement
Tweets can sometimes be much too short and I think (1/2)	Part one of 2
... that they should be allowed to be longer ...	Continuation before and after
That's not even what I mean tho?	Sarcastic rhetorical
Dis bed feelin 2 gud 2 get out of rite now	Phonetic spelling
Listening to Craig David.... Life is goood!	Ellipsis & Word lengthening
women! on bbc2 is v good	Abbreviation

Table 2.1: Example microblog posts from Twitter

Twitter is prone to abuse by spammers, who post automatically for promotion or to degrade the service, often repeating the same text many times. Lee et al. (2011) attempt to classify spammers while Sharifi et al. (2010) use Bayesian filtering to identify spam tweets themselves.

2.1.2 Communication

Users on Twitter can interact in a number of different ways, some of which are private and take a form closer to instant messaging, and others which are entirely public and take place in a way that is visible to all.

Direct Messages These are hidden short messages that can only be seen by the author and the reader. They are functionally similar to posts on online instant messaging services, though they are limited to 140 characters as with public tweets.

Mentions Users may also choose to ‘mention’ one another by prepending their username with ‘@’ within a tweet. Threads of mentions are automatically shown to anyone who has subscribed to the author and participants can drop in and out of a thread of discussion at any time, often leading to very long and complex interactions

(Honeycutt and Herring, 2009). While mentions do not appear by default in other timelines, they are public and can be found by searching or viewing the author’s profile.

Replies A specific type of mention, in which the username of the intended user is placed in the first position. By default these mentions are displayed to users who are subscribed to both the author and recipient, as well as the recipient themselves. As with mentions, these can be seen by other users by searching for the user or browsing their past tweets.

Retweets A user copies another’s status update, rebroadcasting it. A retweet may be carried out either through the client itself, in which case it appears intact, with the name of the original author, or by copying and pasting its contents and adding the token ‘RT’ followed by the original author’s name. Authors receive notifications when they are retweeted.

Retweeting is used for information sharing. Commonly, retweets contain links to interesting content. Such links help relate microblogs to traditional news and blog services. Raux et al. (2011) even used Tweets containing URLs as a feature in the task of URL grouping, as a substitute for the content of the pages themselves.

Favourites A reader endorses a tweet. The number of times a tweet has been favourited is counted and is reported through the Twitter user interface. Authors receive notifications when their tweets are favourited.

Hashtags Tweets can be made globally visible and open to anyone in the whole of Twitter by marking them with hashtags. These typographical markers, created by placing a hash (#) before a word, are usually semi-unique identifiers and they can be used to enter a tweet into a wider discussion (Laniado and Mika, 2010). Users can search for hash tags and see all messages which contain them, discovering what the whole of Twitter is saying on the subject.

2.1.3 Types of Tweet

Table 2.2 shows some types of tweets which are classified following the schema of Naaman et al. (2010), who show that users could be divided into classes called ‘meformers’ and ‘informers’ according to how often they posted each sort of mes-

Code	Meaning	Example
IS	Information Sharing	Student fights US extradition bid (URL removed)
SP	Self Promotion	New issue out around the Union shortly!
OC	Opinions/Complaints	Listening to Martin Simpson makes me happy.
RT	Statements and Random Thoughts	If I married Mr Boe my name would be Harry Boe. Haribo.
ME	Me now	Just got home after 15 hours
QF	Questions to followers	*sighs* Why is my PC freezing again?
PM	Presence Maintenance	Am going to try and get some sleep. Later, twitters!
AM	Anecdote (me)	Last night I learned that I sheep walk when I sleep..sorry about da chewed up toilet paper rolls...
AO	Anecdote (others)	Heard (user removed) got engaged! Congratulations!

Table 2.2: Categories of Twitter message intentions as described by (Naaman et al., 2010)

sage. When adopting these classes, they found that 80% of users were meformers, indicating that they tweet primarily about themselves and their own experiences.

While Naaman et al. (2010) went some way towards categorising tweets for intention, their method was not automatic and it remains unclear whether or not similar categories could be assigned by an automatic classifier.

Tweets can be classified according to apparent intent (the user is simply expressing an anecdote), or according to the value that a reader may gain from it (this is informative, this is a recommendation etc), the latter type of classification will be discussed further in Section 3.4.

2.1.4 Followers and Friends

Users of Twitter can “follow” one another to see tweets. Following a user indicates an interest in their tweets, which will subsequently appear in a view of Twitter known as the personal or home timeline as they are posted. The home timeline contains messages from everyone a user follows. Unlike in many other services, the act of following in Twitter need not be reciprocated; that is, one user can follow another without being followed by them. In fact, the level of reciprocity in Twitter is lower than other social networks, indicating that most people follow without being followed back (Kwak et al., 2010).

A friend is the inverse of a follower, otherwise known as a followee. This is an account that the reader follows.

When users mutually follow one another, they will be notified of public tweets directed at them, and will be able to send each other direct messages. One view of Twitter is as a large graph in which these mutual follower/following relationships comprise the edges (Kwak et al., 2010).

The Home Timeline

The home timeline is a reverse chronological list of posts by the authors that a user has subscribed to. Users invest time curating their home timeline by choosing twitterers to follow based on their extant friendships, interests or location (Sousa et al., 2010; Kwak et al., 2010). They sometimes unfollow others who post too much or who post on topics they are not interested in, amongst other reasons (Kwak et al., 2011). Twitter’s makes it very easy to end a subscription to another user.

Some tweets are more interesting than others. As a part of this work, discussed in Section 3.3.1, a sample of around 70 users were asked to indicate through a

web interface which of their recent incoming tweets they considered interesting. Some indicated that as many as 60% or beyond were worth reading, but most selected fewer, with the mean number of interesting tweets at 16.34%. A low ratio of interesting to uninteresting tweets demonstrates a poor signal to noise ratio, where uninteresting tweets are considered noise. Nobody stated that every tweet in their home timeline was interesting.

Connectivity

Twitter users do not exist in isolation. As Java et al. (2007) showed, they form social connections and build communities, discussing shared topics of interest with one another and engaging in socialising, information seeking and information sharing.

For many users it is their interests which govern who they follow. A fan of television comedy is more likely to follow a comedic writer, for example. In many cases, however, users interact with one another on the basis of real world or virtual social ties, regardless of shared interests (Sousa et al., 2010). The notion that users in classical social networks are likely to be connected to others with whom they share common traits is known as homophily (Easley and Kleinberg, 2010).

It is useful to regard the egocentric network of a user (that is, a subset of the overall social network representing their immediate and close connections) in light of the apparent homophily within Twitter. Java et al. (2007) find that connected groups tend to share a common topic of interest and also that users were more likely to follow others on the same continent as their own. The latter assertion was investigated by Kwak et al. (2010) who found that an overwhelming majority of connections were between users in the same time-zone as one another. Users who have more friends in common are also more likely to live in the same city (Rout et al., 2013b).

Topic homophily was also explored by Sousa et al. (2010). They find that those with large networks tended to compartmentalise their social connections by topic. This is consistent with the earlier findings that cliques are formed according to shared interests (Java et al., 2007).

Early research found a graph made of connections between Twitter users to contain much higher reciprocity and degree correlation than a corpus of weblogs (Java et al., 2007). Krishnamurthy et al. (2008) stated that the number of followers as well as the number followed by a user grew with the number of posts a user had made.

In comparison, Kwak et al. (2010) later found that 77.6% of user connections were not reciprocated. Additionally, they found that most users were not followed by anyone they themselves followed, meaning they did not use the service in any social way at all. They offer the explanation that the majority use Twitter as a source of information alone. Their findings somewhat contradict those of Java et al. (2007), who stated that the graph was highly reciprocal, though Twitter grew rapidly in the three year period between the two studies, suggesting that its use may have simply changed over time.

2.1.5 User profiles

Twitter users each have a profile which reveals details of their identity. The profile is semi-structured, able to store a textual biography field, a full name, the user's location, a profile picture, a time zone and a homepage URL. The user's attributes affect the content of their posts, for example their physical location can help determine the language they use (Fink et al., 2008; Cheng et al., 2010) or the events on which they comment (Yardi and Boyd, 2010).

There have been efforts to discover information about users which is not available in the fields in their profile. Burger et al. (2011) classify users as male or female based on the text of their tweets, their description fields and their names. They report better-than-human accuracy, compared to a set of annotators on Mechanical Turk. Pennacchiotti and Popescu (2011) present a general framework for user classification which can learn to automatically discover political alignment, ethnicity and fans of a particular business, with the potential to be retrained for other classification tasks.

2.1.6 Discussion

This section outlined a number of technical and social aspects of the Twitter platform. Twitter is a fast moving stream of information with an unusual method of structuring posts. There are, however, other networks that have adopted the same or similar conventions. The following aspects of Twitter are considered key to this work:

Readers are Authors: Most Twitter readers also post their own messages, and celebrity authors are not privileged by the service over other users, though they can be expected to receive more attention (Wu et al., 2011).

Connections are explicit: Unlike blogging services, for example, the list of users that a reader follows is explicit, and can be seen by third parties. Moreover, connections between users are not always reciprocated, and might be formed for a variety of different reasons.

Posts are short and timely: Posts are limited to 140 characters by convention (though multiple tweets are sometimes issued in a row). They are ordered chronologically, and every post is shown. Users are not necessarily expected to read every tweet.

Interactions between users are often public: Unless conversations are of a private nature, personal discussions between users happen publicly and can be seen by the friends they share. Users may also join in with one another's discussions.

The networks Weibo², and Ello³ also possess most of these features, and much of the evaluation applied in this work could generalise to those other networks. Facebook⁴ has less in common with Twitter; authors are also readers, and connections are explicit, but since the restructuring of the UI users primarily communicate via private messaging, to the degree that many had forgotten that they had previously had public conversations, evidenced by a privacy scare in 2012 when these messages became easier to find⁵. Facebook is also far less timely than Twitter, ranking posts by default through a recommender system rather than ordering them chronologically.

2.2 Text Summarisation

Textual summaries are extremely common in online and printed media. Table 2.3 lists some examples of types of summaries which are often encountered. These classes have been described as either *critical*, *informative* or *indicative*, where critical summaries appraise and evaluate a work in some way, informative summaries capture the content of the original document, and indicative summaries enable a reader to decide whether or not to read the whole document (Nenkova and McKeown, 2011).

Textual summaries are traditionally also described as *extractive* or *abstractive* (Carenini and Cheung, 2008). The former type of summary, the extract, is composed entirely from text which can be found in the original document. The latter form, the

²<http://weibo.com>

³<http://ello.co>

⁴<http://facebook.com>

⁵<http://pando.com/2012/09/24/how-facebook-got-pummeled-over-a-fake-privacy-scandal/>

Summary	Purpose
Journal abstract	Indicative
News report	Informative
Movie review	Critical
Novel blurb	Indicative
Football highlights	Informative

Table 2.3: Example classes of summary

abstract, contains at least some text which was not present in the source. Extractive summarisation is the focus of the work in this thesis, because whilst abstractive summarisation can require detailed information extraction and text generation, much of which is not well developed for microblogs, extractive summarisation merely requires that existing text units can be scored and combined automatically (Nenkova and McKeown, 2011).

Summaries can be classified by whether they are derived from single or from multiple documents (Radev et al., 2000). Multi-document summarisation must address slightly different challenges to single document summarisation such as redundancy, which can be caused by identical or semantically similar text appearing in several documents documents, the order of publishing, and inter-document references (Goldstein et al., 2000).

Summaries may also be *topic-centric (generic)*, *user-focused* or *query-focused*. The former class of summary is meant simply to summarise the content with no bias as to whom can benefit from it. Query focussed summarisation, on the other hand, involves building a summary to meet a specified information need; a task which is key to information retrieval (Manning et al., 2008). User focussed summarisation models the information needs and interests of a specific user, arranging a summary containing details which they alone may find salient.

2.2.1 Automatic Textual Summarisation

While many summaries in every day use are created manually by humans, it is desirable to generate summaries automatically. Automatic textual summarisation aims to produce shorter, human readable representations of longer text. Extractive summaries can be produced in two phases (Nenkova and McKeown, 2011):

1. Score textual units (sentences, phrases, paragraphs etc) according to some representation of the document or document set.

2. Generate summaries by selecting high scoring textual units until the desired compression ratio has been achieved.

The textual unit to be included in a summary could be a word, phrase, sentence or whole paragraph depending on the application, though in the majority of existing work, the summary is composed of sentences.

2.2.2 Frequency Based Summarisation

The central idea behind frequency based automatic summarisation is that terms which are more salient in a document are often repeated; that is, a document which discusses a presidential election may reuse terms like ‘party’ several times (Brandow et al., 1995). Since textual units containing these terms might well be more important than others, they are favoured for inclusion in the summary.

The simplest term-frequency method considers the words that represent the topic of the document to be those that occur more frequently in the document. This is known as term-frequency (TF) weighting. The term frequency score for an individual term w is given for document D as follows (Manning et al., 2008):

$$\text{tf}(w) = |\{w' \in D : w' = w\}| \quad (2.1)$$

With this definition of term frequency, individual sentences from the document are scored for relevance according to the sum of the term frequencies for the terms that they contain. Given a sentence S , the relevance score is as follows.

$$\text{TF}(S) = \sum_{w \in S} \text{tf}(w) \quad (2.2)$$

The term frequency method suffers from the problem that many frequent terms are not useful in capturing the nature of a document (Nenkova and McKeown, 2011). These include terms which appear in any document, regardless of topic or domain, because they are members of the closed classes of words that give structure to language (determiners ‘an’ or ‘that’, prepositions ‘in’ or ‘over’, auxiliary verbs ‘is’ and ‘has’ etc). Terms which would appear in any document for the domain, (‘model’ or ‘algorithm’ in computer science) are also not useful for summarisation.

These terms can be eliminated by setting a threshold on their likelihood of occurrence, or the number of documents in which they occur, to form a stop list. However, they may also be negotiated by weighting the term frequency score against the number of documents from a collection in which they occur. This notion was described

in very early work by Jones (1972) as term specificity, in a paper which also demonstrated term specificity’s importance in information retrieval. A formulation of this was later described as inverse document frequency, or IDF, by Buckley and Salton (2009), who gave the formula for this value, given a collection of documents D , as:

$$\text{idf}(w, D) = \log \frac{|D|}{|\{d \in D : w \in d\}|} \quad (2.3)$$

Term frequency weighting can then be combined with inverse document frequency to give a TF.IDF score for the sentence.

$$\text{TF.IDF}(S, D) = \sum_{w \in S} \text{tf}(w) \times \text{idf}(w, D) \quad (2.4)$$

There are other similar approaches to scoring sentences for summary generation including weighting them by probability of appearing in the document as opposed to the entire collection, or counting the number of topical keywords that appear in the sentence, determining these words probabilistically (Nenkova and McKeown, 2011).

The Sum TF.IDF method tends to prefer longer sentences because it takes the sum of the word scores. An alternative to this, the centroid method (Radev et al., 2000), scores sentences by their similarity to the centroid of all the sentences in the document. The centroid is a hypothetical proto-sentence containing exactly the average counts of each term from all sentences in the document. Centroid is described in detail in Section 4.5.1, and is used in this thesis. The most common calculation of centroid is very similar to sum TF.IDF, with the key difference that scores are normalised by the length of the centroid and of the sentence.

While the Centroid method scores sentences according to their similarity to one global centroid, a document or collection of documents may contain multiple topics, all of which should be represented in the summary. The graph-based methods, TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004) achieve these summaries by creating a graph of the sentences in the document. Edges are weighted the similarity between pairs of sentences, and nodes in the graph are assigned scores according to their random-walk centrality, which is the chance that a random walk from any point in the graph will arrive at that node (sentence). A further explanation of the TextRank algorithm is given in Section 4.5.2.

Though word frequency information can be applied to tweets as with longer documents by counting the number of terms, consideration must be given especially to the short nature of tweets, which are close in length to single sentences than to the

documents used as the target for summarisation in generic newswire summarisation.

2.2.3 Context and Term Position

There are many genres of text where documents all share some elements of structure. For example, in Wikipedia documents the first paragraph generally contains a succinct summary of the object being described, and news that is written in the pyramid style will contain key information in the leading paragraph, to the extent that the original Document Understanding Conference task of generic summarisation of news articles was discontinued after two years because no system could outperform the baseline of simply selecting the lead paragraph (Nenkova and McKeown, 2011). These common structures can be leveraged as features for summarisation.

Additionally, certain types of document contain strong textual hints of salient information. Scientific papers may have phrases beginning with ‘In conclusion...’, for example. These cues have been exploited in extractive summarisation (Baxendale, 1958).

The structure of a collection of tweets is nothing like an article, blog post, essay, abstract of any such traditional media. Tweets are more conversational, with replies and topical shifts. They do not organise themselves into coherent discourse, and there may be many such discussions occurring simultaneously.

There is no linguistic continuity between one tweet and the next in a standard timeline, except where a user has created a run-on by posting several tweets in a row (optionally signalled by the use of ellipsis). There are no subsection headings, no paragraphs, no titles.

Tweets lack traditional structure that may be seen in other kinds of documents. Unlike the context of a sentence in a paragraph, the context of a tweet consists of its author and the preceding and following tweets, as well as potentially other tweets that share the same hashtags.

2.2.4 Machine Learning for Summarisation

There are a great number of different kinds of information that can be used to summarise documents automatically. Frequency-based scores, key phrases, positions in documents and in paragraphs may all need to be combined to create an optimal scorer.

In much work, this is achieved by using the weighted sum of multiple features to give an overall relevance score (Edmundson, 1969; Saggion, 2008). This requires

considerable empirical work to determine the correct weights for individual features, meaning that in practise the number of features which can be used is limited. The linear addition of features used also assumes that the relationship between each feature and the overall relevance score is both linear and independent.

Machine learning for summary generation offers the ability to combine many more features with more powerful models, but with the downside that these models require additional training and can not always be inspected by humans to give a deeper understanding of nature of summarisation. Another key problem is the provision of training data required for supervised models - often this takes the form of annotators marking certain sentences for inclusion or exclusion from a ‘gold standard’ summary, which is not the natural way in which summaries are usually formed (Nenkova and McKeown, 2011).

The ease with which diverse features can be combined has spawned a great deal of work in the area of machine learning for summarisation. Early learning algorithms used for this purpose included decision trees such as C4.5 (Quinlan, 1986; Chuang and Yang, 2000; Mani and Bloedorn, 1998), which have the advantage of producing rules which are easily inspected by humans (Mani and Bloedorn, 1998), and Naive Bayes classification (Mitchell, 1997b; Chuang and Yang, 2000; Neto et al., 2002).

More recently, nearly every possible machine learning algorithm has been applied to the task of automatic summarisation (Nenkova and McKeown, 2012). Extensions to Support Vector Machines (Joachims, 1998) have been made specifically to deal with ranking, for the task of information retrieval (Joachims, 2006), though this ranking approach may also be applied to summarisation.

The strength of machine learning algorithms for summarisation is in the ability to combine large groups of diverse features without large amounts of experimentation. This will be exploited in this work to use the meta-data attached to a tweet alongside its actual text. There are some limitations for machine learning in summarisation. The need for training data creates a labour intensive annotation task. Supervised machine learning algorithms cannot easily be trained on all of the criteria that form good summaries, such as coverage and coherence, and instead must be trained on simple judgements on whether or not a sentence would be included in a hypothetical summary.

2.2.5 Generating Summaries

Redundancy can be a problem when multiple text units occur which closely repeat one another. Having too much of the same or similar information in a summary will give a result which does not cover all of the key topics. This is especially likely in multi-document summarisation, where authors cannot be assumed to have made any effort to avoid repeating one another. Diverse summaries have been achieved in some work by greedily selecting sentences which have little in common with those already in the summary, using a simple threshold on similarity between a new sentence and those already in the summary (Saggion, 2008), or by weighting the similarity against relevance scores (Carbonell and Goldstein, 1998).

2.2.6 Sentence Modification

The problems with sentence selection as the sole basis for extractive summarisation are numerous. Phrases which structure the discourse do not make sense when re-ordered. For example, “in conclusion” should not appear at the start of a summary. Sentences can contain anaphora - references that are resolved in other sentences. For example in the sentence “He can’t be trusted with this”, what is meant by ‘he’ is unclear. Conjunctions that appear at the start of sentences (such as however, or but) make little sense without the preceding sentence for context.

Sentences in long-form text can contain detail that is not essential to the overall meaning of the summary, or which is included elsewhere in the summary. For example, the sentence:

“The British Prime Minister, David Cameron, made a controversial statement today on Twitter that the left-wing leader of the opposition, Jeremy Corbyn, is a threat to national security due to his policies on unilateral nuclear disarmament and other areas of foreign policy.”

Could be reduced to give:

“Prime Minister David Cameron made a statement today that Jeremy Corbyn is a threat to national security due to his foreign policy”

This short form contains much of the key information but far fewer words.

The process of modifying text to form stronger summaries was described by Mani (2001) as text revision, and broken down into three distinct methods - shallow coherence smoothing, to deal with anaphora and gaps in the discourse, text compaction to shrink text using simple methods, and full revision, to both compact the discourse and streamline different information within sentences.

Coherence smoothing can take the form of anaphor substitution, though as argued by Nenkova and McKeown (2011), readers are very sensitive to errors caused by this kind of substitution. Anaphora resolution and substitution was carried out by Dalli et al. (2004) to help summarise email threads, where, they argue, anaphor presents a greater problem than in many other media. Some of the work excludes all sentences which appear to contain anaphor (Brandow et al., 1995), or delete anaphor if the required sentence is not already in the summary (Nanba and Okumura, 2000). The latter work also removes conjunctions which may start a sentence, such as “However”, if the previous sentence is not included in the summary.

Sentence compression can help form summaries which are closer to those which would be created by human annotators (Nenkova and McKeown, 2011). Sentence compression has been carried out using human-created rules (Conroy et al., 2006; Dorr et al., 2003) and probabilistic models over grammar tree-structures (Wang et al., 2013; Turner and Charniak, 2005; Knight and Marcu, 2000). A number of shortening rules were applied by Corston-Oliver (2001) to compress emails using ‘text-speak’ for viewing on mobile devices.

Given that tweets are already made extremely short by limitations on their length in characters, it is unlikely that sentence compression will be needed to make them shorter. To our knowledge there have been no attempts to shorten tweets further using sentence compression techniques. Tweets do, however, contain anaphora and can begin with conjunctions, though the extent to which this affects the summarisation task is as yet unknown.

2.2.7 Evaluation

Efforts to evaluate the quality of summaries have covered both objective scores about how much of the target document’s content they cover, and subjective judgements about aspects of the quality of the summary. Key questions in the latter include the cohesiveness, which includes whether or not the structure of sentences in the summary is effective and flows naturally and coherence, which addresses the extent to which the topics in the summary flow naturally from one to another without disorienting the reader with sudden topical shifts or interjections.

Judgements of the content of a summary can be calculated automatically or manually. In the Text Analysis Conference (TAC), a major shared task for automatic summarisation, summaries were evaluated for their content manually using the Pyramid method, in which model summaries are compared to automatic summaries using

a somewhat rigorous approach to comparison (Nenkova and Passonneau, 2004). The Pyramid method consists of extracting summary content units (SCUs) from a model summary and calculating the coverage of those SCUs in an automatic summary. The TAC also considered subjective evaluation scores for summaries through “Readability/Fluency” and “Overall responsiveness”, though Nenkova and McKeown (2011) argue that the emphasis on content coverage in the shared task discouraged research on linguistic quality of summaries.

The task of manually evaluating summaries using the Pyramid method is time consuming. Automatic evaluation using textual content overlap was therefore proposed by Saggion et al. (2002) using cosine similarity or longest-common substring between a peer and a model summary. This approach was later extended as part of the ROUGE package for automatic evaluation of summaries, so named after the BLEU package for automatic evaluation of translations (Lin, 2004a). These automatic evaluation methods were initially shown to correlate well with manual judgements of content coverage. Further discussion of ROUGE and its application within this thesis can be found in Section 4.7.

The ability to quickly evaluate summaries against a gold standard allows for the rapid comparison of the content selection carried out by many different systems or system variants, and as such can be invaluable for developing summarisers. However, automatic evaluation cannot by itself account for subjective elements of summaries such as fluency and coherence.

2.3 Twitter Summarisation

This section describes the state of previous work in Twitter summarisation, broken down into a series of related problems for which the possible approaches and the evaluation methodology differ. Many of the summaries of tweets are actually rankings, though some limited attempts have been made to generate phrase-based summaries of tweets.

2.3.1 Phrase Generation

These are methods that attempt to take a collection of tweets related to a topic and produce a short and coherent summary of the information in that collection. This is the most similar to general summarisation in the sense that it produces continuous text. All of the approaches of this type carry out text generation.

Much of the work has focussed on summarising trending topics, which are popular discussion threads from Twitter as a whole. The Phrase Reinforcement algorithm (Sharifi et al., 2010) builds phrases out of common bigrams by constructing graphs of connections between them, before generating optimal paths through this word graph to create phrases. Phrase reinforcement has also been extended to take into account the grammaticality of generated phrases (Judd and Kalita, 2013).

The SumTweets system described by Wang et al. (2010) evaluated a number of algorithms for tweet summarisation by comparing candidate summaries of trending topics to those provided by the What The Trend service, through which descriptions for the topics are crowd sourced. The algorithms evaluated included term frequency, LexRank and ‘longest subsequence’, similar to Phrase Reinforcement.

2.3.2 URL Recommendation

Tweets have been summarised using information from the web that is not contained in the tweets themselves. This is possible because tweets often contain URLs, which point to web resources. In one example of this, Abdel-Hafez et al. (2014) collected tweets which refer to films on IMDB and ranked them for interestingness on the basis of the film’s budget and IMDB scores. This work was very limited because it could only be applied to tweets that refer to films on IMDB.

In contrast, Chen et al. (2010) propose a recommender system for URLs in tweets, ranking both by predicted topical relevance and by social voting. While this is not as restricted as only ranking tweets that contain IMDB references, this variant of the task is fundamentally very limited, since it ignores the role of Twitter users are authors as well as readers. Only content that is generated by a professional author, copywriter or blogger can be recommended in this way. Much useful and interesting content on Twitter is not generated by celebrity or professional authors, but by everyday users of the platform.

2.3.3 Extractive Summarisation of Tweets

This section views the task of summarising tweets as an extractive problem of finding units of text that best represent the interesting content from a larger collection of tweets. One might extract sentences, or whole tweets, or collections of tweets, in this manner.

Sentences: Simply using sentences for tweet summarisation would allow us to cast the problem in a way that is similar to traditional textual summarisation. However, tweets are often very informal and it is not uncommon to communicate in fragments rather than complete sentences. Additionally, discourse markers that do not exist in general language are commonplace on Twitter (RT, via, @mentions etc), and the poor quality of language including misspellings and ad-hoc contractions would make it difficult to combine sentences from tweets in paragraphs.

Tweets: An individual tweet is typically be self-contained (though sometimes tweets “run-on” or form part of a dialog), as such, they could be used as candidate units for summarisation, without requiring modification or analysis to ensure for example that they contain no unresolved anaphora. The main implication of choosing tweets as units for summarisation would be that the summary itself becomes more of a ranking of tweets according to estimated relevance to the reader. To form a summary, an entire collection can be shown in order of relevance, or a subset of the most interesting tweets can be shown. Summaries consisting entirely of tweets mitigate the problem of coherence that is found with those formed from sentences, though they are still at risk of redundancy.

Multiple Tweets: Although tweets are generally self contained, it is possible that a single tweet in isolation cannot be read because it relies on context which would normally be available in the home timeline. For example, it might refer to previous tweets by the same author, or a general ongoing discussion which has, at the time of reading, already ended. Some sets of related tweets can be trivially connected together using metadata from the Twitter API. Tweets which belong to chains of replies can be identified by the ‘in_reply_to’ field. Clusters of tweets in the same argument by a single user could be identified by the use of ellipsis, or by the time between posts. Tweets from a single user are likely to be part of a single discussion if they occur in quick succession.

The remainder of the related works in this chapter address the tasks of tweet ranking and tweet recommendation, in which whole tweets are scored according to quality or relevance. In practise it may be possible to take promising tweet ranking methods and apply them specifically to sentences within tweets, since the mean number of sentences in a tweet, according to a sample of the Spritzer API, is 1.8.

The ranking of input units, either sentences or tweets, could be viewed as only one part of summary generation. Summaries may be produced by ordering units by

their relevance rank, or by introducing some method to reduce redundancy (Inouye and Kalita, 2011). To our knowledge no existing work has attempted to produce extractive summaries for tweets which are designed to be read as continuous prose; to do so would likely require large amounts of modification of tweets as part of the summary generation process to remove discourse markers which don't normally appear in prose (like retweet markers) and Twitter-specific dialogue acts (callouts to followers, references to earlier tweets) which would hinder readability in this context.

2.3.4 Tweet Recommendation

Recommender systems are those which provide users with a list of items which they may prefer, helping users to decide between them (Shani and Gunawardana, 2011). The term tweet recommendation is used in this work to describe the act of recommending tweets which may be of interest, with the additional distinction that tweets are not drawn from search results or from the target user's home timeline, but rather the whole of Twitter (or a sample of it).

Collaborative filtering is a class of method used in recommender systems whereby objects that are endorsed by a number of users are recommended to other, similar users (Sarwar et al., 2001). These are used by online retailers and streaming services to recommend content based on personal and global viewing habits; recommendations may be presented in the form 'other users who bought product X also bought product Y' (Shani and Gunawardana, 2011).

Kim and Shim (2011) generate these recommendations using a generative model for the probabilities of various Twitter activities, trained using Expectation Maximisation (EM). Their system operates in a selection of limited domains, recommending both users and tweets. They do not incorporate any linguistic or quality based features of the tweets themselves. Ren et al. (2013) also produce generative models for tweet propagation, going beyond collaborative filtering to consider the social circles of users, as well as time and novelty. Meanwhile, Chen et al. (2012) model the likelihood of one tweet being more relevant than another, and incorporate a number of diverse tweet features into their models, including information about the reader's own social network.

Yan et al. (2012) recommend tweets based both on the relationships of the reader and the semantic relatedness between pairs of tweets. User interests are represented using topic models, and a graph is formed which is then co-ranked using a random-walk algorithm. Their work bears a similarity to that of Sun and Zhu (2013), which

compares the immediate social network of a candidate author with that of the target reader.

The likelihood of a given tweet being retweeted was also studied by Suh et al. (2010). Considering a number of features, they found that URLs and hashtags play an important role there in determining this likelihood. Our own experiments, as described in Section 3.3.2, showed that URLs and hashtags were not much more prevalent in relevant tweets than in Twitter as a whole (Hashtags: 33.79% relevant vs 35.29%, URLs: 52.11% relevant vs 53.37%).

All collaborative filtering systems are limited to recommending tweets on the basis of observable endorsements of content, or other publicly viewable material. Most commonly, recommender systems consider retweets as the visible signal of approval of posts (Suh et al., 2010; Chen et al., 2012; Yan et al., 2012), though not all kinds of tweets that are interesting are retweeted, and very new tweets are not yet retweeted (Section 3.5).

This thesis differs from work on generic tweet recommender systems, in that problem of ranking the incoming posts of a specific social user is considered. Whereas recommender systems assume that interestingness can be calculated on a global scale by matching users topically, tweets in the home timeline may be interesting for personal and social reasons that occur on a very local scale.

2.3.5 Search Ranking

When searching through the whole of Twitter, a large number of tweets may be nearly identical. For a simple Boolean search, especially when the query is a trending topic or popular hashtag, one would expect to see a lot of redundant or irrelevant information. As such, it is useful to rank tweets in the query according to relevance automatically.

Using a keyword Twitter search as an input set, a number of different solutions have been proposed to the problem of prioritising tweets within search rankings. A shared task of this type was administered in a TREC 2011 Microblog retrieval task (Ounis et al., 2011).

A number of authors (Choi et al., 2012; Duan et al., 2010; Huang et al., 2011) introduce the concept of tweet quality and reliability. In the context of a collection of search results, Choi et al. (2012) rank tweets according to a model of quality based on feature such as originality (not a retweet), the number of terms, fraction of stopwords, TF.IDF and a series of basic linguistic statistics such as the fraction

of English terms, presence of a question mark and fraction of unique terms. They employ the Kullback-Leibler divergence score (Kullback and Leibler, 1951) between tweets that have been retweeted and those that have not as a further feature.

The quality of a tweet might be measured in terms of the number of out-of-vocabulary words, as well as the PageRank popularity of Twitter authors as well as the popularity of the hashtags it uses, as in Duan et al. (2010), who carry out greedy feature selection based on normalised discounted cumulative gain scores, which are intended for evaluation of ranking systems, and learn rankers using RankSVM. Using similar features, but incorporating both labelled and unlabelled data in their model, Huang et al. (2011) demonstrate that they can outperform this earlier work, whilst also incorporating new features including user profile completeness and sentiment scores.

The measures of quality in the related work fall broadly into a few general categories; the popularity of the author, the popularity of the content, and the linguistic distribution of terms in the tweet itself. Much of this work makes a number of assumptions about tweet quality, including that high-quality documents share similar content, known as the content conformity hypothesis (Huang et al., 2011). The quality of a tweet is also assumed to be independent of the reader, and as such is implied to be objective, though later work has since shown that this is not the case, and that it is extremely difficult to obtain high agreement between annotators for tweet quality (Alonso et al., 2013).

In Huang et al. (2012) it was shown that the number of URLs that a tweet contains can form a strong signal for relevance even on its own.

There exist measures, such as TextRank and Centroid (Mihalcea and Tarau, 2004; Radev et al., 2004), which do not attempt to directly model the quality of a text unit but rather its centrality within a collection; usually by establishing commonly used vocabulary and prioritising content which best addresses that vocabulary. Both of these have been applied to the task of ranking Twitter search results (Becker et al., 2011; Inouye and Kalita, 2011). Selecting tweets in this way can lead to redundancy, since tweets that are retweets or near-retweets will score the same as one another, especially under Centroid. This can be mitigated somewhat by taking measures to reduce redundancy (Inouye and Kalita, 2011); a step which was empirically demonstrated to be of importance by De Choudhury et al. (2011). This thesis also considers Centroid and TextRank, and enforces diversity using Maximal Marginal Relevance (Carbonell and Goldstein, 1998).

2.3.6 Home Timelines

The Twitter home timeline is the reverse-chronological collection of all the tweets by all the authors to whom the logged in user has subscribed. Given that home timelines are the default view of Twitter, and they are highly personalised, acting as a hub for much of the communication on the service for many users, they present a very important target for summarisation. More discussion of Twitter home timelines appears in Section 2.1.4.

A body of existing work has addressed the task of ranking tweets in Twitter home timelines according to their likelihood of being retweeted. In this version of the ranking task, the results are personalised for a specific user, though the type of personalisation that is possible through retweet prediction is limited in terms of the types of tweets that are likely to be recommended, and the immediacy of the recommendations, as discussed in Section 3.5.

Features used to predict retweets in personal Twitter timelines represent information about the relationship between author and reader and the content of the tweet itself. Some of the work also used author authority features (Uysal and Croft, 2011; Feng and Wang, 2013), the previous writings of the reader (Hong et al., 2013), and the novelty of the tweet as compared to all others by the same author (Uysal and Croft, 2011) as features for tweet relevance.

These features were combined to produce tweet rankings through graph factorisation by Feng and Wang (2013) and Hong et al. (2013), similar in spirit to collaborative filtering. A learning to rank approach, Coordinate Ascent, was used by Uysal and Croft (2011) along with decision tree classification. The latter work is the only which reports results by feature groups, reporting that in predicting retweets, features which were based on the content of the tweet (including novelty) and features which were based on the user's own posts were outperformed by simple counts of tokens like question marks and URLs.

Home timelines, which are ranked in this thesis, present an unusual challenge compared to other popular types of tweet summarisation because they are both very dynamic and very personal. It is legally challenging to distribute evaluation data for the task, and as such none of the mentioned work has been evaluated on shared data.

The Twitter home timeline is personal to the reader, so the use of popularity methods for evaluation is not appropriate. This thesis will instead evaluate home timeline ranking on a subjective data set generated manually by the readers to which

the timelines belong. Carrying out evaluation in this way changes the task since further use of personalised features is possible, and these are used extensively in this work. The reasons for which tweets are relevant can be quantified, which is impossible when scraping retweets directly (some reasons for retweets were discovered via survey by Boyd et al. (2010)).

The data in this thesis is also focussed around specific topics, since readers were asked to filter their timeline prior to annotation, whereas existing work on retweet prediction does not consider specific topics, instead predicting retweets from the whole timeline. This additional step allows for the social media monitoring use case (Chapter 5), and for the evaluation of text-based methods from topic-focussed summarisation (Chapter 4).

2.3.7 Evaluation

The type of evaluation that is used to measure the quality of summaries of tweets depends on the specific variant of the task that is being considered.

Human generated gold standards and manual evaluation are used for approaches that generate summary phrases (Sharifi et al., 2010; Judd and Kalita, 2013), though Wang et al. (2010) use definitions from What The Trend (which are crowdsourced) as gold standard summaries. Both types of evaluation are carried out on large scale social media trends, and in the latter case could not be adapted to a personal scale.

Likewise, for the summarisation of search results data sets are annotated in response to queries by human annotators, as in the TREC 2011 microblog track (Ounis et al., 2011). On the other hand, tweet recommender systems often recommend tweets based on the likelihood of their being retweeted, and thus use retweet counts as the gold standard (Suh et al., 2010; Chen et al., 2012; Yan et al., 2012), though some also use self-reported judgements (Chen et al., 2011). Retweets are created for content promotion (Boyd et al., 2010), and are sparse, so these works cannot be assumed to predict general tweet interestingness. Where work attempts to personalise tweet summaries, this personalisation is limited by the impersonal nature of retweets.

For textual summaries of relatively small and topically coherent collections of tweets, researchers have used the standard DUC evaluation measures. For instance, Sharifi et al. (2010) gather 100 tweets per topic for 50 trending topics on Twitter, and then ask volunteers to generate summaries which they feel best describe each set of tweets. The automatic summaries produced by their algorithm are then tested

using a manual content annotation method, evaluating whether or not the same content is available as in the gold standard, and automatically using ROUGE-1. Similarly, Harabagiu and Hickl (2011) make use of model summaries, although they eschew ROUGE in favour of the Pyramid method (Nenkova and Passonneau, 2004), in which summary content is compared by hand by human annotators.

All of the discussed work that has attempted to rank content in Twitter home timelines has used retweet counts as the gold standard. Section 3.5 questions the utility of retweets as a gold standard for this task, arguing that the task is substantially different when judgements are reported by the readers themselves. Manual annotation of home timelines was used for Facebook by Paek et al. (2010), as in this thesis.

For tasks which use human generated judgements as a gold standard, subjectivity in the generated results may prove problematic; Alonso et al. (2013) ask users of a crowd-sourcing service to judge the relevance of a series of tweets, and found very poor agreement between annotators. Mackie et al. (2014) argue that evaluation of tweet summaries might be carried out without supervision data, using SIMetrix (Louis and Nenkova, 2013), which compares summary similarity to the source text.

2.4 Social Media Summarisation

In much of the related work, as in this thesis, the social network of choice for summarisation is Twitter. As discussed in Section 2.3, this summarisation can take a number of forms, including content recommendation, home timeline ranking, and search result ranking. There exists a related body of work which attempts similar forms of content recommendation, but outside of Twitter itself.

These works have addressed platforms which vary in terms of their similarity to Twitter, from microblogging services such as Sina Weibo (Bian et al., 2013) and Plurk (Weng et al., 2011), to large general purpose services like Facebook (Paek et al., 2010; Koroleva and Röhler, 2012; Bourke et al., 2013; NTALIANIS et al., 2013). Social media ranking and summarisation has also been extended into professional networks such as LinkedIn (Hong et al., 2012; Agarwal et al., 2014), among other, bespoke networks (Guy et al., 2015).

Though these networks differ in terms of their functionality and purpose, the purpose of previous summarisation work is generally to discover and promote interesting content, in order to capture user interest. There are exceptions to this, for

example (Weng et al., 2011) specially worked to rank tweets which were intended for question answering.

The existing approaches to social media summarisation all attempt the ranking of status updates or content (such as links). This is generally achieved through some method to combine multiple kinds of features in order to rank or recommend content.

The work of Hong et al. (2012) is unusual in that it is, to our knowledge, the only approach to summarising a network outside of Twitter that explicitly uses collaborative filtering and matrix factorisation to recommend content on a large scale. Other work manually weights social and textual features to produce rankings (Berkovsky et al., 2011; Guy et al., 2015) or uses anonymous graph based scores for ranking (Namaki et al., 2013). The remaining approaches are all based on some form of machine learning.

Machine learning is carried out using discriminative models such as SVM (Paek et al., 2010; Weng et al., 2011) or logistic regression (Agarwal et al., 2014), models with latent variables (Hong et al., 2012; Bian et al., 2013) and learning to rank approaches (Hong et al., 2012; Bourke et al., 2013). Koroleva and Röhler (2012) also demonstrated the use of neural networks in ranking Facebook posts.

The types of features that are exploited in ranking content do not vary much between platforms. The majority of the work incorporated social features in some way, including the strength of connections between users, estimated using friends in common and previous posts (Paek et al., 2010; Berkovsky et al., 2011; Hong et al., 2012; Koroleva and Röhler, 2012; Namaki et al., 2013; Agarwal et al., 2014). Other features relied on the overall popularity of the posts themselves in predicting interestingness, based on counts of interactions such as clicks, likes or shares (Paek et al., 2010; Bourke et al., 2013; Guy et al., 2015).

Features based on the text of posts themselves have been less widely adopted. Hong et al. (2012) attempt to estimate the ‘professionalism’ of a post and use this as a ranking feature. Similarly Koroleva and Röhler (2012) predict the readability of candidate posts. Paek et al. (2010) found that their best features were similar to those that may be used in summarisation of longer text, including n-gram frequency and Sum TF.IDF. Ranking according to social features, in their work, gave far inferior results.

Other work exploits the connection between posts and their context, such as previous posts of the reader and author (Weng et al., 2011; Bourke et al., 2013; Guy et al., 2015). These features were not found to be particularly effective in the work

of Weng et al. (2011).

Koroleva and Röhler (2012) presented a survey in which they asked users whether their preferences were driven by affective (like/dislike), cognitive (boring/interesting) or instrumental (useful/useless) views of the content, with different features giving different correlations with each of these dimensions of relevance, though the differences were not large.

The summarisation of networks other than Twitter is less well developed or expansive than work with Twitter, with less exploration outside of standard machine learning and learning to rank methods, and fewer variations of the problem itself. Nonetheless, the features that have been successfully exploited are similar both between platforms and to those used for tweets. This is encouraging, as it demonstrates that work in Twitter summarisation is likely to be transferable to other networks.

2.5 Discussion

As new kind of medium, microblog networks and Twitter in particular are fast moving and unique in terms of both the types of communication they facilitate and how they go about it. Beyond limiting the size of individual posts, they encourage users to adopt new conventions such as reposting, hashtags and mentions, and create new structures for browsing data such as the home timeline.

Microblog summarisation is worthy of study not only because it differs dramatically from summarisation of text collections linguistically, but because its subjective and personal nature presents a strong use case for user-focussed summarisation.

This chapter presented a brief discussion of text summarisation, including the notion that the context of a text unit can be important in extractive summarisation, in terms of its position in the document and its representativeness of the document as a whole. The context of a tweet in a tweet ranking task is completely different. For example, the context may be a wider collection of tweets that all contain a searched keyword, or it may be the relationship between the author and reader, depending on the exact variant of the task at hand.

An overview was given of the types of summarisation that have been carried out for tweets in the related work, including the generation of explanatory phrases for trending topics, recommending tweets or URLs from the whole of Twitter, ranking search results and ranking home timelines. Much of the work uses retweet counts as the target for summarisation, and as such can be thought of as retweet prediction.

The remainder of this thesis will address the task of Twitter home timeline sum-

marisation, by ranking tweets according to their predicted interestingness. While this is just one type of Twitter summarisation, it is one of the more challenging variants of the problem, since timelines are not focussed around a single topic, and since readers can reasonably expect to see tweets of a social nature and tweets by their friends (not just celebrities) in summaries of the home timeline.

There exist a number of other challenges that are present for many types of automated Twitter summarisation. Common text summarisation algorithms such as Centroid and TextRank assume that collections are coherent and that ideal sentences are those which share the most vocabulary with the rest of the document, though this is not always true, especially when tweets can easily be exact replicas of one another (retweets), boosting such scores. Inouye and Kalita (2011) compared several extractive summarisation algorithms and reported that Centroid and graph-based text ranking algorithms did not perform well. It is also difficult to correctly weight out-of-vocabulary words according to importance, and such terms are very common within the language of Twitter.

There is great deal of *redundancy* in a Twitter stream, since tweets are contributed by a number of authors. Large news events may garner many independent tweets with very similar themes, and the same article or event could be shared dozens of times. Tweet redundancy has been exploited by some (Sharifi et al., 2010) and mitigated by others (Inouye and Kalita, 2011), but it has been shown to be a significant issue in the ranking of Twitter search results (De Choudhury et al., 2011).

Naaman et al. (2010) found over 40% of their sample of tweets were “me now” messages, that is, posts by a user describing what they are currently doing. Next most common were statements and random thoughts, opinions and complaints and information sharing such as links, each taking over 20% of the total. Less common tweet themes were self-promotion, questions to followers, presence maintenance e.g. “I’m back!”, anecdotes about oneself and anecdotes about others. They group users into *informers* and *meformers*, where meformers mostly share information about themselves.

A key purpose this thesis work is to allow for Twitter home timelines to be judged by the authors for whom they are intended, creating a greater need for personalisation than in existing work which relies exclusively on the popularity of the tweet for home timeline ranking (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013). The ranking of home timelines requires that summaries are personalised to the reader. For tweets such as “me now” posts, the relationship

between author and reader will be important in determining whether the reader is interested. There are also many different reasons to use Twitter (Chen et al., 2011), and the types of content a user wishes to see depends on their reason for using the service.

Twitter posts are often very context-dependent. While larger texts may reference one another, the type of reference is usually semantically clear, for instance an inline link between articles providing context, a citation providing proof of a statement. References between tweets can have many purposes (getting attention, direct reply, group reply, public criticism). Tweets can rely on the user having read the previous tweet (or retweet) by an author, and this is only sometimes signalled by the relevant typography (“...” or “LRT” meaning Last ReTweet). The semantic role of a mention between users is not explicit either; they can be recommendations, direct messages or attempts to provoke a response, for example.

The remainder of this thesis will explore these challenges in order to understand their relevance to the task of home timeline ranking, and the ways in which they affect the performance and evaluation of such ranking methods.

Chapter 3

Task Definition and Gold Standard Creation

3.1 Introduction

There are many different forms of Twitter summarisation; a full review of which can be found in Section 2.3. This work, however, specifically addresses the ranking of tweets in Twitter home timelines. Home timelines contain all incoming tweets from all authors to whom a reader has subscribed, or ‘followed’. This is the default view of Twitter for every user.

Home timelines are not filtered in any way, and as such contain discussion of very diverse topics and posts which are created for any number of different reasons. The timeline might contain discussion of news and current events, information about the reader’s hobbies, friends announcing life events, and social chat. This diversity, and the resulting potential for information overload (Beaudoin, 2008), increased by the lack of tools to filter such timelines (Hargittai et al., 2012), also make this a practical and useful target for summarisation.

When search results or trending topics are summarised, the information need of the user is expressed in terms of a keyword query. On the other hand, there is no clear single information need for the home timeline; users may have an implicit business goal, or they may simply wish to be entertained, or to socialise. A study of the different reasons for which tweets can be interesting is presented in Section 3.4. Information needs for the home time are inherently subjective, and can be expected to change from user to user, because the accounts being followed are personal to the reader. This subjectivity should define how the task of home timeline ranking

is evaluated.

This chapter presents a definition and exploration of a particular home timeline ranking task. A process is defined by which the very general home timeline can be filtered to discover specific topics of interest prior to ranking. The problem of evaluating these rankings given that the home timeline is likely very personal and subjective in nature is discussed. Two new gold standard data sets are curated for learning and evaluation, one of which captures the interests of a professional organisation which uses Twitter, and the other a general set of annotations from 148 others.

3.2 Task Definition

This section describes the specific scenario of use, or the setting in which this ranking will take place. In previous work, which relied on sparse counts of retweets, the candidates for summarisation have been drawn from the entire user home timeline (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013). In this thesis, users are instead asked to indicate the relevance of posts from their own timeline, so the setting in which tweets are annotated is viewed as a two stage retrieval process, composed of tweet collection followed by summary generation.

3.2.1 Retrieval Process

While an algorithm may attempt to summarise the entire home timeline, this presents the problem that most timelines contain many discussions by many different authors. Many text summarisation algorithms makes the assumption that all given documents are about the same topic or event (Nenkova and McKeown, 2011). However user timelines are diverse, not just in terms of topics, but also in terms of the types of tweet they contain (Naaman et al., 2010).

An undirected summary would need to be extremely robust, and very long, to be a useful indication of the content of the timeline. Therefore, the first step towards successful ranking of topically diverse Twitter timelines is to generate filters or clusters for key topics of interest.

In much of the related work, candidate subsets of tweets for summarisation have been generated using keyword search by researchers or by third party volunteers, who are not the users for whom the tweets were curated initially (Harabagiu and Hickl, 2011; Sharifi et al., 2010; Duan et al., 2010). Prior to ranking, users could

be asked to perform a similar keyword search over just the tweets in their home timeline. However, doing so would require that they already know what they are looking for; instead, an interface is developed to allow users to discover topics of interest in their timeline before creating a query using term selection.

The home timeline is presented chronologically in the Twitter UI, with users not expecting to see older tweets (Wu et al., 2011), so time-limited windows of the timeline are targeted for gold standard generation.

The basic retrieval process is as follows:

1. The user’s home timeline is fetched using the Twitter API and split into time windows.
2. A user chooses a temporally bounded window of tweets, or the most recent window is selected.
3. A tag-cloud representation of the terms in the window is shown to the user.
4. The user chooses one or more topics and entities of interest, and thus narrows down the set of tweets to be ranked.
5. The summary is generated and shown to the user.

The effectiveness of the system to support exploration of the data is not directly evaluated. The focus of this work is instead on the quality of the summary itself. A modified version of this process is also used to assist in the development of a gold standard, wherein the final step of summary generation is performed by a volunteer rather than automatically.

3.2.2 Tweet Filtering

In order to form coherent topics from the incoherent home timeline, users are first shown a tag cloud representation of a single window from the timeline. This view is naturally imperfect, containing noisy irrelevant terms and not necessarily promoting the most indicative ones. However, by asking users to select terms from the cloud (thus indirectly selecting the matching tweets), a sample of tweets is created that is topically coherent.

The tag cloud interface allows users to select several terms and the tweet sample is drawn from the union of tweets containing their selections. The size of the input set for ranking is fixed at 50 tweets. In cases when there are fewer than 50 tweets

containing the chosen terms, users are asked to keep selecting more terms until the sample is complete. If there are no more related terms left, the remaining tweets are drawn at random from the rest of the timeline. If there are more than 50 tweets in the sample, then 50 are chosen at random and the remainder discarded. An input set size of 50 was determined through a pilot study, as described in Section 3.3.2

Tag clouds are built either from single terms and usernames or from named entities taken from the home timeline. The colouring of the terms was weighted through a combination of IDF score the frequency with which the participants themselves use those terms or interacted with users with the given usernames. The clouds are animated in a moving 3D sphere to aid visibility when a large number of tokens are included.

When a term is clicked, the user sees live feedback on how many tweets contain terms from the current selection. The term selection interface is shown in Figure 3.1.

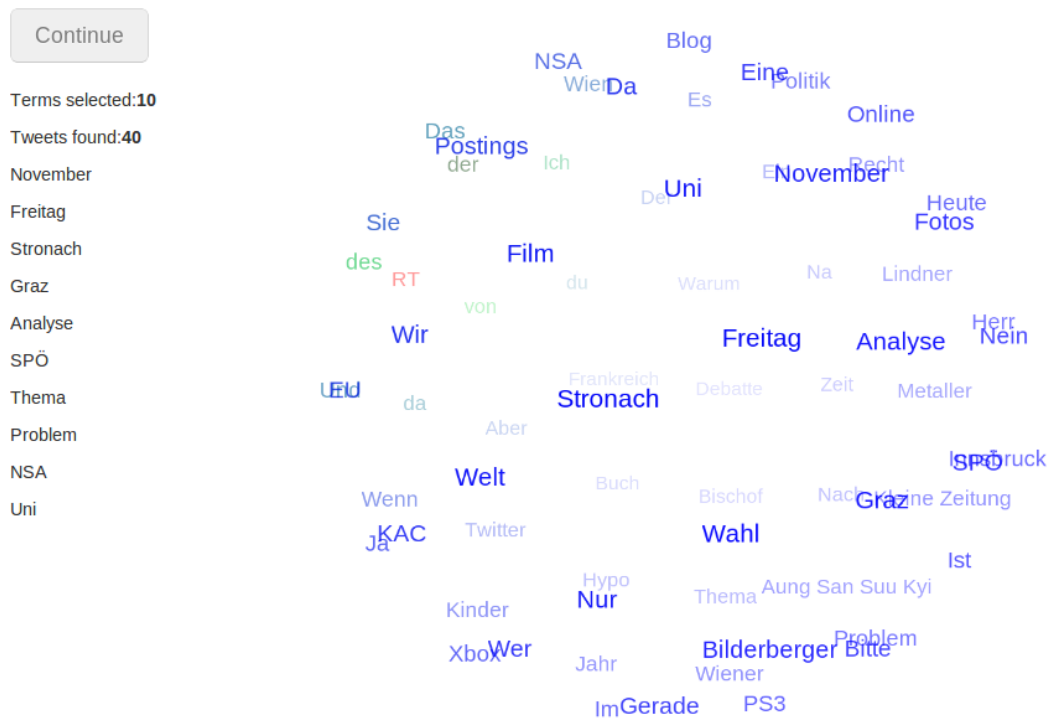


Figure 3.1: Selecting terms to indicate a broad topic

3.2.3 Creating Summaries

This work will use a variety of approaches to estimate the relevance of tweets to the target reader, prior to generating summaries. In practise, these estimates are interestingness scores, or an ordering of tweets from most to least interesting, as predicted by the algorithm.

The actual summary generated from this ordering of tweets can take one of two forms. The collection can either be truncated to a fixed size to form a summary, or the entire ordered collection of tweets can be shown to the user as a complete ranking. In the former case, it is important to ensure that the limited collection of posts shown to the user are of interest and represent the complete set well (Nenkova and McKeown, 2011), in the latter, the most interesting posts should be shown first, but if the system fails to do this, the reader can continue to inspect posts which were predicted to be of less interest. This is a key idea in the evaluation of search results in information retrieval (Manning et al., 2008).

Both types of summary generation will be used for this task, depending on the ways in which summaries are being evaluated (Section 4.7).

Once tweets have been ranked according to interestingness, or relevance to the query, it may be possible to form paragraphs of text by taking sentences or phrases from the top ranked tweets. To do so would likely require large amounts of additional processing, including pruning grammatical constructions that would not scan in extended text, and detecting run-ons between pairs of tweets. This additional step would also require a different kind of evaluation, manually rating the quality and coherence of the summaries. As such, this type of summary generation is considered to be outside the scope of this thesis.

3.3 Data Collection

Constructing a gold standard data set for home timeline ranking will allow the comparison of a large number of algorithms without going through expensive manual evaluation for each (Lin, 2004b). The availability of infinitely repeatable evaluation allows for comparison between many variants of a method, meaning that the impact on performance of individual features or parameters can be better understood. The statistics of the gold standard will drive the development of ranking features and algorithms.

The gold standard is created by asking volunteer users to annotate their own

home timelines, though they are not able to annotate their entire historical timeline since joining the service. A sampling or filtering process must be used to select a subset of the task to annotate for relevance. The method by which these samples is generated is influenced by the task definition from Section 3.2. Were retweets used instead, the entire timeline could be exploited, but retweets are sparse, limiting the ability to find annotated tweets around specific topics.

This section details the execution of a pilot study which led to the final definition of the annotation task. Two final tweet relevance gold standards are discussed, a social media monitoring data set produced by a political analysis organisation, and a general Twitter dataset produced by volunteer users from The University of Sheffield. Users annotated their own timeline because relevance annotation for tweets is subjective (Alonso et al., 2013).

Where allowed by the terms of ethical consent and the Twitter terms of service, the data described in this study will be shared with the wider research community to aid in replication and comparison for future work.

3.3.1 Pilot Data

For the pilot study, volunteer twitterers were recruited through a university mailing list to participate in a gold standard annotation process. The goals of this initial study were as follows:

1. To develop a more robust, representative evaluation data set, based on an analysis of the problem of re-ranking tweets.
2. To limit the data set as little as possible, so as to meet realistic information needs (not queries defined by the researchers).
3. To evaluate Twitter summaries on real timelines, as curated naturally by Twitter users.

The initial data set is not restricted by type of user, and forms a pilot study, allowing us to understand the quirks of creating such a gold standard for the problem, before further refining the annotation task and carrying out two further studies on larger selections of volunteers.

	Ed	Victory Carriages seems a pretentious name for a minicab office.	Interesting? <input checked="" type="checkbox"/>	Comment (optional): Sort of funny
	James	RT @kitation: Anyone who claims their hairspray is indestructible has never tested it in a Yorkshire wind	Interesting? <input type="checkbox"/>	Comment (optional):
	Ed	Q: If I stare into the sun is there a chance it will burn out my soul?	Interesting? <input type="checkbox"/>	Comment (optional):
	SheffieldSU	RT @ShefSabrecats: @SheffieldSU we have our GIAG after Easter and with a new squad we need a RT pleeeeeease!	Interesting? <input checked="" type="checkbox"/>	Comment (optional): Interested in event
	Craig	Whenever I have an amount of time off work, I destroy myself with lie ins, videogames and competitive swingball.	Interesting? <input type="checkbox"/>	Comment (optional):
	SheffieldSU	RT @BummitHitchHike: Huge well-done to our #bummit teams who have made it to Sofia! Great work; we can't wait to hear more stories! #tri ...	Interesting? <input type="checkbox"/>	Comment (optional):

Figure 3.2: Interface used to annotate interesting posts in pilot. Selected tweets are highlighted

User Study

The pilot data set was collected by recruiting 71 Twitter users and asking them to complete a very informal annotation task. The study was advertised via Twitter and Facebook and on a University mailing list. As a result, most participants in the study had a high degree of literacy, and received tweets largely in English. Several of the users also received and sent tweets in other languages.

After agreeing to take part in the study, the volunteer twitterers authenticated a simple web application to access their account. Their most recent 800 incoming tweets were downloaded and stored. Additionally, the first 100 of these were presented to the user for annotation (see Figure 3.2). They were then asked to read the feed and select the most ‘interesting’ tweets. By default, all tweets were marked as uninteresting and users were asked to select those that stand out.

In the early stages of the pilot study, no prior information was available about the kinds of tweets that users would find interesting. The annotation interface was modelled to resemble the original Twitter interface as closely as possible, in order to minimise the possibility that the experiment design could affect their behaviour. To drive further experimentation, and to better understand the behaviour of the volunteers, an optional free comment field was added to each judgement for users to enter the reason that they selected a tweet as interesting. About 5.8% of the interesting tweets were commented on by the participants.

Tweets were displayed in reverse chronological order. The usernames of their subscribed users were kept intact, and the thumbnail images from the associated profiles were shown. Links were expanded and made clickable. The target link content appeared in a second window to allow users to return to the study easily.

Variability

No limit nor lower bound was placed on the number of tweets that could be marked as interesting. As a result of this unrestricted approach, different users marked vastly differing portions of their streams. Some users selected only 3 tweets, while others selected as many as 50%. The mean number of interesting tweets was 16.34 out of 100.

The evaluation methods used to estimate summary quality will give very different scores where there are different ratios of relevant to irrelevant units, for the same system. Randomly ordering the tweets, for example, will score more highly when 50% of tweets are relevant than when 5% of tweets are relevant. This leads to a dramatically different performance score for each user, such that statistically significant performance improvements are difficult to reach and the task is inconsistent. Predicting the relevant 50% of tweets is a much easier task than predicting 5%.

In order to eliminate the variance in the number of tweets selected as interesting by the different users, the study was later revised to ask users to select a fixed number of 8 tweets as interesting, out of a sample of 50. This is similar to the word-based restriction on summary sizes, commonly used in evaluation initiatives for multi-document news summarisation, e.g. the TAC Guided summarisation task¹.

Whilst restricting the number of relevant tweets to 1 in 6 does force conformity onto participants and limits the dataset, this accommodation will make evaluation of prospective approaches far more meaningful. The chosen ratio matches the mean number of interesting tweets per annotation set in the pilot study.

Temporal effects

The pilot data set showed a strong bias towards newer tweets. Given Twitter's streaming nature, more recent tweets should naturally be of greater interest to the user. However, the effect observed was very strong, with users much more likely to choose tweets towards the very start of the list, as shown in Figure 3.3, which shows the distribution of ages for interesting tweets, and Figure 3.4 which shows the length

¹<http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html>

in minutes of annotated tweet sets. The former distribution is much more skewed towards newer tweets.

This effect is seen despite the relatively short period of time covered by each of the annotated sets. As Figure 3.4 shows, most collected samples of Twitter covered less than an hour of tweets, yet the skew to more recent tweets still exists.

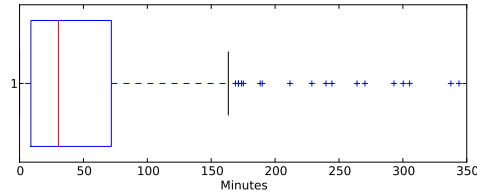


Figure 3.3: Age of interesting tweets in minutes at the time they were accessed by study volunteers. Outliers above 6hrs (8% of points) have been removed for readability. Actual upper limit is 48hrs.

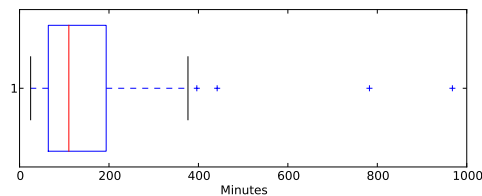


Figure 3.4: Total duration in minutes covered by tweets in annotated gold standard sets. Calculated as as the difference in time between oldest and newest tweet in set.

The temporal effect may not be as strong in reality as the data from this pilot study would suggest. The performance of newer tweets could also be attributed to annotator fatigue, where annotators do not read the later tweets in as much depth, or do not wish to change their earlier decisions when they reached the end of the set. To ascertain whether any preference for newer tweets is genuine, the annotation task was revised as follows:

- The size of the candidate set was reduced from 100 to 50, to prevent annotator boredom.
- A view of previous selections is made available all times, to aid with deselection.
- The tweets in the set were ordered randomly prior to display to the user.

The former change, the size of the annotation set, is to help compensate for the potential boredom of annotators, and to make it easier for them to review the

entire set. Allowing the removal of previous selections makes it easier for annotators to reverse earlier decisions when they find more interesting tweets. Ordering the candidates randomly divorces the position in the set from the time at which the tweet was posted; any tendency to select newer tweets after this change can therefore no longer be attributed to annotator fatigue.

3.3.2 Redesigned Personal Timelines Data Set

Addressing the weaknesses of the pilot study, and expanding beyond the unconstrained annotation task, a second study is carried out, drawing from the same pool of volunteers, albeit considerably later. Aside from a change in the wording of the questions asked of annotators, which will be covered later in this section, the design was evolved from the pilot study in the following ways:

- The size of the annotation set was fixed at 50.
- The size of the ‘relevant’ set was fixed at 8.
- Tweets to rank were shown in a random order, rather than reverse chronologically.
- Users were asked to select reasons for relevance from a list rather than a free text field.
- Tweet sets were filtered by topic prior to annotation.

The fixed annotation set size and random ordering are directly influenced by the results from the pilot data set, in which variation of the size of the interesting set, and a tendency to prefer tweets that appear earlier in the annotation view stymied comparison between algorithms.

The use of fixed relevance categories is an improvement to aid quantitative discussion of the motivations of the annotators. The categories are explained in Section 3.4. For this version of the study, most users (but not all) were also asked to filter their tweet set to a particular topic prior to annotation, creating a more coherent, constrained version of the gold standard creation process.

Creating coherent topics

An extra step is introduced into the annotation process to allow tweets to be filtered by topic prior to annotating tweets for interestingness. This filtering brings

the annotation task in line with the multi-stage timeline summarisation process described in detail in Section 3.2.2, in which tweets are first filtered by topic and then summarised automatically. Home timelines typically contain discussion of a large number of inter-twined topics, however, a useful summary may instead be formed on a subset of the timeline addressing a single topic.

After a window of tweet candidates for summarisation had been downloaded, a 3D term cloud representation was generated. In this version of the study, the terms in the cloud were selected according to a threshold on their frequency score. The frequency score for each term was calculated as the number of times it appeared in the timeline for annotation, weighted by the inverse document frequency score based on the corpus of tweets in the pilot study. A further discussion of this weighting scheme, known as TF.IDF, is given in Section 4.5.3. After applying a threshold for these scores to get key vocabulary, the screen names of authors in the collection are also added to the cloud. Terms in the term cloud were coloured with an intensity value that depended upon the scores.

Annotators were asked to interactively select terms from the collection and immediate feedback was given as to the number of tweets which would be matched by a binary union of the selected keywords. Term selection continued until at least 50 tweets were matched to the keyword set, at which point exactly 50 posts were selected at random prior to the next stage of the process, in which interesting tweets are selected.

The filtering interface was shown to the majority of users in the study. However, a random subset (less than 50% of the first run) of volunteers was not prompted to define a topic of interest in this way, in which case 50 recent tweets were selected at random. This provides a gold standard data set for a much harder and more general summarisation problem, which is not addressed in this work but which may prove useful in future experiments.

Selecting salient tweets

The interface through which users select interesting tweets was redesigned. Tweets were shown vertically using the Twitter bootstrap’s ‘media’ class, which produces a display very similar to the way tweets are shown on the Twitter website itself. User thumbnails were shown and links could be clicked.

The user interface was designed in such a way as to make it as easy as possible to both select interesting tweets and review the tweets that have already been marked

as interesting. Judgements could be removed easily, so that users could update their decisions as they make their way through the whole set. A screen shot of the interface is shown in Figure 3.5.

Annotating Reasons for Salience

In the pilot study, users were given the opportunity to give a reason for relevance as a note attached to a judgement. As described in Section 3.4, these reasons were codified and the interface was updated to allow users to select from a fixed selection of reasons.

Dataset Statistics

In March 2013, a group of 130 volunteer Twitter users took part in the second version of the user study. This first round of volunteers created a total of 322 annotated sets. In 2013, 215 of the sets were annotated using the full version of the annotation task including topic generation, whereas 107 were not filtered by topics and instead consisted a random subset of the annotator’s recent incoming tweets. 92 unique Twitter users completed the full annotation process including topic selection.

The user study was repeated in April of 2014, almost a year after the initial annotation. Given that the use of Twitter apparently changes over time, a second round of annotations removed in time from the first will give a more robust data set. A further 56 users took part this time, along with 5 from the 2013 set. These 148 users annotated a total of 172 tweet sets using the full annotation process, including topic selection.

The complete data set consists of 387 filtered tweet sets, annotated by 148 unique users. Counts of tweets in the corpus over time are shown in Figure 3.6. The spikes in tweet count are at weekly intervals, because all volunteers were emailed a week after their initial annotations to ask them to participate for a second, third and fourth time.

The tweets in the study were largely in English, though a few volunteers used Twitter in other languages. According to LangID (Lui and Baldwin, 2012), 92.78% of the tweets were in English, though in reality this portion may be higher or lower; the reported accuracy scores of LangID on two Twitter data sets are 0.941 (6 languages) and 0.886 (5 languages).

The mean length of a tweet in the data set was 109.58 characters, and 18.57 tokens. At least one hashtag was present in 35.29% of tweets. Other very common

features were URLs, which appeared in 53.37% of tweets, and user mentions, which appeared in 47.67% of tweets in the dataset.

In this version of the study, 72.50% of all tweets had been retweeted at least once, compared to 73.32% of tweets marked as relevant by participants; however, relevant tweets were retweeted on average 160.6 times, whereas irrelevant tweets had a mean retweet count of 84.4, a much larger difference. However, the retweet counts appear in an enormous range with a logarithmic distribution (a few tweets have very high counts, most have very low counts), so this difference does not by itself indicate a useful ground truth for relevance.

Although some of the volunteers did share some common interests (10.11% of the twitter accounts in the dataset were followed by more than one member of the study), only 1.34% of tweets appeared in multiple annotated sets. Therefore, it is not possible to use the judgements of one home timeline owner to corroborate the relevance judgements of another based on this data.

TF.IDF was calculated for each term to discover overall trends toward specific topics within the data. A list of the top scoring terms from 2013, 2014 and the combined data set is shown in Table 3.1. Stop words were not included in the list. Topics include sport, technology, music, politics and food. Few of the top vocabulary terms remained the same across the two years, although, in general, the types of topics under discussion did not change drastically.

Many of the popular terms in both data sets are locally focussed, with common topics including #SUVERSITY, a Sheffield university sporting event, and Hallam, the name of a university in Sheffield. Some of the terms stem from spam tweets, such as ‘win’ and ‘#followtrick’.

Table 3.2 lists popular users, with counts of their tweets seen in the study along with the number of those considered interesting by volunteers.

The terms that are much more likely to appear in interesting tweets than uninteresting, or vice-versa whilst still appearing with some regularity throughout the rest of the data can help show the topics that are of interest or that are not interesting to user. Normalised point-wise mutual information (NPMI) (Bouma, 2009) was calculated. This score is highest for features that are both very common and very predictive of a particular class. Unfortunately, there were very few common tokens that were only seen in the set of interesting tweets and the terms found occurred very infrequently in the dataset, e.g. ‘2005’, seen in 4 tweets, and ‘almond’, seen in 5.

Given that there were few terms which appeared only in interesting tweets, and

2013	2014	Both
#suvarsity	dropbox	#suvarsity
uni	#iomtt	dropbox
hallam	sheffield	sheffield
pope	#sheffield	uni
#travel	#teamfollowback	hallam
#bbcfootball	ar	pope
#chairman	di	book
england	cup	england
easter	matcha	easter
sheffield	#pepsiipl	#bbcfootball
international	#mivsrh	#sheffield
book	#whyimvotingukip	#travel
#weareinternational	#askstephandjamie	win
de	video	research
football	gaeilge	health
race	followers	video
#news	uk	#chairman
@mi_abaga	research	international
@sheffielduni	win	via
madrid	clegg	students
watch	hull	cup
via	ugh	#iomtt
win	university	season
tickets	STAR EMOJI	tickets
#driving	season	football
fans	#mgwv	uk
health	#retweet	madrid
varsity	health	music
#weareproud	book	amazing
real	#followtrick	@sheffielduni

Table 3.1: Terms from data set with high TF.IDF scores

Author	Interesting Tweets	Total Tweets	Origin
BBCSport	23	156	UK
forgesport	18	156	Sheffield, UK
sheffielduni	21	138	Sheffield, UK
digitalspytech	11	92	London, UK
BBCR1	8	83	UK
AP	4	65	New York, USA
BBCBreaking	9	58	UK
twirlyswirly	5	53	Edinburgh, UK
guardian	5	50	London, UK
WhatTheFFacts	11	48	Unknown
nationalrailenq	6	48	UK
gmpolice	8	48	Manchester, UK
stephenfry	9	46	London, UK
OMGFacts	6	45	Chicago, Illinois, USA
BBCNews	4	44	UK

Table 3.2: Most popular users according to volunteers for general timelines

then very rarely, NPMI was again used to find prominent terms in the negative class. A selection of common terms which occurred very commonly but always in uninteresting tweets is shown in Table 3.3. These terms to address a variety of topics, including status updates and spam. Specifically, the user ‘@_o_marielle_o_’ has since been banned by Twitter for spam. Several brand names make an appearance, including ‘Chanel’ and ‘@teapigs’, which are attributable to spam.

Two British TV shows are represented in this list with the terms ‘@joeyessex_’ and ‘e4chelsea’, as well as sporting and political events.

Many of the topics are short lived or relate to current political or sporting events, a static list of uninteresting terms would not be useful in the direct ranking of tweets, without being continuously updated.

Development, Training and Testing Sets

The gathered data is divided into development and testing sets. The development subset was then further divided into training and testing sets for selecting between machine learning methods. When evaluating on the held out test set, models were trained on the entire development set. The division of the data into these sets is shown in Table 3.4.

Term	NPMI	Count
#driving	0.0209	35
STAR EMOJI	0.0205	30
chanel	0.0205	30
@joeyessex_	0.0205	30
northern	0.0204	29
elections	0.0200	25
las	0.0198	23
#mgwv	0.0198	23
updates	0.0197	22
@e4chelsea	0.0197	22
#followtrick	0.0197	22
patients	0.0196	21
comic	0.0196	21
bong	0.0196	21
@teapigs	0.0196	21
#f1	0.0196	21
secretary	0.0195	20
relief	0.0195	20
atos	0.0195	20
@_o_marielle_o_	0.0195	20

Table 3.3: Terms which exclusively predict uninteresting tweets according to NPMI.

Tweetlists	
Total	387
Development - Training	145
Development - Testing	121
Held Out Testing	121
Users	
Total	148
May 2013	92
May 2014	56

Table 3.4: Size of annotated tweet corpus with topic filtering

3.3.3 Political Timelines

For comparison and to allow analysis on a single use case, a smaller, secondary data set was created with the help of SORA, an Austrian political analysis organisation which monitors social media during election campaigns, key political debates, and for evidence based policy. Targeting professional social media monitoring in particular, a restricted use case was developed to help understand how effectively a set of algorithms can aid in political analysis. More details about this use can be found in Section 5.1.

The general Twitter timelines data set was created by recruiting volunteers from the University of Sheffield mailing list; as such it is varied in terms of the type and demographics of users represented. Given that users are recruited in this manner, there was no direct interaction with any of the users during the study. Whilst reasons were given for which volunteers found particular tweets interesting, no demographics were collected. For any given user, only a few annotation sets have been created. The specific SORA use case will allow for more in depth interaction, including a detailed user study.

There are many differences between SORA's use of Twitter and that of the general public. They are primarily tweet consumers, not producers. They have built a timeline of tweets authored by Austrian political and journalist figures. This timeline consists only of tweets from other users; the SORA Twitter account does not interact socially nor publish any tweets.

The importance of SORA for this work is in the introduction of a controlled use case for summarisation, where that use case is one of social media monitoring to assist in political analysis. The inclusion of this use case allows analysis of how the content ranking task differs for professional users as opposed to the general case. Since SORA are monitoring the tweets of German speakers, it is also useful as a secondary language for the evaluation of term-frequency based approaches. SORA were approached for this task as they were use-case partners on the TrendMiner EC project (no.287863), on which the University of Sheffield was a technical partner.

A historical archive of 1,799,924 tweets in the SORA timeline was collected over a period of one year starting from the 1st of September 2012. A smaller subset of these tweets were annotated for topics and interestingness. The tweets were authored by nearly 2000 Twitter users and are mostly in German with few in English. Tweet volume varies over time and covers numerous topics.

By carrying out tweet ranking on the accounts of a professional Twitter user

as well as those of a general collection of collection of Sheffield-based volunteers, this thesis can highlight that the type of algorithms and features that are effective depend on the use case of the reader.

The primary annotation task is the same as that described in Section 3.3.2, though some modifications are made to the tweet retrieval and tweet filtering process.

Selecting windows

Tweets from the timeline are polled regularly, and made available for annotation via a web interface. Windows are formed by the polling process, and a window contains as many recent tweets as could be found in one request to the Twitter API, with cut-offs to prevent overlapping with previous windows. A calendar interface was used to assist in selecting time windows. This differs from the personal timeline annotation task in which only the most recent window is available.

Selecting topics

Once a time window has been selected, an interactive term cloud is generated. The term cloud is populated automatically with the named entities discovered by LODIE, a system for named entity disambiguation (Damljanovic and Bontcheva, 2012), and the annotator is instructed to click on a number of entities related to the topic. Clouds are generated from the surface forms of the entities, as they appear in the documents themselves.

This process is similar to that used for general timelines, except for the use of a named entity disambiguation pipeline in order to produce the term clouds. Using multi-word expressions and a limited vocabulary in this manner gives term clouds that are appropriate for the professional domain, including political actors and the names of companies and parties.

Selecting tweets

The process by which the tweets are selected is the same as was used in annotating personal timelines. The researchers were asked to select 8 out of 50 relevant tweets using an interface which makes it easy to review existing decisions and update them as desired.

Alexander, Marko, Christine, Sek, Sen, wJ, fck, gt, amp, de, pic, 2M, Muc, who is, Gita, I, Now, new, Gregor, georg, Top, lol ZiB, Donau, ZIB, Groe Koalition, Faymann, Stegersbach, Wien, EU, orf, Zib, de-ported, Ungarn, NR Michael Douglas, Pamela Anderson, mayer, Smiths, Faymann, Graz, Rudi, Schieder, Alexander, RT Karl, karl, ZIB, Beatrix, RT, MEP, Mursi, Jane, ha!

Table 3.5: Example queries used to filter tweet sets

#pressestunde	@julia_ortner
#zib24	trkei
@birgit_riegler	#oehwahl
#neuland	raiffeisen
hinzugefgt	@kosak_daniel
neuland	novotny
erdogan	#lampedusa
dietrich	mikl-leitner
#lindner	monika
#bures	mai
lindner	#imzentrum
#zib2	

Table 3.6: Terms from data set with high TF.IDF scores

Collected Data

Over the period of April 2013 to October 2013, analysts independently annotated a total of 62 sets of tweets. Each set contains 50 tweets, of which 8 were marked relevant. In total, the data set consists of 3100 tweets, with 496 positive and 2604 negative examples.

The queries used to create these 62 manually annotated subsets included combinations of names of politicians, parties and/or places that were involved in events of interest to the analysts which occurred during the chosen time window (e.g. ZIB, Grosse Koalition, Wien, Donau). Of the tweets marked interesting by volunteers, only 36.36% had been retweeted at least once. Table 3.5 lists some created filters.

The tweets are primarily written in German, with 92.07% guessed as ‘de’ by LangID (Lui and Baldwin, 2012). The actual portion may be higher, though inspection showed that there is certainly some portion of the discourse which takes place in English. Some of the top terms, according to TF.IDF, in this data set, are shown in Table 3.6

Author	Interesting Tweets	Total Tweets
ChristophFreina	4	27
LisiMoosmann	1	25
mabogsi	7	20
steinlechnerdan	6	18
Vilinthril	5	17
krone_at	2	17
AnChVIE	4	16
DiePressecom	1	16
WSJLA	3	15
EPichlbauer	2	13
pbern	0	13
rudifussi	3	12
orf_at	2	12
rupprECHT	0	10
FSchweitzer	4	10

Table 3.7: Most popular users according to volunteers for political timelines

The retweeting behaviour observed for this political data set is different to that for the general data set. Only 31.83% of tweets were retweeted, where relevant tweets were apparently only slightly more likely to be retweeted than irrelevant ones (36.36% vs 31.83%). This would appear to indicate that counts of retweets alone are not suitable as a gold standard for ranking.

As before, lists of terms are given that often appeared in interesting tweets (Table 3.8), and terms that often appeared in uninteresting tweets (Table 3.9), using NPMI (Bouma, 2009). The uninteresting set differs from those in the general timelines data set, in that more of them are terms one would expect to see in many kinds of discourse (‘gehen’ / to go, ‘zurück’ / return, ‘dort’ / there, ‘uhr’ / (o’)clock). There are relatively few tweets in this data, so the frequency counts will be lower than those for general timelines. The most prolific authors in this data, including counts of interesting tweets they produced, are shown in Table 3.7

3.3.4 Validation

The relevance annotations in the corpus were produced by SORA analysts, with the self-declared aim of guiding and informing political analysis. In order to determine the repeatability of these annotations, the task was repeated one year later, on a random selection of 15 tweet sets from the originally annotated dataset (the 62 sets described above). The observed agreement between the new annotations and the

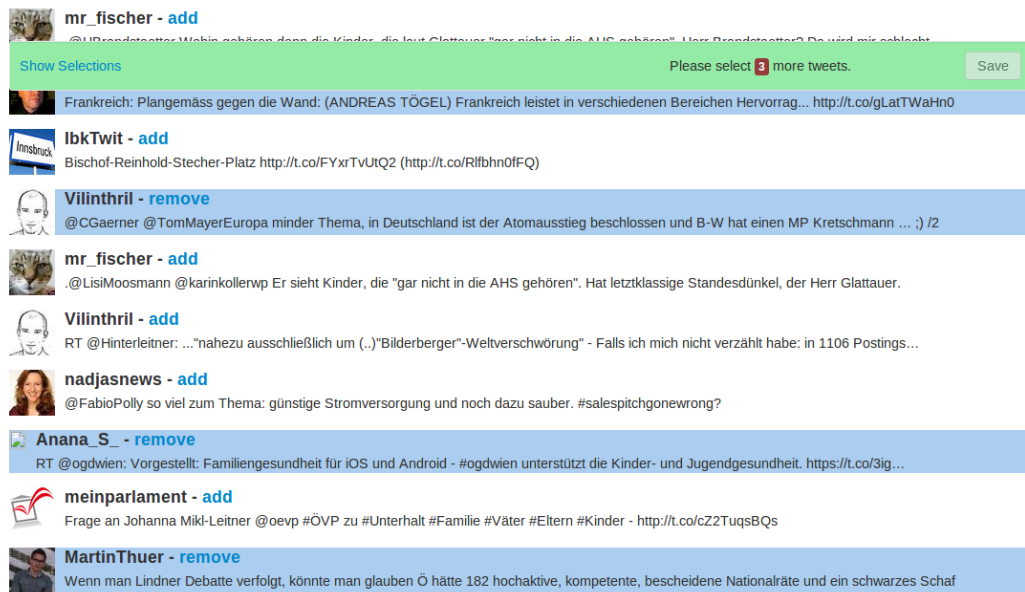
Term	Translation	Count	
		Interesting	All
award	Award	8	10
nominiert	Nomination	8	10
@gabiwaldner	–	4	4
gepflogenheiten	Customs	4	4
twitter-kampagne	Twitter campaign	4	4
DOUBLE QUAVER	4	4	4
hamann	Name	4	4
lansky	Brand/name	4	4
geschftsanbahnung	Marketing	4	4
bricht	Bright or Break	8	9
lehre	Lesson	8	9
grimme	Grimme Award	8	9
#swift	–	6	6
sechste	Sixth	8	8

Table 3.8: Terms which predict interesting tweets according to NPMI for professional data set

Term	Translation	Count	
		Uninteresting	All
—	—	—	—
@kosak_daniel	–	53	53
@wild_urb	–	28	28
vs	–	27	27
bucher	Book	26	26
uhr	(o’)clock	22	22
@youtube	–	21	21
-playlist	–	20	20
@sigi_maurer	–	20	20
hinzugefgt	Added to	20	20
#graz	Region in Austria	19	19
zurück	Return	19	19
dort	There	19	19
@vilinthrill	–	18	18
sehen	See	18	18
gab	Gave/was	16	16
istanbul	–	15	15

Table 3.9: Terms which predict uninteresting tweets according to NPMI for professional data set

original gold standard was 0.0743 (Fleiss' Kappa). The reasons for this relatively low agreement are discussed in Section 5.5.6.



- France: run into the ground (as planned): (ANDREAS TOGEL): France excels in various areas...
- Bishop Reinhold Stecher Square
- Poor subject - Germany has decided to phase out nuclear power and Baden-Württemberg has Kretschman for an MP.
- He sees children that do not belong in high school at all. Mr Glattauer is the worst kind of snob.
- Nearly exclusively about the global Bilderberg [a political conference] conspiracy. There are 1106 posts, if I count correctly.
- So much for that: a convenient energy supply, and a clean one at that. #salespitchgonewrong
- Presenting: Family health for iOS and Android, #ogdwien supports children's and young people's health
- Question for Johanna Mikl-Leitner #OVP about #earningaliving #fathers #parents #children
- If one follows the Lindner debate, one could think that Austria has 182 highly active, competent and modest national representatives, and one black sheep.

Figure 3.5: Interface used to annotate interesting posts in final study. Selected tweets are highlighted (with translations)

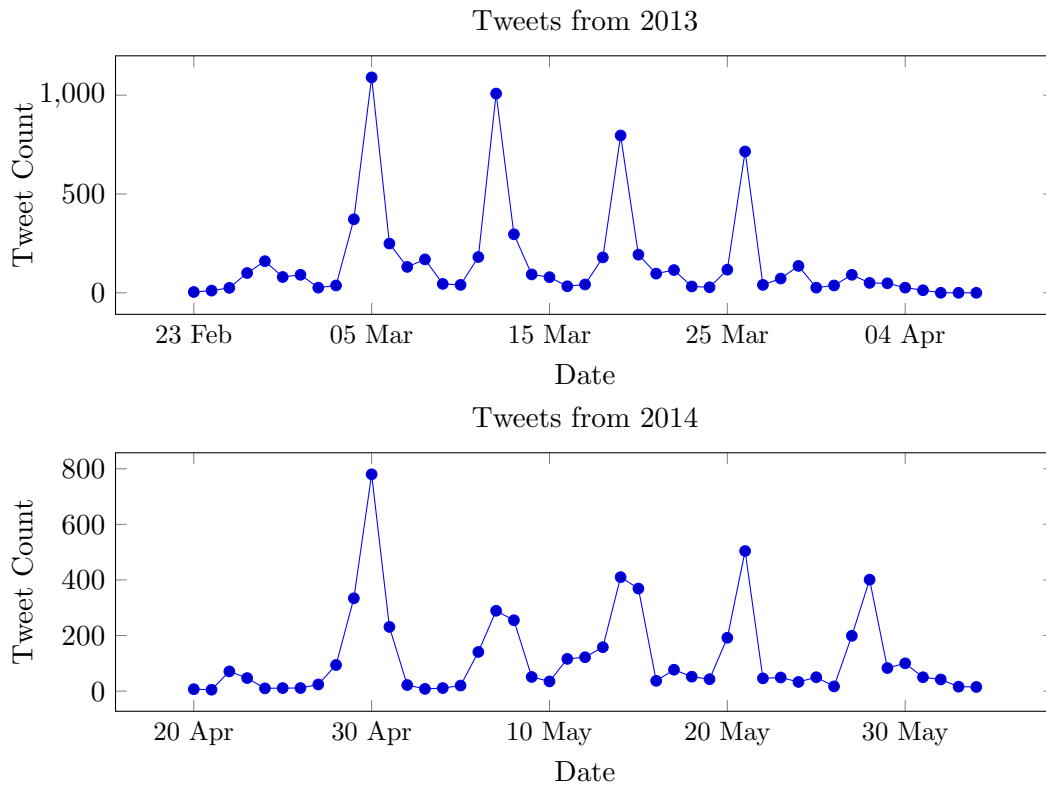


Figure 3.6: Count of tweets in the gold standard by date, in 2013 and 2014 respectively.

Available data sets

Clicking a range will immediately start a new annotation task based on tweets in that range.

Oct 13 — 19 2013

	Sun 10/13	Mon 10/14	Tue 10/15	Wed 10/16	Thu 10/17	Fri 10/18	Sat 10/19
all-day							
12am	11:55 - 1:45 TMSORA12	11:55 - 1:50 TMSORA12	11:58 - 1:20 TMSORA12	11:56 - 1:54 TMSORA12	11:55 - 1:49 TMSORA12	11:55 - 1:51 TMSORA12	11:56 - 1:55 TMSORA12
1am							
2am	1:56 - 3:53 TMSORA12	2:01 - 3:52 TMSORA12	1:59 - 3:55 TMSORA12	1:58 - 3:54 TMSORA12	1:58 - 3:54 TMSORA12	2:03 - 3:55 TMSORA12	1:58 - 3:54 TMSORA12
3am							
4am	3:59 - 5:55 TMSORA12	4:00 - 5:55 TMSORA12	3:57 - 5:54 TMSORA12	4:01 - 5:55 TMSORA12	3:56 - 5:54 TMSORA12	3:59 - 5:55 TMSORA12	3:56 - 5:55 TMSORA12
5am							
6am	5:56 - 7:55 TMSORA12		5:56 - 7:55 TMSORA12	5:56 - 7:55 TMSORA12	5:56 - 7:55 TMSORA12	6:15 - 7:55 TMSORA12	5:55 - 7:55 TMSORA12
7am							
8am	7:56 - 9:55 TMSORA12	7:39 - 9:55 TMSORA12	8:18 - 9:55	8:15 - 9:55	8:21 - 9:55	8:22 - 9:55	7:55 - 9:55 TMSORA12

Figure 3.7: Choosing a time window to annotate

3.4 Reasons for Relevance

Since one of the goals of this work is to identify and classify the different reasons for which tweets are interesting, more detailed rationale information was also gathered from the volunteers. The first version of the study asked users to supply optional ‘Other Comments’ alongside their interestingness judgements. This was well received and was used by the volunteers to explain their decisions, with roughly 5.8% of the judgements being accompanied by comments.

Stated reasons included agreement with the content, liking the products and services mentioned, being a fan of the author (e.g. a celebrity) or knowing them personally. A manual analysis of these textual responses was carried out, and they were generalised into a classification, describing different types of tweet relevance.

After generalising the responses from the first version of the study, 7 salience reasons were determined. These were validated by asking 6 of volunteers to apply them to a set of 100 responses. The task was difficult because the users had to guess the intent of the original volunteer’s judgement from what were often unclear comments. The original tweet was also supplied though in many cases it was difficult to interpret without context.

An ‘other’ category was shown during annotation and was used in cases where it was unclear what the author meant by their comment.

A Fleiss’ Kappa score of 0.53 was calculated for the volunteer responses, indicating moderate agreement. These labels were then used in the final version of the study, allowing annotators to select from them when giving reasons for the relevance of their selections.

The labels used and the proportion of comments assigned to that label (averaged across annotators) are shown in Table 3.10. The table shows the distribution of classes created by annotators for the pilot study, and the observed distribution from the final version of the study.

The distribution of these classes show that tweets are interesting for a variety of different reasons, and that readers have varied motivations. No one class dominates the others.

The more common, broader classes do not see much variation between the counts in validation study and the counts when users were asked directly, though some of the rarer classes did increase, perhaps due to a reduction in the number of ‘other’ tweets and reflecting the greater knowledge of the annotator. The ‘addresses friend’ class did not apply to any of the tweets in the second version of the study.

Comments	Pilot	Final
Relevant to work/location/interests	29%	29%
Funny/Clever observation	20%	20%
Contains interesting or surprising information	18.7%	20%
Mentions something they like	6%	10%
Sympathise/agree with the sentiment	6%	9%
They like or trust the author	4%	8%
Addresses reader or a friend directly	4%	0%
Other	12.3%	4%

Table 3.10: Distribution of reasons for which interesting tweets were chosen

In the final study, 40% of tweets belonged to classes that might be considered to be universal (Humour, Surprise), with 56% of interestingness classes not only subjective, but are also specific to each timeline owner, indicating that an effective summary of such timelines would need to be personalised to the reader.

These classifications for reasons for relevance are not the same as classifications for the tweets themselves, such as those seen in (Naaman et al., 2010). They do not make sense for tweets that are uninteresting, as they only attempt to explain positive judgements. Moreover, what they seek to capture is a property of the user’s attitude towards the tweets, not the tweet itself.

3.5 Retweet Limitations

The gold standard data set gathered in this thesis will allow for the comparison of ranking algorithms for Twitter home timelines. In other work, however, no such gold standard was created, and instead retweets were used as a ground truth for tweet relevance (Uysal and Croft, 2011; Hong et al., 2013; Feng and Wang, 2013). The act of retweeting requires that the reader has engaged with the content, lending weight to the idea that a retweet is an expression of interest.

The use of retweets as a gold standard implies the assumption that all interesting tweets are retweeted. There are many reasons why this may not be the case, however. Retweeting is a public act, visible to all of the retweeting user’s followers as well as the original author of the tweet. This public approval would not be appropriate for private conversations, or for personal status updates.

Home timelines are personal to the reader, and contain private tweets (which cannot be retweeted) as well as tweets which are unlikely to be of interest to all followers, such as personal updates (Naaman et al., 2010). Some users feel that

authors retweet too frequently, engaging in attention seeking behaviour (Boyd et al., 2010), creating an expectation that counts of retweets should be moderated. This sparsity of retweets means that users are unlikely to retweet every interesting tweet on a particular topic.

Similar to retweeting is the “favourite” feature on Twitter, which expresses approval of the tweet and notifies the author, but does not share the tweet on the reader’s timeline. None of the existing work on home timeline ranking has used favourites as a ground truth for ranking. Gorrell et al. (2014) found that tweets are favourited for a number of different reasons, including to bookmark them, as thanks for mentioning them, to express that they like the post, as a dialogue act in a conversation (a wordless form of response), and for self-promotion (as authors are notified when their tweets are favourited).

As with retweets, the act of favouriting has side effects beyond just expressing interest in a post. Favourite posts are compiled in a list for a user to refer back to, so those that use this feature as a form of bookmarking may use it sparingly. When a post is favourited, both the author and other readers can see a list of who favourited it, so many may choose not to do so for privacy reasons, or because of the nature of their relationship with the author.

A study of reasons for retweeting was carried out by Boyd et al. (2010). Likewise, a study of reasons for relevance stated by users was given in Section 3.4. Of the reasons given by users, only ‘Sympathise/agree with the sentiment’ was seen in the list of stated reasons for retweets.

The gold standard data can also be used to empirically demonstrate the limitations of retweets as a ground truth for relevance. Several issues with using retweets in this manner are revealed by the data:

Retweet counts are less reliable for very new tweets: This is a variant of the cold start problem faced by recommender systems (Schein et al., 2002). Given that no user has yet endorsed a given tweet, it can be assumed either that the tweet is not interesting, or that it has not existed for long enough to have garnered any attention.

Figure 3.8 shows the distribution of retweets and favourites against the age of the tweet in the development portion of the user-selected gold standard. Both tweet age in seconds (at the time of annotation) and counts are shown on a logarithmic scale as the data encompasses tweets of wide ranging ages and popularities. This data clearly demonstrates that retweet counts for new tweets are much lower than those for older tweets.

The relation between tweet age and interestingness was also investigated, and

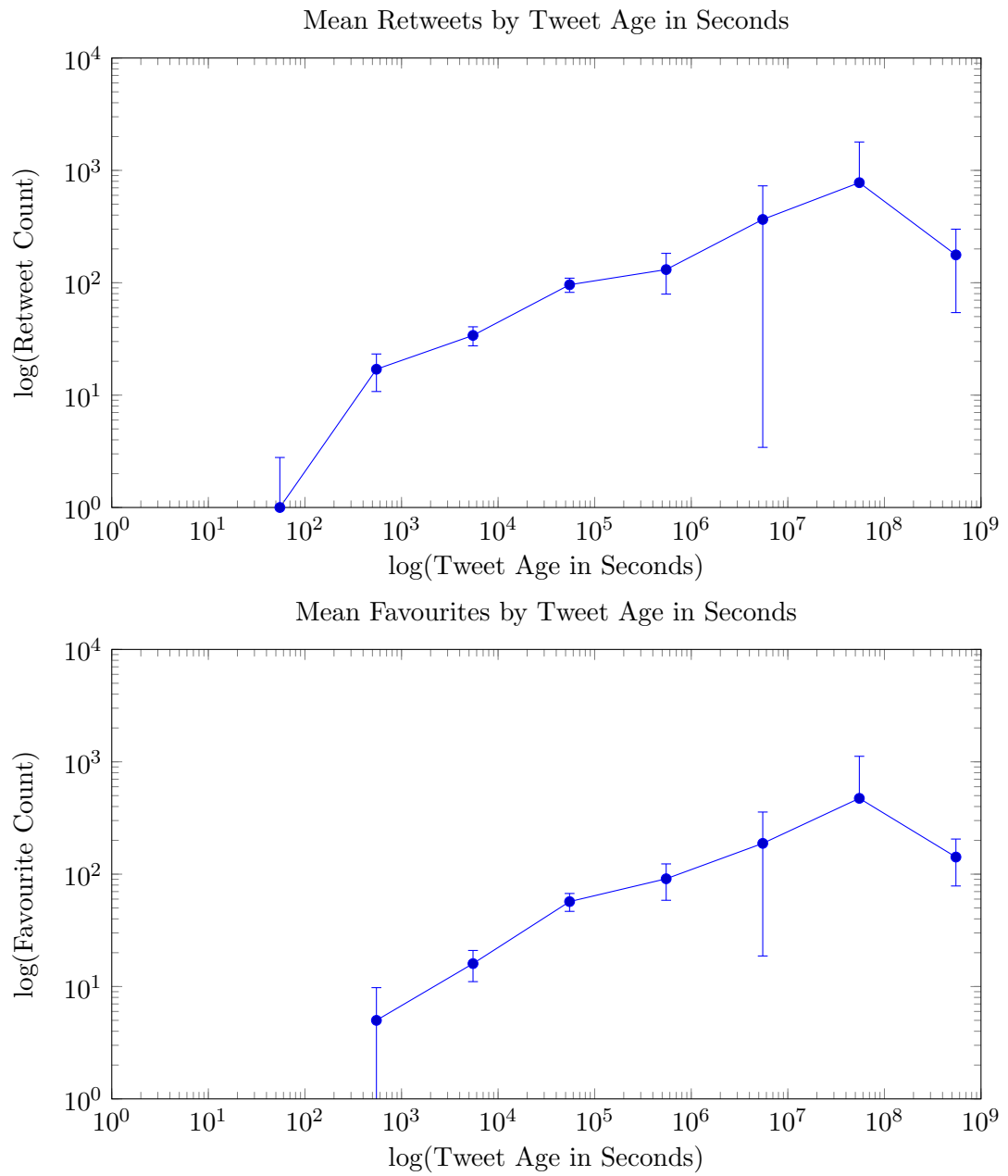


Figure 3.8: Distribution of retweet and favourite counts by age, showing 95% confidence intervals assuming normal distribution

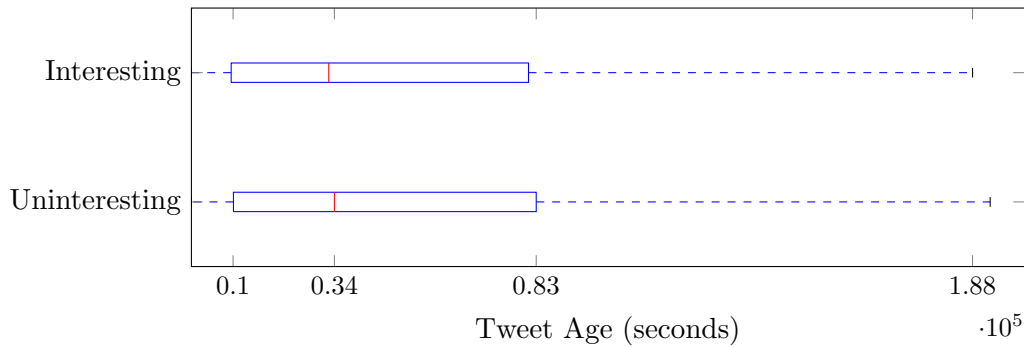


Figure 3.9: Distribution of tweet age in seconds for interesting and uninteresting examples (outliers not shown)

no strong preference for more recent tweets was observed. The age in minutes of interesting tweets and the age in minutes of the uninteresting ones is compared in Figure 3.9. Outliers are not shown, as tweet sets would occasionally contain retweets of very old tweets, making visual comparison difficult. The distributions are broadly identical, demonstrating immediately that there is not a direct connection between the age of the tweet and whether it was found interesting by the respective timeline owner, unlike the correlation seen for retweets and tweet age.

These two observations make the cold start problem an important issue to consider, when relying exclusively on retweets as the oracle indicator of tweet relevance.

Retweets are biased towards certain kinds of tweets: This was demonstrated using a crowd sourced classification process to understand the types of tweets that users marked relevant. It would not be possible to properly carry out this analysis based on the self-reported ‘reasons for relevance’ (Section 3.4) for tweets in the gold standard, as these reasons do not capture the kind of tweet itself nor do they apply to tweets that are not interesting.

A sample of 599 tweets were taken from the gold standard data set (both interesting and uninteresting tweets), and annotated using the CrowdFlower service. The tweet classes used for this task are shown in Table 3.11. These classes were loosely based on those proposed by Naaman et al. (2010), with modifications to the classes to assist in crowd sourcing by untrained annotators, and pairs of rarer classes merged. Classes were not assumed to be exclusive, so tweets could be both humorous and opinionated, for example.

Table 3.11 reports three statistics: the overall class distribution by type for tweets in the sample; for tweets selected as interesting by one of the volunteer home

Tweet Class	Overall	User-selected	Retweeted
Information sharing (tips, news etc)	72.8%	57.1%	77.1%
Opinions	38.9%	51.0%	37.4%
Story, joke, or humour	16.9%	26.5%	14.4%
Update about author	17.2%	24.5%	14.0%
Message to a friend	17.0%	9.2%	14.4%
Question to followers	6.7%	7.1%	7.3%

Table 3.11: Distribution of Tweet Classes. Tweet classes are not exclusive.

timeline owners; and for tweets that have been retweeted at least once. Distributions based on retweets by the volunteers themselves were not reported, as only one tweet in the sample had been retweeted by a volunteer.

As can be seen in the table, the distribution for retweets alone differs from that of interesting tweets. The home timeline owners selected more status updates, opinions and anecdotes/jokes from their timeline as interesting, whereas information sharing and messaging tweets were more likely to be retweeted. These results indicate that a user-reported gold standard (which reflects more accurately what timeline owners consider interesting) differs substantially from a purely retweet-based evaluation dataset, and thus, the use of the latter as gold standard should be avoided, when possible.

Many interesting tweets are never retweeted by anyone: Tweets that are of interest to an individual do not necessarily garner attention from Twitter in general, and 34% of the interesting tweets in the gold standard were not retweeted by anyone at all, and only 0.1% were retweeted by the participant themselves.

3.6 Discussion

This chapter introduced a specific version of the problem of Twitter timelines summarisation. By defining the problem as one of ranking, a clear framework is provided for automatic evaluation, whilst leaving the possibility for other types of summarisation to be targeted in the future. The evaluation framework that will be used this work has been discussed, including two gold standard data sets with which to automatically evaluate a large number of algorithms.

The need for a user-created gold standard for tweet ranking of Twitter home timelines, as opposed to the use of retweet counts as a proxy for relevance, was argued in Section 3.5.

The annotation process used in this chapter is subjective, as is the task of tweet

ranking itself (Alonso et al., 2013). Rather than attempting to transform a deeply subjective task into an objective one, a large cohort of 148 Twitter users were instructed to annotate 387 samples of their own timelines (Section 3.3). The use of a large number of annotated tweet sets, including multiple from each user, will help to overcome the subjectivity of the task itself.

The task of identifying interesting tweets is defined in this work in subjective terms - the interesting tweets are those that appeal to the specific readers involved in our study. This does not imply that the task is ill-defined, rather that it is defined in broad terms. This is reinforced through the introduction of a specific use case for the general task, and a study of the reasons for which certain tweets may be found relevant by readers.

The specific use case for the identification of interesting tweets is one of assisting professional analysis. There is a clear need in this case to discover tweets of relevance to a specific topic automatically in order to inform and guide further experiments. On the general case, a number of reasons for which tweets can be found relevant were coded and validated in Section 3.4. These reasons further identify the kinds of decisions that must be made as part of the task (discovering humorous tweets, tweets with social relevance, tweets with topical relevance etc).

The task is very general in terms of the types of tweets that are to be identified, but is otherwise restricted in terms of the ratio of interesting to uninteresting tweets, and the size of the input set. The input set size is determined by technical limitations - it is the number of tweets that Twitter will return through their API. On the other hand, the ratio of interesting to uninteresting tweets was determined empirically through a pilot study.

Chapter 4

Methods and Representations

4.1 Introduction

This thesis aims to understand the type of information and the kinds of algorithms which are effective in ranking tweets in Twitter home timelines. A restricted version of this task was proposed, and data sets produced to reflect both a specialist professional consumer of tweets, and a general population of users (See Chapter 3). Throughout this thesis, the differences between these two experiments are highlighted, applying different features to each as appropriate.

Nonetheless, there are methods and text representations which are used to answer both research questions. This chapter gives an overview of these methods.

In the context of the social media monitoring use case, social connections and past posts are not exploited, because the target users did not create any. These social connections are, however, available and exploited in the more general home timeline summarisation task. The methods which are common to both areas of work are those which treat tweet ranking as a variant of text summarisation or information retrieval. They rank tweets according to the text they contain. Where features that are specific to social media are described here, they are those which do not depend on the reader, such as retweet counts and counts of favourites.

This chapter is structured as follows: a general description of the approach to tweet summarisation is given, and the evaluation metrics which will be used to compare methods are defined. A series of appropriate baseline methods are then presented, which will be used in this work to provide comparison to candidate rankings, and multiple text based methods will be presented for summarisation and content ranking, explaining how they will be applied to tweets in particular. The

ways text is represented in tweets in order to apply these methods will also be described.

4.2 Approach

This section describes the approach taken in this thesis to the task described in Chapter 3.

Key to this work is the notion of ordering tweets according to interestingness and relevance scores. Some of the algorithms used here score tweets according to a specific kind of relevance, for example relevance to the filter terms, or relevance to the reader. Others, including those which simply reflect the popularity of the tweet, can be thought of as universal interestingness scores. No attempt was made in the data collection process to differentiate these types of scores. Users were not prompted to specifically find the most representative tweets, nor those which are most interesting.

Relevance scoring in this thesis falls into several categories.

- Baseline scores based on the tweet metadata, such as its popularity.
- Text-based comparison of tweets to the terms in the query or the collection to which they belong.
- Scores which reflect the popularity of the author who posted the tweet.
- Models of the relationship between the author and reader.
- Text-based comparison of tweets to the previous posts by the reader.

The baseline scores are used in all areas of this thesis, both for the social media monitoring use case and in the case of general Twitter users. These will be described in Section 4.3. The text based algorithms, which predict the relevance of the tweet to the query or to the collection as a whole, are also applied to both data sets. These scores address the first hypothesis of this thesis, which is that summarisation and information retrieval methods can be applied to Twitter home timelines (Section 1.2.2).

Other features in this work rely on the social nature of Twitter users. These are personalised to the reader, and as such can be thought of as reader relevance scores. These social features and algorithms reflect the second thesis hypothesis,

which states that personalisation can lead to improved ranking performance (Section 1.2.4). SORA did not use Twitter socially, nor did they create their own posts, so these scores are applied only to the general timelines dataset, and as such will be introduced in Chapter 6.

After being assigned scores, tweets are then sorted by that score in order to form the summary. Where a single feature is used for ranking, such as in the baseline methods, tweets are ordered by the value of that feature. When ranking by retweet count, for example, tweets are sorted descending by those counts.

Once tweets have been ranked according to relevance, a tweet selection method is applied in order to form the final summary. This work considers the following methods:

- The tweets are simply reordered according to their relevance scores.
- The top 8 most relevant tweets are taken to form the summary.
- Tweets are ordered according to a weighted score of relevance and diversity.

The decision of whether or not to truncate the summary depends on the evaluation method. For some automatic evaluation methods, and for manual evaluation, it is necessary to produce an actual summary of a limited size as the candidate. When the score under consideration is one which scores ranking quality, the entire set may be supplied.

4.3 Baseline Methods

For purposes of comparison, and to help gain an understanding both of the problem and the performance of methods that address that problem, this thesis implements and evaluates a number of baselines.

4.3.1 Random

The simplest baseline is random ordering. A random relevance score between 0 and 1 is assigned to every tweet in the set. This results in a random ordering of the tweets. The use of a randomised baseline is not always appropriate, however in this data set the ‘interesting’ class is much smaller than the ‘uninteresting’ class (8 out of 50), so it is entirely possible to devise approaches which perform worse than random. Since tweet sets are already filtered, random ordering forms a strong baseline (Mackie et al., 2014).

4.3.2 Tweet Meta Data Baselines

Counts of retweets and counts of favourites are used as simple baselines for relevance, as in related work (Duan et al., 2010; Uysal and Croft, 2011). Neither retweet counts nor favourite counts are sensitive to the context in which summarisation is taking place, because they are global recommendations, and as such can be applied in all parts of this work.

In the pilot data set (Section 3.3.1), users were far more likely to mark very recent tweets as relevant. This may have been caused by the decision to display recent tweets higher up in the user interface, so in later experiments, posts were shown in a random order. In light of this, a reverse chronological baseline is added, in which tweets are ordered by their age, rather than by any relevance measure.

The reverse chronological baseline re-establishes the tweet order as would be shown by Twitter in the web interface, and as such represents the status quo. Strong performance here would indicate a preference for newer tweets, validating the results seen in the pilot study.

4.3.3 Text Baselines

The cosine similarity between the terms in a document set’s filter (the words which were chosen from a term cloud to generate the set initially) and a candidate tweet is used as a baseline. This is a simplistic measure from information retrieval which is commonly used as part of the vector space retrieval model (Manning et al., 2008).

Since the tweets are drawn according to whether or not they contain terms from the corresponding filters, and the filters contain many terms, in practise the cosine baseline may score similarly to other algorithms which don’t rely on the query but instead rely on the most common terms in the document set.

4.4 Comparison to Related Work

When seeking related work for comparison, there were a number of key issues which limited those which could be selected. Though many models have appeared since the start of this work to address related tasks, at the time of the experiments there were few which specifically addressed the task of ranking from home timelines. The nature of related tasks means that many of the models could not easily be translated to this domain, for example retweet prediction work which assumes that tweets are public and can be retweeted at all.

The most closely related work was carried out using complex graph-based models (Feng and Wang, 2013; Hong et al., 2013), or learning to rank approaches (Uysal and Croft, 2011) which would have created technical problems for reimplementation and which would have dramatically expanded the scope of this work, due to the need to implement bespoke algorithms and train unusual models. In the case of the graph factorisation work, these approaches were also published two years after the start of this thesis work, when their inclusion as benchmarks for comparison would have represented a significant expansion.

The Hybrid TF.IDF method of Inouye and Kalita (2011) is, however, re-implemented here for comparison to related work. Hybrid TF.IDF was proposed as a method for ranking tweets in response to search queries, in which one tweet is considered a document for the purpose of estimating IDF, but TF scores are considered at the scope of the document. This approach characterises how significant the terms in the tweet appear to be and then normalises that score by the length of the tweet.

Importantly, Hybrid TF.IDF not only expands the use of TF.IDF for ranking in tweet summaries, but also includes a threshold for diversity. Tweets are not prioritised for inclusion in the summary if they are very similar to existing tweets in the summary. The threshold is determined in this thesis using grid search.

4.5 Text Summarisation

The algorithms in this section are text-based, and are derived solely from the content of tweets in the topic-focussed collections.

Given the lack of position information in tweets, the methods tested from extractive text summarisation are those which use term frequency. TextRank and Centroid attempt to find the most characteristic parts of a document, which are considered to be those which most closely reflect the vocabulary of the document as a whole.

4.5.1 Centroid

The Centroid algorithm prioritises text units from a document that are most representative of the document as a whole, according to counts of terms (Radev et al., 2004). The algorithm as defined originally was used to select sentences, though tweets are selected using Centroid in this work. They are turned into bag-of-words representations, or vectors, with one vector per tweet and one dimension per word.

The use of tweets rather than sentences is not a significant deviation from the original method since tweets are not usually much longer than a sentence would be in extended prose.

When applying the centroid algorithm, the document is converted, unit by unit, into a series of counts of individual words. A brief example follows:

	Got	go	to	Sheffield
Got to go to Sheffield	1	1	2	1
To Go	0	1	1	0
To Sheffield	0	0	1	1
Got to go	1	1	1	0

The score for each term is summed for the entire document. Where no weighting is used, this gives the number of occurrences for each term in the document, known as the term frequency.

Got	go	to	Sheffield
2	3	5	2

The word scores are then divided by the number of text units, to give a centroid, which is an average text unit to be used as a prototype for an ideal summary of the entire document. For each word in the centroid, the calculated score is the mean number of times that appears in sentences in the document.

Got	go	to	Sheffield
0.5	0.75	1.25	0.5

Since the centroid is just a series of weightings over terms, it does not form a legible summary on its own. The individual sentences are therefore scored according to their textual similarity to the centroid, and those which score highest are used in the summary.

The similarity measure used in this thesis for comparison to centroid is cosine similarity. Centroid will be evaluated with and without stoplists, IDF weighting and case information.

4.5.2 TextRank

For TextRank summarisation (Mihalcea and Tarau, 2004), tweets are also converted into vector representations, which are compared for similarity. The authors of early

work on TextRank define their own text similarity scores, but cosine similarity is used here for consistency with Centroid and other approaches.

A graph is drawn in which each sentence is represented by a node. Edges are added to the graph between pairs of sentences where the similarity between the two is above a threshold. In some variants, the edges are also weighted by the similarity score.

The PageRank algorithm (Brin and Page, 1998) is then applied to the constructed graph. PageRank begins by assigning random scores to all nodes, and then updating those scores iteratively until they represent the likelihood that a random walk from any other point on the graph will arrive at that node. The final PageRank scores are a form of betweenness measure. When PageRank is used in the context of web search, an influential page may be linked to by many others, and that page's outgoing links contribute more to others.

Whilst one sentence does not explicitly endorse another in the same way that pages do, TextRank uses the PageRank algorithm to instead score sentences that have more in common with the others in the collection. As with Centroid, it is possible to use a variety of vector representations. The cosine similarity is again used for text comparison, and methods are evaluated with and without stoplists, case information and IDF weighting.

4.5.3 Sum Term Frequency (TF)

In this algorithm, each term in the document is counted to give the term frequency (TF) (Manning et al., 2008). Then, for each candidate text unit, the term frequencies of all of its words are summed to give an aggregate score. Tweets are then ordered by this score. IDF can be used to provide additional weight to rarer terms. For a word w in document D , term frequency is given by $tf(w) = \text{count}(w, D)$ and the TF weighting for a sentence W by:

$$\text{TF}(W) = \sum_{w \in W} tf(w) \quad (4.1)$$

This algorithm is nearly identical to Centroid, except that the scores are not normalised by the length of the collection nor the length of the tweet. The sum of TF scores will select long text units over short ones. As with Centroid ranking and TextRank, sum term frequencies are calculated on various representations, including n-grams, idf weighting, stoplists and text case information.

4.5.4 BM25

Okapi BM25 provides a ranking function in which to weight various features including term frequency and document length as part a probabilistic retrieval framework (Robertson et al., 1995). There are two parameters to BM25, which are k , a weight for term frequency and b , a weight for document length. This is in contrast to Centroid where unit length is normalised and to TF scoring where the longer units are preferred. For BM25 this weighting can be set empirically to give the strongest results.

There are several variants of the Okapi BM25 scoring function. In this work the slightly simplified formulation of BM25 intended for short queries is used. The BM25 scoring function is similar to the baseline Cosine distance to query score, in that it considers only terms from the query. Important differences include the probabilistic nature of BM25, and the ability to manually weight term frequency and query length.

The following formula for BM25 is used, where tf_{td} is the term frequency (number of occurrences) of term t in document d , L_d is the length in tokens of document d and L_{avd} is the average length of documents in the entire collection.

$$\text{BM25}_d = \sum_{t \in Q} \text{IDF}(t) \cdot \frac{(k+1) \times \text{tf}_{td}}{k((1-b) + b \times (L_d/L_{\text{avd}})) + \text{tf}_{td}} \quad (4.2)$$

When k is set to 0.0, the scores returned by BM25 are just the sum of IDF values for words in the query set, which is formed from the filter generated by the reader. This score is similar to the baseline for the IDF weighted cosine similarity to the query, with the exception that it does not normalise tweets for length.

$$\text{BM25}_d = \sum_{t \in Q: t \in d} \text{IDF}(t) \quad (4.3)$$

Unlike Centroid, TextRank and TF scoring which are summarisation methods, BM25 is an information retrieval technique. Units are ranked according to their similarity to the query, rather than their representativeness of the document as a whole.

BM25 is applied in this work with the filter terms which were generated by the users during annotation. The representation used for BM25 is the lower case IDF weighted term vector with stopwords removed.

4.5.5 Maximal Marginal Relevance

Tweets are prone to repetition thanks to retweets. In the methods above, where tweets are ranked according to their similarity to a collection of terms, nearly identical tweets will receive the exact same score and appear next to one another in the rankings, leading to the inclusion of redundant tweets in the generated summaries.

Maximum Marginal Relevance (Carbonell and Goldstein, 1998) is used to reduce this redundancy. MMR is a greedy approach to ranking in which new posts added to a summary take into account the existing posts. A hyper-parameter λ is set to a value which weights the selection between a relevance score (in this case, centroid), and a score which is the inverse of similarity to the documents already in the set. As such, MMR gives a combination of some ranking score and a measure of diversity and can be thought of as a method of summary generation.

Tweets were selected according to the maximum of the following formula, where t is the tweet, $\text{rel}(t)$ is the relevance of the tweet to the document (cosine similarity to centroid), and $\text{sim}(t)$ is the maximum cosine similarity of the tweet to any already in the summary:

$$\lambda \cdot \text{rel}(t) - (1 - \lambda) \cdot \text{sim}(t) \quad (4.4)$$

When the hyper-parameter $\lambda = 0$, the summary generated is maximally diverse according to the distance measure used. At $\lambda = 1$, MMR is simply equivalent to whichever relevance measure was used, which in this thesis is the Centroid.

Maximal Marginal Relevance has the disadvantage of being a greedy approach where the decision made at each stage depends on the previous state of the algorithm. Because MMR has a hyper-parameter, λ , the method requires tuning to select the ideal value for this parameter. In this thesis λ is tuned using grid search. The inclusion of such parameters requires extra computation during training, and can reduce the generalisability of the result, so this can be considered a disadvantage over centroid alone, which has no parameters to tune.

MMR will be applied using the cosine similarity measure for text. For political timelines, MMR will be calculated using IDF-weighted unigrams with stopwords removed. For general timelines, MMR will also be evaluated for unweighted unigrams with and without stopwords, bigrams and trigrams.

4.6 Representations

The text scoring methods Centroid, TextRank and MMR, described in Section 4.5 require numerical representations of tweets. These representations consist of a mapping of terms, (or sequences of terms), onto counts of the number of times they appear in the tweet.

This mapping of terms onto frequency is viewed as a term vector (Manning et al., 2008), where each possible term in the collection is indexed sequentially to form a new dimension of the vector. A corpus with n unique words would then be represented by a vector with n dimensions. The method used in this work to compare term vectors is cosine similarity, given for two term vectors \vec{T}_1 and \vec{T}_2 by the following:

$$\text{similarity}(\vec{T}_1, \vec{T}_2) = \frac{\vec{T}_1 \cdot \vec{T}_2}{|\vec{T}_1||\vec{T}_2|} \quad (4.5)$$

Counts of single words, or tokens, are known as unigram frequencies. Term vectors can be formed from unigram frequencies, or from frequency counts of sequential pairs of tokens (bigrams) or sequential triplets of tokens (trigrams). Each bigram or trigram has a unique index in the vector representation. Unigrams cannot differentiate, for example, between “Please RT” and just “RT”, longer n-grams can be used to count the two separately, capturing dependencies between terms which can be useful in text comparison or language modelling (Manning and Schütze, 1999).

As sequences grow longer, however, the likelihood that they are seen in a given corpus grows smaller. A back-off model can be used in language modelling to estimate likelihood of N-grams given the shorter sequences they contain, allowing for unseen sequences to be assigned a non-zero likelihood (Manning and Schütze, 1999). However, these probabilistic models are not appropriate when a non-probabilistic method such as Centroid or TextRank is to be used for ranking, as in this work.

With the use of bigram and trigrams, the start and end of a sentence may be marked by special ‘START’ and ‘STOP’ tokens, such that a representation of the sentence “Bigrams can be useful” would contain the bigrams ‘START, bigrams’ and ‘useful, STOP’. The start and end of sequences is marked in this manner in all experiments in this thesis.

The upper and lower-case variants of words can often have the same referent, but at different points in a sentence or typed differently by a hurried user, words can be converted to their lower or upper case representation automatically, as a

preprocessing step.

In this thesis, n-gram representations are formed from lower case variants of terms, using ‘START’ and ‘STOP’ delimiters.

4.6.1 Stoplist

Centroid and sum TF ranking assume that the most common terms in the set tweets represent the overall topic under discussion. This is not true if the common words unrelated to the theme or content of the document. Problematic, but common, terms include ‘stopwords’, which are words which do not relate to the subject of the document, such as conjunctions (‘and’, ‘but’) and pronouns (‘she’, ‘they’) and which are likely to appear in any tweet, regardless of the topic.

When calculating Centroid with stop words left in, for example, auxiliaries and determiners which appear in every single text unit will dominate the centroid, and the units which are selected will be those which contain greatest ratio of ‘meaningless’ tokens. This can be prevented by removing stop words prior to further analysis (Manning et al., 2008).

The majority of these terms belong to closed classes, and as such there is a small enough quantity of them to list, so that they may be removed automatically. Such a list is known as a stoplist.

The set of stop words is generally consistent within a language regardless of domain. There are conventions in Twitter orthography that lead to additional stop words beyond those that occur in other examples of English. Examples include ‘RT’, which signifies that the post is a retweet, and contractions such as ‘gonna’, as well as punctuation chains such as ‘!!!’.

Stoplists are applied everywhere unigrams are used in the work on summarising political timelines. In the experiments on general timeline summarisation, stoplists are applied where specified.

4.6.2 Inverse Document Frequency

When term counts are used directly, it is implicitly assumed that every term in the vector has the same discriminative value when determining for example term similarity. This is often not the case because tweets not only contain stop words, which are removed using stoplists, but also vocabulary which, despite being meaningful, one would expect to see universally regardless of the topic at hand. Examples might include ‘minister’ in a political corpus or ‘patient’ in a medical corpus.

A background corpus can be used to discover such terms and to weigh them less heavily than very unusual words when using term frequency vectors. By doing so, rarer terms are highlighted and assumed to be more ‘meaningful’ where they do occur. Terms may be weighted in this way using Inverted Document Frequency, abbreviated as IDF (Manning et al., 2008).

One classic formulation for IDF is as follows:

$$\text{IDF}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (4.6)$$

Where t is a token and D is the collection of documents in the index. In some cases, terms in the target document are not contained in the index itself, leading to division by 0 (because the document frequency is 0). IDF in this form is not defined for terms which hadn’t been seen when the index was constructed, but where $\exists d \in D : t \in d$. Smoothing for IDF, to deal with the other cases will be discussed in Section 6.2.2

The term frequency of a token in the document (the number of times it occurs) may be multiplied by the inverse document frequency for that term in order to arrive at a final, weighted score to be used in the document vector. These scores, known as TF.IDF can then be used in place of the frequency vector and the algorithm may be applied using this vector in place of the term counts.

IDF indices are used where specified in all text ranking methods. IDF indices are not used with bigrams or trigrams. IDF indices are always accompanied by stoplists in this work.

4.6.3 Dimensionality Reduction

Frequency based approaches to tweet recommendation are likely to suffer from issues related to sparseness thanks to the constantly shifting vocabulary of online broadcasts, where different entities, slang, emoticons and hash tags enter common usage over time. Likewise, misspellings are extremely common and can cause problems where exact matching is required (Ritter et al., 2011).

While there may be a single, prominent topic in a collection, it is possible that different terms are used to refer to the same entities. This is especially plausible in tweets, where hashtags, long names or abbreviations can be used to refer to the same organisations or people or events. Vocabulary mismatch can be addressed through dimensionality reduction with topic models (Dumais, 1991).

The method used in this thesis method to reduce vocabulary mismatch is to

train the topic models latent semantic indexes (LSI) (Deerwester et al., 1990) and latent Dirichlet allocation (LDA) (Blei et al., 2003) on a large background corpus of suitable tweets. Topic models are applied to tweets, giving a distribution of probabilities for each topic. This distribution is then used as a feature vector and as the input for centroid and MMR, giving a representation with fewer dimensions and less opportunity for vocabulary mismatch.

Latent Semantic Indexes

Latent Semantic Indexes are an older form of dimensionality reduction which rely on matrix representations in order to reduce the feature space of vectors generated from documents (Deerwester et al., 1990). There are multiple ways of preparing the term vectors, but in the work in this thesis they are formed from term frequencies; alternatively, they could have been represented on a binary or logarithmic scale.

A matrix X is formed with a column for each term and a row for each document, giving a dimensionality of $t \times d$. This matrix is approximated by a singular value decomposition, consisting of three matrices.

$$X = TSD' \tag{4.7}$$

In this matrix, T represents the mapping from term onto concept, S the singular value decomposition of the document, and D the mapping from document onto concept. The dimensionality reduction step of LSI is then carried out by taking only the highest K values of the SVD S . A new representation is then given for X' , an approximation of the original but with only the key concepts retained. This lower dimensionality representation may then be used as the input for related tasks from information retrieval or summarisation (Nenkova and McKeown, 2011).

Latent Dirichlet Allocation

Whereas Latent Semantic Indexes are created by manipulating matrices in a purely frequentist manner, Latent Dirichlet Allocation, or LDA, takes a probabilistic approach to dimensionality reduction. A generative model is formed for the creation of documents, and latent variables in this model, often referred to as the topics, are used as a representation with lower dimensionality (Blei et al., 2003).

The generative process for each document in LDA is given by Blei et al. (2003) as follows:

1. Choose $N \sim \text{Poisson}(\xi)$, where N is the number of words in a document.
2. Choose $\theta \sim \text{Dir}(\alpha)$, where θ is a distribution of topics in the document.
3. For each of the N words in the document, w_n :
 - (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
 - (b) Given $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic, Choose a word w_n for the document.

The three values ξ , α and β are parameters of the model. The parameter ξ gives the distribution of the lengths of documents. Parameter α contains weights for topics in documents, and β weights for words in topics. These may be trained using various Bayesian approximations.

Given this generative model, a document may be represented by its topic distribution θ , rather than its surface word itself. This representation, which depends on the topics of the terms in the document, can be of a much lower dimensionality than the original. As with LSI, this topic vector can be used in place of the term frequency vector as input to information retrieval and summarisation algorithms.

4.6.4 Overall Preprocessing

The following steps are carried out for the tweets in this work prior to other analysis. In many cases, results are given both with and without these processing sets.

1. Tweets are tokenised using Twokenize (Gimpel et al., 2011).
2. Unless otherwise specified, uppercase and lowercase forms of terms are merged.
3. Unless otherwise specified, stopwords are removed from the tweets.
4. Sequences of terms are joined as needed to give bi-grams or tri-grams.
5. Term vectors are produced from frequencies of occurrence of each n-gram.
6. Where specified, term vectors are weighted according to TF.IDF.
7. Where specified, term vectors are transformed using a topic model for dimensionality reduction.

In all experiments on summarisation for political timelines, stopwords are removed prior to further analysis. In the general timelines work, results are also given for unigrams with the stopwords left in place. Details of the creation of lists of stopwords, IDF indices and topic models will be given in the respective chapters for each area of work.

4.7 Evaluation Metrics

As discussed in Section 3.3, a key contribution of this work is a user-reported gold standard for tweet relevance in home timelines. This is in contrast to much of the related work, which relied upon retweet counts as the gold standard and did not directly ask users to express their interests (Uysal and Croft, 2011; Hong et al., 2013; Feng and Wang, 2013). The gold standard was motivated initially by the subjective and reader-sensitive nature of the task. Home timelines have only one intended reader, and that reader should be the one to annotate their interests.

There are several key challenges which must be addressed when using gold standard judgements to evaluate the quality of generated summaries. The systems under test produce a relevance ranking of tweets, whereas the users were asked to selection 8 relevant tweets, leaving 42 tweets which are then assumed to be irrelevant. A scoring algorithm must be used to compare the ranked candidates with the binary judgement scores.

The generated summaries take the form of either a limited selection of the posts which are selected as most interesting from the reader’s timeline, or a re-ranked version of the entire collection of tweets. The summary generation method to be used depends upon the type of evaluation to be carried out, and as will be demonstrated in this work, these methods differ in their treatment of redundancy and summary length, as well as the assumptions that make about the activities of the reader.

Where summaries are fixed in length, the presence of redundant information in a summary can lower its usefulness (Nenkova and McKeown, 2011), as interesting tweets may be excluded from the selected posts. This reduction is less problematic for simple re-rankings of the entire collection, as readers may ignore redundant posts and continue reading until they find tweets of interest. The assumption that readers will do so is built into ranking evaluation here (Manning et al., 2008).

The key challenge in selecting an appropriate method of automatically evaluating summaries is in ensuring that the scores will reflect the actual utility of the summary to a target user. The evaluation chosen in this work should measure the completeness

of the summary (does it contain all of the interesting tweets, or equivalent?) and redundancy (does it contain tweets that are not needed in the summary).

The scoring algorithms used for evaluation in this work are ROUGE-1 (Lin, 2004a) and MAP (Manning et al., 2008), which will be presented here.

4.7.1 Mean Average Precision

The Mean Average Precision (MAP) score (Manning et al., 2008), a method for evaluating information retrieval systems, is used to indicate the quality of a tweet scoring algorithm as a ranking task. This quantitative measure was also adopted by Ramage et al. (2010); Yan et al. (2012). Other work has compared algorithms using normalised discounted cumulative gain (nDCG) (Yan et al., 2012; Huang et al., 2011; Duan et al., 2010; Järvelin and Kekäläinen, 2002). Both of these measures calculate the quality of generated rankings, dealing differently with the idea that the number of results shown can affect the satisfaction of a reader with the results.

Mean Average Precision calculates the quality of rankings of tweets. When evaluating under MAP the ranking is generated by sorting according to relevance scores.

Given the precision score at position k in the ranking, $P(k)$, which is the fraction of documents that are relevant out of those appearing before position k in a ranking, and given that $rel(k)$, an indicator function, is 1 if the document in position k is relevant and 0 otherwise, average precision for a specific query is given by:

$$\text{Average Precision} = \frac{\sum_{k=1}^n (P(k) \times rel(k))}{\#\text{Relevant Documents}} \quad (4.8)$$

When the hypothetical user reaches an interesting post, the precision is the ratio of interesting over uninteresting posts they would have had to read to get there. The same calculation is carried out for all relevant posts in a set, and averaged to give Average Precision. The mean of these averages across all sets is the MAP.

MAP tests the strict position of post and it does not give credit to tweets which are close, but not identical to the gold standard. As such, if a collection contains many similar posts, a system may be penalised for not predicting what could, in effect, be a random decision by the user, since annotators were not instructed to select both tweets in the event that two are equally of interest. This is mitigated in this work by assuming that retweets of relevant tweets in the gold standard are themselves relevant, though this also increases the number of relevant documents.

However, MAP is forgiving in instances where the interesting posts appear rel-

atively high within but not at the very top of the generated rankings, based on the assumption that a user who is searching for interesting content is likely to keep searching until they find it; a strong system is one that requires the user to search less for the desired content. As such MAP can be expected to be more generous when summaries contain duplicate tweets near the top of the ranking.

Parts of this work involve calculating the value for tweet ranking of a single feature; in which case the tweets are sorted by the value of the numeric feature. In all cases, ties are broken by ordering randomly, which prevents systems preserving the original order shown in the annotation interface.

4.7.2 ROUGE

Whilst MAP evaluates candidates as in terms of ranking, this work can be also be considered a form of personalised summarisation. For this reason, evaluation is also carried out with ROUGE (Lin, 2004a), a common automatic evaluation method for summaries.

ROUGE is a collection of summary scoring functions which measure the textual similarity between candidate summaries, which are generated by automatic systems, and target summaries, which are generated by annotators as part of the gold standard. ROUGE is able to compare a system to a number of target summaries, allowing for multiple annotators, though in this case each collection of tweets is annotated by only one user.

Prior to applying ROUGE, a candidate summary must be generated for evaluation. When using MAP for evaluation, the summary is the entire ranked collection of tweets, which is obtained simply by sorting tweets by relevance scores. This is not a suitable summary for ROUGE, as summaries must be limited in length, so the collection is truncated to the 8 top-ranked tweets to form a summary matching the size (in number of tweets) of the target summaries.

Scores are reported in this work for the variant ROUGE-1, which is based on the number of unigrams the candidate summary has in common with the gold standard and is the same variant of ROUGE that is used in comparable work (Sharifi et al., 2010). ROUGE-1 consists of three measures, which are recall, precision, and a combined F score.

The problem of redundancy, where several tweets appear in the candidate summary which are identical or very similar to one another, will affect ROUGE more than MAP. When the summary size is fixed to 8 out of 50 tweets, it is possible to

fill it with 8 copies of the same tweet and nothing else. While MAP would continue down the ranking and evaluate the ordering lower in the set, ROUGE cannot consider tweets outside of the initial 8, and as such will penalise candidates which contain repeated tweets and fail to include other content.

ROUGE-1 has limitations which affect its indicativeness of quality in this work. Whilst ROUGE does not expect exact matching of content, it relies instead on vocabulary. Given that tweets are pre-filtered for the vocabulary they contain, many of the tweets can be expected to overlap in terms of unigrams, regardless of how relevant they are. The evaluation score is also not normalised by summary length. Summaries in this work are fixed by number of tweets, not by number of words, so longer tweets will overlap more with the vocabulary of the gold standard tweets by virtue of being longer, and ROUGE will favour summarisers which prefer longer tweets.

4.7.3 Significance

Since the results here vary by a large amount between tweet sets, it is inappropriate to discover the significance of results by comparing error bounds assuming a normal distribution. A repeated measures difference test where documents are paired is a better fit. Since the differences may not be normally distributed, this work implements a Wilcoxon Signed Rank test, which is a pair-wise non-parametric significance test.

In order to avoid errors due to multiple comparisons, the significance is reported only for differences between selected pairs of results, or between algorithms and baselines. A held out test set is used for the purpose of calculating significance and for testing the generalisability of the results seen in development to a wider context.

4.8 Discussion

This section presents the general approach used in this thesis to home timeline ranking, presenting an overview of the text based and social network based features which will be used.

As discussed in Section 4.7, algorithms will be evaluated using both MAP and ROUGE. By considering the results both as a ranking task (with MAP) and as a part of a summarisation task (with ROUGE), this work will demonstrate both the usefulness of the system on its own, as an information retrieval system, and in terms

of its potential applications.

All of the methods discussed in this chapter are only concerned with the text of the tweets. In reality, tweets are not just text, and they occur in a larger social context which includes both the author and the reader. This context is lacking in general topic-based summarisation because not every reader of tweets uses the service socially. Many users, such as SORA in the political timelines data set, are interested in consuming posts whilst never posting any of their own (Java et al., 2007).

While the text features described in this section are relevant to all of the experiments in this thesis, and will be evaluated both for the professional use case and for the general home timelines, a variety of features will also be derived from the relationship between the reader and the author or the tweet itself. These are not applicable to the political use case but will be relevant for general timelines, and as such will be introduced later in Section 6.4.

The algorithms for ranking text are drawn from existing work on text summarisation. Evaluating these algorithms specifically will discover any similarities between the problem of summarising Twitter timelines and that of summarising text in general. They will test the first hypothesis of this thesis, that text summarisation and information retrieval algorithms can be used to summarise Twitter home timelines (See Section 1.2.2).

Chapter 5

Ranking for Social Media Analysis

5.1 Introduction

The task in this work is one of ranking for tweets in home timelines, as discussed in Chapter 3. The purpose for which ranking takes place, however, is not defined, and different readers are interested in different content (Alonso et al., 2013). Chapter 6 will attempt to address the summarisation task for a general collection of users, without a defined use case. Prior to doing so, however, this chapter will detail the summarisation of home timelines for a particular, defined use case, which is that of social media monitoring.

The task of social media monitoring is represented in this work by the example of SORA, an Austrian political analysis organisation. According to Java et al. (2007), Twitter users are either sources of information, friends or information seekers, and SORA fall firmly into the latter category. They use Twitter to investigate trends, reactions and announcements surrounding political figures and events, monitoring social media in order to drive further detailed analysis. In their use case, tweet ranking takes place in a context where the desired outcome is known, even if the specific information need is not. They wish to gain an understanding of emerging political events to inform their work.

Many full text search systems assume that an information need already exists in the mind of the user (Pirolli, 2009), whereas in this case the reader must first explore the available topics in order to discover their information need. The two stage ranking process, which is defined in Section 3.2 and which forms the basis of

this work assists in this exploration. Common vocabulary is shown prior to tweet ranking in a tag cloud interface, allowing users to select topics of interest.

This two stage ranking process forms an exploratory search interface. Exploratory search, rather than focussing on answering perfectly crafted queries as accurately as possible, acknowledges the interactive and repeated nature of search on the web (Marchionini, 2006). Collections of tags may be used to drive this exploration, allowing the filtering of results in real time as they are shown (Kammerer et al., 2009). These tags can, in the same setting, be replaced by automatically generated word clouds, based on term frequency (Dörk et al., 2008).

The use case presented by SORA for social media timeline summarisation is valuable because it is restricted, compared to the much more general problem of timeline summarisation, providing an initial test case in which to gain an understanding of the problem. Not only is the specific use case known ahead of time (social media monitoring for political analysis), but the topical domain is also limited to political tweets.

It is argued throughout this work that Twitter timeline summaries are not reader-agnostic. Different readers will have very different requirements. Early discussion with SORA revealed that they require a form of exploratory search, finding topics and tweets within those topics that are good examples of discussion surrounding those topics. As such, methods are taken from information retrieval and text summarisation to foreground representative examples in response to a query or a collection of documents. These methods are described in detail in Section 4.5.

There are differences between the readers in this case study and those of Twitter as a whole. The SORA account does not itself engage with the social features of Twitter. It does not retweet, favourite or author its own posts, mention other users or engage in direct messages with other users. These differences, and their effects on the summarisation task, will be discussed further in Chapter 6 where the experiments for the SORA timeline are repeated on the general set of Twitter users.

The remainder of this chapter will present performance results for several baselines and a selection of approaches from text summarisation and information retrieval. This work will also demonstrate the use of dimensionality reduction to beat these baselines under automatic evaluation. A manual evaluation is presented to corroborate the results for automatic evaluation and to stimulate discussion of the qualitative aspects of the problem.

5.2 Methodology

A number of text summarisation and information retrieval methods are investigated in this chapter with regards to their effectiveness in assisting social media monitoring and in summarising tweets to this end. These methods used here are described in Chapter 4.

The task definition, gold standard data and evaluation methodology is described in detail in Chapter 3.

5.2.1 German Language Stoplist

The tweets in this data set are primarily in German, so a German language stoplist is used. The chosen stoplist is the one distributed alongside the Porter stemmer (Porter, 1980). No attempt was made for this part of the work to expand the stoplist to include German-specific Twitter orthography. Stoplists are used for all methods in this chapter.

5.2.2 German Language IDF

During the process of gathering information for the political timelines analysis, a collection of all the incoming tweets that the SORA account had seen over the course of a year was archived. This year-long home timeline contains 1,799,924 tweets, from which all of the gold standard sets were drawn. This corpus was used to create IDF indices.

Some consideration was given to the definition of a ‘document’ for the purposes of calculating IDF. This part of the definition is important because IDF assumes terms to be meaningful if they occur in few documents. If documents are too large, they contain every term and the weighting cannot differentiate between the frequency of terms. The actual number of times a term appears in a document is not considered for document frequency, which counts only the number of documents in which it appears, allowing IDF to work effectively for documents of varying lengths.

The scope of a document for the purposes of calculating IDF is considered in this work to be a one hour window of the SORA home timeline (the chronological collection of tweets from curated political figures). This scope is defined by time, not by topic, however since SORA follow current events, which change over time, the topic is assumed to change from window to window.

Since the entire timeline is used for IDF estimation, the background corpus

contains all of the selections used for the gold standard, so all the vocabulary in the annotated documents will be included in the index, meaning that smoothing for out of vocabulary terms is not needed.

5.3 Baselines

A number of general metadata and information retrieval baselines were implemented, as well as random ranking and one method from existing work. An effective tweet ranking method is one that outperforms these baselines.

5.3.1 Results

Algorithm	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Cosine	30.20%	32.45%	35.38%	33.59%
Cosine with IDF	25.99%	26.68%	31.01%	28.42%
Number of Favourites	27.58%	36.29%	36.88%	36.28%
Number of Retweets	23.44%	31.79%	28.01%	29.53%
Reverse Chronological	21.32%	27.12%	28.16%	27.37%
Random	22.75%	27.55%	28.40%	27.69%
Hybrid TF.IDF	28.65%	30.92%	39.94%	34.50%

Table 5.1: Performance for baselines on political data

Table 5.1 shows the baseline results. The number of times a tweet has been favourited is the best performing social baseline according to ROUGE-F, and third strongest under MAP. Retweets are somewhat weaker, giving the second best performance under ROUGE-F and the third worst under MAP.

Both favouriting and retweeting express approval of a post and notify the author, but only the latter causes the tweet to appear in the reader’s timeline. Lower scores indicate that the interesting tweets in this gold standard are not the ones that an average reader would retweet, even if they enjoyed and favourited them.

In contrast to retweet and favourite scores, text based ranking was stronger for this data under MAP than under ROUGE. Cosine and Hybrid TF.IDF are the first and second best performing methods according to MAP, but third and second under ROUGE-F.

Both Cosine and Hybrid TF.IDF discover tweets that are close to the query, and identical tweets are equally similar to these terms, so will be placed next to one another in the ranking. Methods like these, which rely on comparison to a

query set, are therefore prone to redundancy which, given the limited size of the summary, is punished more harshly by ROUGE than by MAP. This treatment of redundancy also helps explain why Hybrid TF.IDF, which attempts to introduce diversity, beats cosine under the combined ROUGE-F score. The cosine similarity method normalises for length and will prefer shorter tweets, which is not rewarded under ROUGE.

Ranking the tweets reverse chronologically (new to old) gives very poor performance, demonstrating that although Twitter displays posts reverse-chronologically, there is no preference for reverse chronological ordering. In other words, the current ordering used in the Twitter web interface is sub-optimal for users who prefer tweets ordered by relevance.

The score for random reordering under MAP is a function of the size of the relevant set, which is fixed at 8, and the length of the filtered input, which is fixed at 50. The score for random reordering under ROUGE is higher, but is determined by the textual variance of documents within the same set; since the documents are pre-filtered by the user search, this variance is low and ROUGE is higher than MAP for the random ordering.

5.4 Ranking with Text

The text-based ranking methods evaluated in this chapter characterise a collection of text units by selecting one unit that best represents the collection as a whole; either by comparison to a query, or to the most common terms, or by comparison between individual text units.

The algorithms in this section are as discussed in Section 4.5. They rely only on the text of the tweets themselves. As such, they are suitable for application where little is known about the reader, without relying on the meta data of the tweet.

5.4.1 Results

The results for Centroid and TextRank are shown in Table 5.2. Even though the ROUGE results for TextRank are higher than those for Centroid, the scores did not differ significantly ($p=0.33$ for unweighted unigrams).

Algorithms which used unigrams with IDF were more effective those that used unweighted terms, though the differences in ROUGE were also not significant ($p=0.72$ for Centroid unigram, $p=0.06$ for TextRank unigram). Likewise, no significant dif-

Algorithm	Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Centroid	Unigram	27.51%	33.00%	30.85%	31.68%
	Unigram (case preserved)	26.08%	31.97%	29.54%	30.52%
	Unigram with IDF	32.97%	35.72%	34.99%	35.11%
	Unigram with IDF (case preserved)	32.00%	35.69%	34.57%	34.86%
	Bigram only	30.66%	34.09%	33.36%	33.35%
	Unigram with IDF & bigram	33.21%	36.32%	35.84%	35.83%
	Trigram	29.70%	35.41%	31.19%	32.93%
	Unigram with IDF, bigram & trigram	33.26%	36.76%	35.64%	35.98%
	TextRank	Unigram	28.28%	33.86%	33.59%
Unigram (case preserved)		26.75%	31.62%	31.05%	31.17%
Unigram with IDF		31.81%	35.62%	39.82%	37.33%
Unigram with IDF (case preserved)		30.63%	35.10%	38.39%	36.43%
Bigram only		27.31%	30.70%	34.61%	32.17%
Unigram with IDF & bigram		33.53%	36.05%	40.35%	37.76%
Trigram		28.30%	32.14%	33.75%	32.71%
Unigram with IDF, bigram & trigram		33.85%	36.33%	40.55%	38.02%

Table 5.2: Performance of Centroid and TextRank for political timelines

ferences are observed when including bigrams alongside weighted unigrams ($p=0.8$ for centroid, $p=0.25$ for TextRank).

Although some of the textual approaches appear to outperform the baselines as described in section 4.3, the best performing, TextRank with unigrams and bigrams, does not give performance that is stronger than the favourites baseline ($p=0.58$).

Both Centroid and TextRank rank according to the common vocabulary held within the document. Where there are redundant tweets with the exact same content as one another, as is common for microblog posts surrounding news events, both methods are very likely to prioritise these posts. The generated ranking will therefore most likely contain many repeated and redundant tweets in the top positions.

Centroid and TextRank do not significantly outperform the baselines, however the sample score given is slightly higher than the baseline.

Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigram	27.40%	41.95%	33.99%	37.38%
Unigram (case preserved)	26.47%	40.29%	32.72%	35.95%
Unigram with IDF	35.73%	44.17%	40.46%	42.10%
Unigram with IDF (case preserved)	32.74%	40.62%	37.05%	38.63%
Bigram only	28.62%	38.72%	32.48%	35.19%
Unigram with IDF & bigram	34.85%	42.78%	38.13%	40.19%
Trigram	27.67%	38.86%	31.58%	34.67%
Unigram with IDF, bigram & trigram	33.75%	43.20%	37.95%	40.28%

Table 5.3: Scores for term frequency ranking on political timelines

The results for sum TF Ranking are stronger than those for centroid or TextRank (see Table 5.3). Term frequency ranking differs from centroid in that it does not assume a central prototypical document for the collection, but rather weights terms in the individual tweets by how often they appear in the collection as a whole. The resulting calculation is similar to centroid. However the centroid implementation in this work uses the cosine similarity measure which corrects for tweet length (Manning and Schütze, 1999), whereas the TF ranking does not, and therefore is more likely to select longer tweets.

In text summarisation tasks where summary length is limited to a certain number of words, longer sentences are a disadvantage if they do not relay as much information per word as a more terse version. In the ranking task, however, summary length is fixed by number of tweets for ROUGE evaluation and not at all for MAP; as such

there is no inherent advantage to shorter text units here. Furthermore, preferring short tweets privileges those which contain only the key vocabulary and no extra information. In the most extreme cases this may be just a single hashtag.

k	b	MAP	ROUGE-R	ROUGE-P	ROUGE-F
0.00	–	20.50%	29.60%	31.17%	30.14%
0.25	0.00	29.17%	32.81%	33.50%	32.95%
0.25	0.25	29.90%	31.93%	34.24%	32.78%
0.25	0.50	30.12%	31.75%	34.30%	32.73%
0.25	0.75	29.79%	31.39%	34.47%	32.60%
0.25	1.00	29.46%	31.73%	35.18%	33.09%
0.50	0.00	30.06%	32.83%	33.51%	32.95%
0.50	0.25	29.97%	31.75%	34.30%	32.73%
0.50	0.50	29.75%	31.90%	34.94%	33.11%
0.50	0.75	29.65%	31.75%	34.89%	32.99%
0.50	1.00	29.63%	31.35%	34.90%	32.76%
0.75	0.00	29.50%	32.85%	33.50%	32.96%
0.75	0.25	30.70%	31.80%	34.84%	33.00%
0.75	0.50	30.06%	32.02%	35.06%	33.23%
0.75	0.75	29.43%	31.48%	34.89%	32.86%
0.75	1.00	29.99%	31.30%	35.10%	32.82%
1.00	0.00	28.96%	32.85%	33.50%	32.96%
1.00	0.25	29.50%	31.75%	34.80%	32.96%
1.00	0.50	29.61%	31.79%	34.90%	33.03%
1.00	0.75	29.54%	31.34%	35.00%	32.82%
1.00	1.00	29.23%	30.92%	34.94%	32.52%

Table 5.4: Scores for BM25 ranking on political timelines

The results for BM25 ranking are weaker than those for centroid and for term frequency weighting, as shown in Table 5.4. BM25 differs from the other ranking methods in this section in that it is an information retrieval method in which tweets are ranked according to their similarity to the query, rather than how representative they are to the document as a whole. In cases where the parameter k , which determines the importance of term frequency, is 0, the result is independent of the length parameter b .

Given that BM25 relies on only the terms from the query, and then trades off between valuing those terms and normalising for document length, showing that either limiting the vocabulary to query terms alone is not useful here, or, at least in the case of ROUGE, there is no need to compensate for the length of tweets because summary length in words is not part of the ROUGE calculation. Longer summaries

can cover more vocabulary. The lack of preference for shorter tweets under ROUGE is also evidenced by the weakness of Centroid, which controls for tweet length, in comparison with the very similar term frequency weighting, which naturally prefers longer results.

Features	λ	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigrams (IDF)	0.0	18.57%	22.30%	24.01%	22.97%
	0.2	20.43%	27.38%	23.73%	25.20%
	0.4	23.59%	32.45%	28.89%	30.29%
	0.6	29.08%	35.85%	33.25%	34.25%
	0.8	31.72%	36.29%	34.24%	34.99%
	1.0	32.97%	35.72%	34.99%	35.11%

Table 5.5: Scores for MMR ranking on political timelines

MMR is used to prevent the placement of groups of very similar tweets at the top of the ranking, and to ensure that at a glance a user would be able to see all of the topics contained in the set. However, performance when MMR is applied was found to be worse than without MMR, and the results worsened as novelty was prioritised (See Table 5.5). In the worse case the MAP for MMR is worse than that of random reordering, showing that MMR is specifically de-emphasising interesting posts.

Redundancy reduction is harmful rather than helpful in this case, which suggests that the gold standard annotators were interested in finding canonical tweets which best represent the topic at hand, regardless of how much they repeat one another. Redundancy and its impact on social media monitoring will be analysed further in Section 5.5.

5.4.2 Dimensionality Reduction

Methods which rely on comparison between text units, such as Centroid and Text-Rank, can be improved through the use of dimensionality reduction in cases where a wide variety of mismatched vocabulary is being used to discuss a general topic (Dumais, 1991).

By creating topic models and preprocessing the tweets with them before applying the Centroid algorithm, tweets are prioritised because they discuss the central topics of a collection rather than because they use the central vocabulary. This can aid performance by making the comparison in question more robust and by favouring less the tweets which repeat the exact same wording as one another.

Since SORA follows current events, which change over time, it is assumed that,

over time, the most popular topics of discussion in their curated timeline will change. For example, discussion of the environment one day may give way to discussion of immigration the next. This assumption implies that tweets which occur closer to one another in time (IE on the same day as one another) are likely to discuss the same topic.

Based on these joint intuitions that topics change over time and that tweets close in time are closer in topic, the SORA timeline is divided into hourly windows, where each window is treated as a document for the purpose of training topic models. This is the same process as was used when developing IDF indexes (Section 5.2.2).

LDA and LSI models are trained from historical tweets posted by a list of over a thousand Austrian political users, as followed by SORA (their home timeline). One year of tweets was used for training, accounting for 1,799,924 posts overall.

After training topic models in this manner, the models are used to map the higher dimensional n-gram feature space onto a lower dimensional space of topics, as a preprocessing step. Centroid is then applied as before, in order to rank the tweets using this topic based representation.

Results

The specific hypothesis under test here is that dimensionality reduction methods can assist with tweet ranking when used as a pre-processing step prior to the Centroid algorithm.

As shown in Figure 5.1 and Table 5.6, using topic models in this way significantly outperforms using Centroid ranking with IDF ($p=0.008$, LSI 200 topics), as well as the favourites baseline ($p=0.017$, LSI 200 topics). The reported figures are for unigrams without IDF weighting.

LSI or LDA are not incorporated as a preprocessing component for TextRank, as the resulting graphs are generally very dense, with some overlap existing between every pair of tweets. TextRank performs well on sparser text graphs. In earlier experiments with TextRank, the graph with LSI and LDA would be so dense that PageRank would not converge within the time available. Future work may consider the use of TextRank here by setting a threshold on similarity empirically based on the density of the graph without dimensionality reduction.

More variation in performance results is seen for Latent Dirichlet Allocation than for Latent Semantic Indexes, though the results for the former are consistently worse than the latter. However, due to the limited size of the data set, it is not possible

to draw a reliable conclusion that LSI is always better than LDA for tweet ranking. Considerable performance improvement can, however, be seen when using topic-based dimensionality reduction methods in general. Although LSI is very similar to LDA, LSI lacks the generative probabilistic basis used for LDA, and the two are not equivalent.

The best performance was achieved by LSI with 200 topics. Given that the documents in question span the course of an entire year worth of tweets from the SORA home timeline, 200 topics represents a considerable reduction in dimensionality.

Model	Topics	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Latent Semantic Index	10 topics	33.50%	38.49%	36.90%	37.41%
	50 topics	33.83%	38.36%	38.43%	38.17%
	100 topics	39.68%	43.43%	44.14%	43.59%
	200 topics	41.44%	44.54%	46.67%	45.36%
	400 topics	40.49%	43.40%	45.40%	44.14%
	600 topics	39.51%	43.14%	45.44%	44.01%
	800 topics	38.37%	41.70%	44.13%	42.62%
	1000 topics	38.34%	42.34%	44.86%	43.30%
Latent Dirichlet Allocation	10 topics	25.74%	30.44%	30.66%	30.36%
	50 topics	28.63%	35.62%	37.87%	36.52%
	100 topics	29.71%	33.63%	34.97%	34.11%
	200 topics	27.85%	32.35%	34.64%	33.24%
	400 topics	34.40%	40.41%	42.94%	41.41%
	600 topics	38.31%	41.47%	44.46%	42.72%
	800 topics	38.24%	40.17%	43.65%	41.63%
	1000 topics	34.60%	37.70%	41.45%	39.27%

Table 5.6: Performance with Dimensionality Reduction

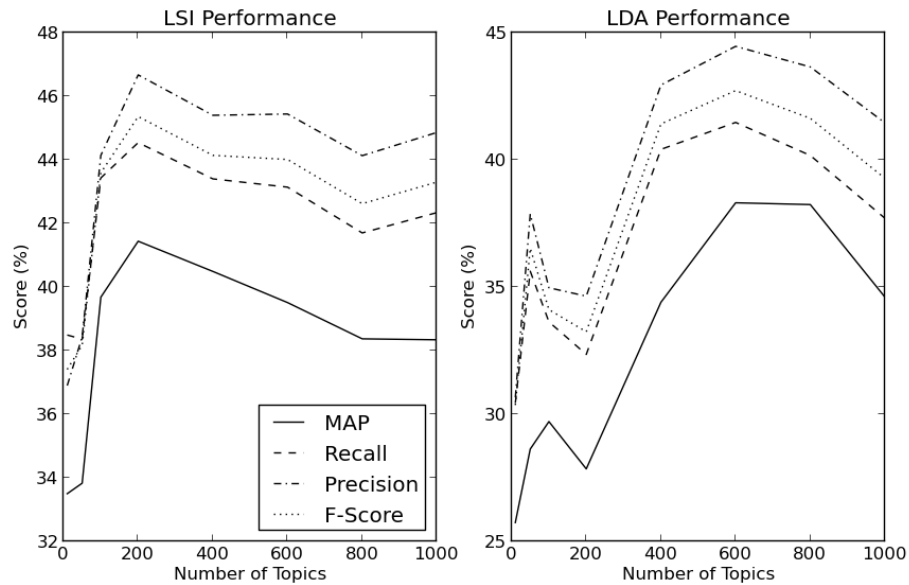


Figure 5.1: Performance with Dimensionality Reduction

5.5 User Evaluation

The type of automatic evaluation carried out so far in this chapter is useful for comparing numerous algorithms and approaches without repeated expensive human annotation (Nenkova and McKeown, 2011). Automatic evaluation is, however, only an approximate substitute for how humans would judge system performance.

The ROUGE and MAP methods used in this work calculate the coverage of the key content from the gold standard by the candidate summary. They cannot accurately measure overlap when synonyms are used in place of the terms in the gold standard, and assume that a strong candidate is one that contains the most vocabulary or tweets in common with the model summary, regardless of redundancy and coherency.

External, subjective evaluation must be used to measure qualities such as coherence, cohesiveness, or redundancy (Nenkova and McKeown, 2011). The treatment of redundant tweets is inconsistent in automatic evaluation. Many of the sets contain verbatim and modified retweets, as well as groups of tweets which all quote the same news headline. With MAP, tweets that are identical to interesting tweets are also treated as interesting. However, this means that MAP does not penalise methods which are highly redundant. ROUGE will score highly redundant summaries poorly

only if they also fail to include important vocabulary in the top 8 tweets.

The manual evaluation process in this work is to first generate example summaries, after asking SORA to once again specify topics of interest. Volunteer users are then asked to rate those summaries. This approach allows the estimation of the real impact of automatic summarisation on the use case, allowing users to evaluate subjective aspects of the summary which are difficult to quantify automatically, such as redundancy. Rather than simply comparing the words of the candidate with the model summary, annotators consider how well the semantic content of the generated summary represents the set as a whole.

This manual evaluation task is practical for the work in this chapter, as opposed to within work on general twitter timelines, because here are relatively few users to ask, and the use case is defined by a business need. The subjective ratings can reflect this use case. As followers of news and political tweets, the tweets received by SORA are not personal to them and as such several annotators can be asked to corroborate manual judgements.

5.5.1 Study Design

Tweet rankings produced by five different algorithms are compared for *completeness, redundancy, utility* and *overall subjective preference*. These dimensions were chosen from work on evaluating recommender systems (Shani and Gunawardana, 2011), though the notions of redundancy and completeness are also important in evaluating text summaries (Nenkova and McKeown, 2011). Only the dimensions which relate to the SORA use case were retained. Areas suggested for evaluation of recommender systems which were not considered included serendipity, robustness, cold start performance, scale-ability, adaptability and privacy.

Although the intent is the direct comparison between summaries, annotators were not asked to rank them directly (e.g. from best to worst), as it could be confusing to perform this task for four different dimensions. They were instead asked to score each criterion separately, using a Likert scale (Likert, 1932), which differs in that users are allowed to assign the same scores to several summaries.

The questions used were as follows:

Please give your opinion on the following criteria, regarding the summary above:

- The summary captures all the important information from the full set of tweets (completeness)

- There were several tweets in the summary that were repeating very similar information (redundancy)
- I could use this summary to study political figures or events (utility)
- Given your responses above, please rate this summary as a whole (subjective preference)

All of the responses were given on a scale of 1 (Disagree) to 5 (Agree), apart from the final rating which was made from 1 to 7, with the responses ‘Very much dislike’, ‘Dislike’, ‘Somewhat dislike’, ‘Undecided’, ‘Somewhat like’, ‘Like’ and ‘Very much like’. The intuition behind the latter, larger Likert scale was to allow users to make finer grained distinctions for their overall summary preferences.


Prior to the user study, SORA were asked to generate a new selection of keyword groups for the political use case from historical data. These filters were then used, rather than those from the gold standard generation process, to simulate a scenario in which information is being discovered for the first time, which would not be the case had the tweets already been considered in detail for a previous annotation task. Each tweet summary was presented to 3 annotators, allowing aggregation of the results from several users with the intention of gaining a more robust evaluation result.


The particular experimental configuration used for the user study does not lend itself well to significance testing. The scores on the Likert scale are ordinal, but arguably not scalar, so differences cannot be quantified unless one algorithm dominates the others, especially with so few users in the study.


The mean of pairwise Spearman’s Rho for each document set is calculated to give inter-annotator agreement. Spearman’s Rho can be used to calculate correlation for ordinal data (Shani and Gunawardana, 2011). To compare systems, they are sorted and ranked amongst each dimension according to the user response (to avoid differences in magnitude between users) and the mean rank is then given for that system across all filter sets.


Figure 5.2 shows an example summary and the interface used for evaluation. Several summaries were placed on the same page in a random order, and evaluated one after the other.


Summary A:


 **StefanHechl**
Why Is Daniel Sturridge the Only Hipster Footballer? | VICE United Kingdom <http://t.co/AgPTW7IRx> via @VICEUK


 **GeorgOstenhof**
Strassburg: Ende der Roaming-Gebühren beschlossen. Danke EU!!! <http://t.co/togPubvYxe> via @SPIEGELONLINE


 **ElmarLeimgruber**
Und Tschüss: #EU schafft #Roaminggebühren ab: #telekommunikation <http://t.co/leHE23N35K> <http://t.co/DqzpxVxrwT>

 **StephanUllrich**
@mehrenhauser Ja, war peinlich. Stimmt. Alles Gute =) @EU_Commission @EUJohnClancy @EuropaAnders @neos_eu

 **Eaglepowder**
Lenkerauskunft/Anzeige wegen Tempoüberschreitung 12 Km/h in der 100er Zone Pack -> 60 Euro. #Oida

 **phil_ipp**
Runde 3: Claudia Schmidt, #ÖVP #EU in 10 Jahren: "Kann noch wachsen." #bjv #EU2014 <http://t.co/PPc96SuUX>

 **phil_ipp**
"Junge Stimmen für die #EU" von der @_bjv_ - jetzt Diskussion mit Politikern (oder solche, die es werden wollen). <http://t.co/lgBI0mfMd>

 **StephanUllrich**
@mehrenhauser Die @EU_Commission führt Wahlkampf? Sie haben dort schon mal gearbeitet, oder? @EUJohnClancy @EuropaAnders @neos_eu

Please give your opinions on the following regarding the summary above:

The summary captures all the important information from the full set of tweets

Disagree Agree

There were several tweets in the summary that were repeating very similar information

Disagree Agree

I could use this summary to study political figures or events

Disagree Agree

Given your responses above, please rate this summary as a whole:

Very much dislike Dislike Somewhat dislike Undecided Somewhat like Like Very much Like

Figure 5.2: Interface for manual summary evaluation

5.5.2 Evaluated approaches

The evaluated approaches are a selection of the best performing methods as evaluated automatically, plus several baselines. These scores are random ranking, retweet counts, Hybrid TF.IDF, Centroid with 50 topics (LSI). 50 topics were used rather than the best performing 200, because at the time of designing the manual evaluation study, ROUGE-L was being used, which suggested stronger performance for 50 topics. ROUGE-L was later dropped in favour of ROUGE-1, for conformity to related work (Judd and Kalita, 2013; Sharifi et al., 2010; Mackie et al., 2014).

Phrase reinforcement Phrase reinforcement (Sharifi et al., 2010) was included here for comparison, despite not being included under automatic evaluation. Phrase reinforcement leverages repeated word pairs in the text to generate textual sum-

maries. Comparison with PR against the gold standard is extremely unreliable, since the summaries generated by this algorithm are short phrases rather than collections of tweets. Producing very short summaries in this way disadvantages phrase reinforcement, but comparison is possible through the user study.

5.5.3 Annotators

SORA know their own domain and information needs, but they were only able to provide a single annotator for this evaluation. Therefore, two additional German speaking users from the University of Sheffield were engaged. They have not been involved previously in the project. Both users were German, not Austrian, and worked in computer science rather than political science, so their inclusion is more useful in evaluating the summarisers themselves, rather than its applicability to the use case. For comparison, for much of its history the gold standard summaries at DUC were produced by single human annotators, or groups of annotators working on different subsets of the data (Nenkova and McKeown, 2011).

5.5.4 Agreement

Inter-annotator agreement was measured using Spearman’s rho, as the ratings are ordered categories and not intervals. These results were calculated using a set of 17 documents, for which 5 summaries each were evaluated. The reported standard deviation is that of the Spearman’s Rho for each document. The mean correlation in each dimension was moderately positive, with a relatively small standard deviation.

The full results for agreement are shown in Table 5.7. This agreement shows that the task at hand could be replicated to at least some extent by annotators outside of SORA.

Dimension	Spearman’s Rho	σ
Completeness	0.58	0.26
Rating	0.51	0.28
Redundancy	0.50	0.31
Utility	0.61	0.17

Table 5.7: Agreement between annotators for manual evaluation

5.5.5 Results

The results for manual evaluation are shown in Table 5.8. The mean rank is reported for each dimension. The random baseline appears to outperform the best performing automated tweet ranking method on all dimensions, and it is the strongest of all the approaches on all considered areas apart from redundancy, where it came second.

The best performing approach, Centroid with LSI, was not preferred over ordering tweets randomly. However, given the particular form in which preferences were expressed here, it is not possible to calculate statistical significance, since the results are not scalar. Nonetheless, it appears here that random ordering is a very strong baseline.

Diversity was very important in manual evaluation, as evidenced by the ratings for random ranking and for Hybrid TF.IDF, the two approaches which are least likely to introduce redundancy in the generated ranking. Additionally, a post-evaluation discussion with the SORA annotator revealed that they felt redundancy was a problem in the summaries that they saw, and that this affected their rankings amongst all dimensions.

This preference for more diverse summaries with lower redundancy is not shown in the automatic evaluation, where methods which attempt to reduce redundancy perform worse than those which simply use Centroid. Had this not been the case, Maximal Marginal Relevance would have been evaluated manually. As it was, under automatic evaluation the best parameters for MMR were those which introduced no randomness whatsoever (equivalent to Centroid).

There are many possible reasons why this difference between automatic and manual evaluation preferences could have occurred, including the passing of time meaning that less current context is held implicitly in the memories of the annotators. It is also possible that the information needs of the SORA users have changed since they created the gold standard for automatic evaluation. Additionally, the ROUGE and MAP evaluations do not consistently penalise summaries which contain redundancy.

The phrase reinforcement algorithm also performed very poorly on all but redundancy. The summaries it generated were very short, usually only a few words. Given that the summaries were only a few words, whereas the other summaries, which are tweet rankings, are much longer. It would be very surprising for such a short summary to have more redundancy. The annotators also claimed that the summaries generated using phrase reinforcement were nonsensical. The tweet sets

provided for this task were too small and contained too little repetition for phrase reinforcement to work optimally.

Algorithm	Completeness	Rating	Redundancy	Utility
Phrase Reinforcement	0.49	0.41	0.76	0.30
Centroid (LSI 50 topics)	2.00	1.86	2.81	2.24
Retweet counts	2.27	2.19	2.38	2.16
Hybrid TF.IDF	2.59	2.59	2.22	2.62
Random	2.65	2.95	1.84	2.68

Table 5.8: Mean rank of 5 systems under manual evaluation

5.5.6 Revisiting automatic evaluation

The discrepancy between the results of the automatic evaluation and the user-based study could result from a change in the information needs of SORA since the gold data corpus was first annotated. This would explain the difference in performance between the centroid method on automatic evaluation, in comparison to manual evaluation. To test this, SORA re-annotated parts of the same gold standard data, without being told that they had already annotated these tweets in the past. Agreement was then measured between the same SORA annotator at two different points in time. The first set of gold data annotations were created some time in October 2013 and the second set a year later in October 2014.

The annotation interface was slightly revised from the one used previously, to give less guidance (since the task had been carried out before). The revised annotation interface is shown in Figure 5.3.

The observed agreement on a random selection of 15 tweet sets from the original gold standard was 0.0743 (Fleiss' Kappa). This low score suggests that either:

- The information needs of SORA have changed in the year's period between the two sets of annotations.
- Or the decisions made by SORA are different when made within a recent time frame and given more available context.
- Or, where a tweet set contains several similar tweets, SORA may have chosen randomly twice on two separate occasions (verbatim tweets were not merged here).

While the annotations have changed to an extent, one common comment from SORA throughout this work is that the sets generated by the keyword selection approach in particular contain many closely related tweets. Where the user has no particular preference between two identical tweets, they can select differently at random, both times.

To determine what effect, if any, the differences between SORA in 2013 and in 2014 have on the performance of these systems, ROUGE could be used for comparison, however the size of the set is very small and it would have been very resource intensive to revisit enough sets to allow robust comparison between algorithms.

Judge Tweets

Please select (by clicking) the most interesting tweets from the set below.

We ask that you choose precisely the 8 most interesting tweets.

Show Selections Please select **6** more tweets. Your name here Save

KlausSchwertner - add
 +1 RT @SusanneSchnabl: In Ktn. wird einiges anders: Nicht nur auf d Saualim: #Asylwerber dürfen nun offiziell arbeiten [kleinezeitung.at/nachrichten/ch...](#)

eminenz - remove
 Markus #Kafka und der @bierpapst als StreithansIn bei @ServusTV - ist das für den Spannungsbogen bei #bierontour?

ipoStandardat - add
 ÖVP - Christine Marek tritt nicht mehr bei Nationalratswahl an: "Zehn Jahre Spitzenpolitik sind genug" [bit.ly/ZLKYQt](#)

datenschmutz - remove
 Der ★Social Media Bote★ vom 29.4.2013 ▶ Topthemen via @SZ_Kultur @panda1616 @futurezoneat [datadirt.net/smb130429](#)

CarFreiTag - add
 @Der_Gregor Zuviel Sport oder zuviel Arbeit?

Figure 5.3: Revised manual annotation interface

5.6 Discussion

This chapter addressed the problem of ranking content in Twitter timelines for social media monitoring for a political analysis use case. Through the creation of a flexible gold standard as part of a multi-stage process, users of this type were able to define their information need, and the kind of results they require in order to meet these needs.

The first hypothesis in this work, which is that text summarisation and information retrieval algorithms can be applied to the specific domain (Section 1.2.2), is partially demonstrated in this section, albeit for the limited use case. It was shown that under automatic evaluation, methods from text summarisation including Centroid and dimensionality reduction can help produce automatic summaries which

beat both the status quo (chronological ordering) and a sensible baseline (counts of retweets and favourites) according to automatic evaluation.

Unfortunately, this evaluation has been limited by its scope. The gold standard was relatively small, and was produced by a single organisation over a short period of time. The use case that was implicitly created by this process was very specific.

The limited nature of the use case was highlighted in the manual user study of the generated summaries. The volunteers appeared to prefer random summaries over the best summariser according to the gold standard, though the results for this experiment were not conclusive. Similarly, even the same annotator repeating the same task a year later did not give judgements which correlated well with those made previously.

The automatic evaluation results show that for a specific use case of social media monitoring, the use of Centroid ranking in combination with LSI for preprocessing can lead to large improvements in performance. In early discussion, annotators indicated that they were looking for the most representative examples from the tweetset. However, later discussion as part of the manual evaluation revealed that they were also concerned with diversity in the generated summaries, finding unacceptable amounts of repetition in the candidate sets where tweets were near copies of one another or direct retweets.

This thesis as a whole aims to prioritise content from the home timeline for all users of Twitter, with varying information needs. This chapter addresses a specific use case, but the remainder of the work presents a more general version of home timeline summarisation, with a gold standard of tweets from many users at The University of Sheffield. The experiments found in this chapter will be repeated on the new, more general corpus of judgements, giving an understanding of whether the results presented here can be replicated for more than just political social media monitoring.

Chapter 6

Ranking for General Twitter Users

6.1 Introduction

The previous chapter presents a very specific use case for prioritising tweets in home timelines - that of assisting social media analysis for politicians and political events. The work therein is a restricted version of the overall effort in this thesis to rank tweets from home timelines; in the general case, the purpose for which ranking is carried it is not known ahead of time, nor is the reader, the kind of authors, or the topics of interest. Whilst the previous work gives an informative case study for this kind of ranking, it is the general ranking task that will be considered in this chapter, and the remainder of this thesis.

The ranking of home timelines in general is not constrained by topic or purpose. Though some of the users in the study were professionals (38% of Twitterers use the service professionally (Bontcheva et al., 2013b)), the majority are not. A sample of 148 volunteers from the University of Sheffield were recruited to filter and annotate their own personal timelines, resulting in a set of 387 document sets with relevance information (See Section 3.3). These annotators were not directed to choose tweets for any particular purpose other than what they found interesting, and no interviews or manual evaluation was carried out.

A series of investigations was carried out into general home timeline ranking using this data set. Section 6.3 describes the repetition of experiments from political timelines (Chapter 5), determining whether these results can be replicated on the general data. Social media information is then utilised in Section 6.4 to improve

tweet ranking, exploiting the interactive nature of Twitter to personalise the ranked tweets. A series of machine learning experiments are carried out in Section 6.5 and Section 6.6, producing models which combine multiple features for personalised tweet ranking.

6.2 Methodology

Tweets in this part of the work are now no longer drawn from a single timeline for all experiments, but rather each annotation set is drawn from the home timeline of the corresponding annotator, giving a much greater breadth of topics and types of tweets.

There is no expectation that tweets are marked relevant because they are of professional value to the user. Table 3.10 shows the distribution of reasons for which tweets were found interesting in Twitter timelines, coded using responses given in free text fields by users in the pilot study and validated in the final gold standard. Whereas relevance to interests was a common category, humour was also the reason for 20% of tweets that were found relevant, and 9% were marked as such because the reader was sympathetic to the author’s point of view, reasons one would not expect in the political analysis task whatsoever.

6.2.1 English Language Stoplist

The tweets in the general dataset were largely in English (see Section 3.3.2), so a stoplist from the Porter stemmer package (Porter, 1980) was used. A collection of the 100 most common tokens from the gathered tweets was manually inspected to discover Twitter specific stopwords, which were selectively added to the stoplist.

6.2.2 English Language IDF

IDF indexes for English tweets could be created using the gold standard that is used for development and for testing of the ranking algorithms. The development portion of the gold standard might be used to create an index in which one tweetset is one document. This unfortunately creates undesirable interactions because nearly every term from a tweet set filter will appear in roughly one document, unless the same term appears in multiple filters, and as such will receive the same IDF score $\log \frac{|D|}{1} = \log |D|$. Being unable to assign different weights to different query terms limits the utility of IDF.

In the SORA study, the IDF index was computed with the entire previous home timeline of the SORA user. To mirror this, the historical home timelines (the entire collection of incoming tweets) were also gathered for every volunteer in the study. It is this collection of timelines on which IDF is created, creating one model for the entire corpus (rather than one per user).

For each reader in the gold standard, a list of the authors they follow is gathered, giving 51841 authors overall. The user timeline for each author is fetched, giving approximately 3200 of their most recent posts. This corpus of user timelines is merged for each volunteer in the study to form their historical timeline. The result of this process is a stream for each user in the data set much like the one used for training topic models for summarisation of political timelines. The stream reflects what the volunteer would have seen on Twitter over the course of several years, though there will be more missing tweets as time passes.

Since many more timelines had to be gathered in this manner, the process was carried out as a batched job over one year after the original version of the personal timelines data set was created. It is unfortunately not guaranteed to contain all of the original tweet sets since several of the users featured in the original study are no longer on Twitter, or have since posted more than the 3200 tweets that can be collected for each user through the Twitter API, so the earlier posts are not longer available.

The historical timelines that are use for training here do not contain the annotated collections of tweets in the gold standard. As such, it is possible for terms to appear in the gold standard that are not present in these timelines. The version of IDF used so far in this work is not defined for these previously unseen terms, because it would require division by zero. As such, Laplace smoothing is applied to IDF scores in the experiments in this chapter (Manning et al., 2008).

IDF is ‘smoothed’ by introducing a fictional additional document which contains every term which could ever exist. This reduces the scores given to each seen term in order to add weight to all unseen terms. The IDF score then becomes:

$$\text{IDF}(t, D) = \log \frac{|D| + 1}{|\{d \in D : t \in d\}| + 1} \quad (6.1)$$

This score is given for all t . In practise, the weight assigned to very rare terms is extremely large, as is appropriate, and the additional document has little effect on very common terms, thanks to the logarithmic scaling. A stop list was applied prior to calculating IDF.

Algorithm	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Cosine	26.46%	27.41%	29.23%	27.98%
Cosine with IDF	25.45%	27.56%	30.08%	28.42%
Favourites Count	26.41%	31.42%	31.20%	31.11%
Retweets Count	26.82%	33.10%	31.41%	32.04%
Tweet Age	22.98%	29.42%	28.93%	28.99%
Random	22.37%	29.04%	28.46%	28.52%
Hybrid TF.IDF	24.46%	30.60%	32.00%	31.02%

Table 6.1: Baseline performance results

Twitter contains many tokens which indicate links to websites using URLs. Arguably, URLs are, by their nature, somewhat short lived. Whatever the theoretical use of an IDF index that contains URLs, such an index will necessarily either be constantly updated or forever out of date. URLs are expanded to their long form (where they have been shortened) and one token is created per domain. This would give a very low IDF for ‘imgur.com’ (a popular image sharing website) and a high IDF for ‘domrout.co.uk’ for example.

Whereas for political timelines, the use of hourly windows as documents for IDF makes sense in the context of a single account that follows current events, assuming that the discussion is focussed and changes often enough for each hour to be a different document. This assumption is not made for the general collection of timelines, instead, each tweet is treated as a separate document.

6.2.3 Baselines

The baseline performance results are shown in Table 6.1. The key baselines here are similarity to the query, popularity (retweet and favourite counts), the age of the tweet, random ordering and Hybrid TF.IDF.

Ranking by the retweet counts alone gave positive results, outperforming all other baselines. In comparable work, retweeted statuses were used as a gold standard (Yan et al., 2012; Uysal and Croft, 2011), so they should form a strong baseline; however, this was not the case in the political timelines use case, wherein retweets were outperformed by favourites, Hybrid TF.IDF and Cosine similarity to the query. Counts of favourites give similarly strong performance on the general dataset, though they are not present for every tweet, since part of the data was collected before favourite counts were made available through the Twitter API.

The random baseline is strong under ROUGE, outperforming tweet age (which is

the status quo, given that newer tweets are currently shown first by the Twitter web client), and both variants of cosine similarity. The documents in question are already filtered according to topic, so one would expect random to be a strong baseline for relevance. One possible interpretation of this result is that diversity is important in tweet rankings, and that methods such as cosine and time tend to cluster together very similar tweets.

Hybrid TF.IDF is comparable with favourite counts under ROUGE, but not under MAP. Hybrid TF.IDF contains measures to help mitigate redundant tweets, yet MAP ignores redundancy in candidate tweets and rewards stronger overall rankings, whereas ROUGE punishes summaries which contain redundant tweets due to the limited summary size (See Section 4.7).

The text baselines, which treat the problem as one of information retrieval and use comparison to the query, are much weaker in the general case than they are for the SORA use case. Whilst SORA stated that they were using the system as an aid to search and exploration, and looking for tweets that best addressed their query, the general population of Twitter preferred tweets for many different reasons (See Table 3.10), not just because of query relevance.

6.3 Text-based Ranking

Text-based ranking features were shown in Chapter 5 to outperform all of the baselines. The general tweet ranking task presented in this work is not restricted to social media monitoring, however, and in this section the text-based ranking features are once again evaluated for the general population of volunteer Twitter users.

The design of the experiments from the previous chapter has been repeated as closely as possible, with the exception that there are now hundreds of Twitter accounts under consideration, rather than one.

6.3.1 Centroid

The performance of centroid for the personal timelines data set is shown in Table 6.2. In the professional timelines data set, centroid alone was found to produce a significant improvement in performance, but these results were not replicated here for general timelines. The results seen in general are instead worse than ordering tweets randomly.

If there are many tweets in the document set that are almost identical, centroid

Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigram without Stoplist	21.80%	27.89%	24.85%	26.10%
Unigram (case preserved)	21.20%	27.29%	24.14%	25.43%
Unigram with IDF	23.03%	25.34%	24.97%	24.85%
Unigram with IDF (case preserved)	23.32%	25.36%	24.65%	24.70%
Unigram with Stoplist	23.20%	27.11%	25.35%	25.95%
Bigram only	20.98%	25.72%	22.34%	23.66%
Unigram with IDF & bigram	22.92%	25.30%	24.92%	24.80%
Trigram	23.33%	27.22%	22.91%	24.66%
Unigram with IDF, bigram & trigram	23.01%	25.30%	24.89%	24.78%

Table 6.2: Centroid performance results for filtered tweets

will rank all of them highly, as they reinforce one another. As such, in documents with much repetition, centroid is prone to creating summaries with large amounts of redundancy.

Given that queries can grow as long as a tweet in many cases, and that every tweet in the document must contain at least one of those terms, the centroid may simply become a weighted version of the query itself, with the consequence that the Centroid score is similar to the cosine similarity baseline.

Centroid makes intuitive sense where the ideal summary is a short indication of the overall content of the collection. This method cannot rank tweets according to a general notion of ‘interestingness’ to the reader, because tweets are instead ranked according to how representative their vocabulary is of the collection as a whole.

6.3.2 TextRank

Results for TextRank are shown in Table 6.3. For the professional timelines data set, TextRank performed similarly to Centroid. Here much the same result is observed though the results are still weaker than the random baseline. While TextRank no longer assumes that a document discusses a single salient topic, it does share several properties with Centroid. The important tweets are assumed to be those that are most similar to others in the set, and pairs of identical tweets can mutually reinforce one another.

Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigram without Stoplist	21.83%	27.72%	25.31%	26.28%
Unigram (case preserved)	21.32%	27.29%	24.71%	25.76%
Unigram with IDF	24.04%	27.06%	28.01%	27.24%
Unigram with IDF (case preserved)	23.47%	26.36%	27.11%	26.46%
Unigram with Stoplist	23.33%	27.69%	27.13%	27.15%
Bigram only	20.14%	25.16%	23.68%	24.13%
Unigram with IDF & bigram	23.88%	26.79%	27.64%	26.91%
Trigram	22.21%	26.54%	24.68%	25.35%
Unigram with IDF, bigram & trigram	23.87%	26.86%	27.70%	26.97%

Table 6.3: TextRank performance results for filtered tweets

Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigram without Stoplist	22.20%	30.40%	25.50%	27.55%
Unigram (case preserved)	21.83%	29.69%	24.78%	26.83%
Unigram with Stoplist	23.05%	30.43%	24.98%	27.24%
Unigram with IDF	22.14%	27.28%	23.09%	24.79%
Unigram with IDF (case preserved)	21.96%	26.47%	22.32%	24.01%
Bigram only	21.35%	27.30%	22.17%	24.27%
Unigram with IDF & bigram	21.99%	27.23%	22.76%	24.58%
Trigram	23.03%	28.34%	22.98%	25.19%
Unigram with IDF, bigram & trigram	22.07%	27.03%	22.43%	24.29%

Table 6.4: Scores for TF weighting on personal timelines

k	b	MAP	ROUGE-R	ROUGE-P	ROUGE-F
0.00	—	22.14%	29.94%	29.20%	29.39%
0.25	0.00	25.95%	30.12%	27.90%	28.76%
0.25	0.25	25.91%	29.58%	28.09%	28.59%
0.25	0.50	25.89%	29.58%	28.12%	28.61%
0.25	0.75	26.04%	29.54%	28.13%	28.60%
0.25	1.00	25.78%	29.53%	28.18%	28.62%
0.50	0.00	26.02%	30.13%	27.91%	28.78%
0.50	0.25	25.90%	29.59%	28.12%	28.61%
0.50	0.50	25.94%	29.70%	28.33%	28.78%
0.50	0.75	25.72%	29.74%	28.50%	28.88%
0.50	1.00	25.87%	29.59%	28.44%	28.78%
0.75	0.00	25.74%	30.22%	28.00%	28.87%
0.75	0.25	26.02%	29.68%	28.23%	28.72%
0.75	0.50	25.84%	29.74%	28.43%	28.85%
0.75	0.75	25.89%	29.52%	28.38%	28.71%
0.75	1.00	25.74%	29.36%	28.40%	28.63%
1.00	0.00	25.83%	30.05%	27.82%	28.69%
1.00	0.25	25.88%	29.59%	28.17%	28.64%
1.00	0.50	26.07%	29.62%	28.38%	28.76%
1.00	0.75	26.00%	29.44%	28.43%	28.69%
1.00	1.00	25.81%	29.51%	28.73%	28.87%

Table 6.5: Scores for BM25 ranking

6.3.3 Sum Term Frequency (TF)

The scores for sum TF, as with centroid and TextRank, are consistently worse than random ordering, as shown in Table 6.4.

Both sum term frequency (without any weighting) and sum term frequency with IDF weighting were evaluated. In TF.IDF weighting, the tweets that are selected are those which contain the most content bearing words, according to the background corpus of all home timelines. The poor performance for TF.IDF scoring indicates that the tweets of interest for a user are not simply those which contain the largest number of contentful words.

Given that these scores are worse than random ordering, they might be improved if combined with measures to reduce redundancy, as is the case with Hybrid TF.IDF, which gave higher scores for both MAP and ROUGE (Section 6.2.3).

6.3.4 BM25

Table 6.5 contains results for BM25 ranking. When $k = 0$, the term frequency and length of the candidate tweets are not used at all, and the tweets are ranked by the sum of the IDF scores for whatever query terms they contain. Under ROUGE-F this gives the strongest combined score. The differences between BM25 and cosine ranking are very small in the case where k and b are both 0. The term frequency is not taken into account with these parameters, so only the IDF scores are summed, though tweets are so short as to contain few repeated terms. Longer text units are preferred, so long as the extra terms appear in the query.

In the same way that Centroid with IDF was weaker under ROUGE than TF.IDF, which also prefers longer tweets, BM25, which prefers longer tweets is slightly stronger than Cosine similarity. Both rely on similarity to the tweet, weighted by TF.IDF, but again the stronger method is the one which does not normalise for tweet length. ROUGE-1 does not normalise according to the length of the candidate summary Lin (2004a), so stronger results can be obtained by selecting longer tweets for the summary.

The results for MAP, whilst stronger than random ranking, are very similar to the information retrieval baseline cosine similarity. These MAP scores do not appear to depend heavily on the choice of parameters, with the exception that they do suffer greatly when term frequency and length are both ignored.

6.3.5 Maximal Marginal Relevance

In the social media monitoring use case, redundancy did not affect performance under automatic evaluation, but it was highlighted as a clear problem in subsequent manual evaluation. Annotators complained that they were being shown the same tweet many times in a short summary. In trying to generalise the results of that work to the general timelines, it has been seen that Centroid, a method which will actually increase redundancy if it is possible to do so, scores persistently worse than random ordering.

By comparison, the Hybrid TF.IDF method (Inouye and Kalita, 2011), which includes a measure of redundancy, performs much more strongly under ROUGE. In order to address redundancy within these experiments, maximal marginal relevance, or MMR, was introduced (Carbonell and Goldstein, 1998).

Under ROUGE, the results for MMR are somewhat encouraging (Table 6.6). Both unigrams with stoplists and bigrams give ROUGE scores that are comparable

Features	λ	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Unigram without Stoplist	0.00	24.24%	28.54%	30.00%	29.03%
	0.25	22.66%	30.40%	28.95%	29.44%
	0.50	21.54%	29.95%	26.52%	27.95%
	0.75	21.46%	29.15%	25.74%	27.15%
	1.00	21.81%	27.91%	24.83%	26.09%
Unigram with Stoplist	0.00	23.03%	28.14%	27.66%	27.70%
	0.25	23.50%	32.75%	29.55%	30.85%
	0.50	22.85%	33.39%	29.91%	31.34%
	0.75	22.91%	32.14%	28.70%	30.10%
	1.00	23.20%	27.07%	25.23%	25.87%
Unigram with IDF	0.00	23.35%	29.32%	28.62%	28.78%
	0.25	22.07%	29.48%	27.69%	28.25%
	0.50	22.55%	26.51%	25.68%	25.78%
	0.75	22.89%	26.05%	25.47%	25.44%
	1.00	23.03%	25.35%	24.98%	24.86%
Bigram	0.00	24.27%	29.58%	29.60%	29.40%
	0.25	22.89%	33.63%	29.49%	31.24%
	0.50	21.55%	31.87%	27.22%	29.16%
	0.75	20.70%	28.59%	24.59%	26.19%
	1.00	21.02%	25.73%	22.34%	23.67%
Trigram	0.00	23.66%	29.87%	29.33%	29.41%
	0.25	22.67%	32.35%	26.66%	29.03%
	0.50	22.47%	32.23%	26.55%	28.91%
	0.75	22.48%	30.13%	24.78%	26.98%
	1.00	23.33%	27.13%	22.86%	24.59%

Table 6.6: Scores for MMR ranking on personal timelines

to retweets in terms of recall, but slightly lower for precision. Under ROUGE, it appears that introducing diversity into the generated summaries deliberately, through MMR, can lead to improvements in performance over other text-based methods.

Unfortunately the MAP scores for MMR are much lower. Since MAP is not applied to short, fixed sized summaries, it penalises redundancy far less because the score indicates the quality of the ranking as a whole, including tweets after the 8th position. This property is clearly evidenced by the way that MAP scores remain much the same regardless of the degree of diversity included by the λ parameter of MMR.

6.3.6 Dimensionality Reduction

The experiments in summarisation for social media monitoring showed that the performance under MAP and ROUGE could be improved using dimensionality reduction (See section 5.4.2). The methodology from these experiments are replicated here, though given that there are over a hundred users in the general gold standard, rather than just one, some changes are made to the methodology, including testing only the best performing parameters from the earlier work, in order to avoid the computing time required to test the many numerous parametisations.

Pre-processing personal timelines here presents a much larger technical challenge than for the political timelines, as to replicate the setting fully one must train a selection of topic models for each volunteer in the study, rather than for just one user.

Training Data

LSI and LDA models are trained using a collection of tweets from 51841 Twitter users, organised into timelines for each volunteer in the study (The process of gathering this collection is described in full in Section 6.2.2). The timelines are split into hourly intervals for training.

This training data is deliberately as similar in form as possible to that used in the training of topic models for political timelines. In the political timelines work, models were trained on hourly slices of the home timeline for a single user. Here, the models are trained for each volunteer on hourly slices of their home timeline. This replication requires large amounts of training data and computation, but is necessary to accurately reflect the experiment setup which gave strong results previously.

Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Centroid	24.02%	27.21%	26.29%	26.51%
MMR ($\lambda = 0.00$)	23.49%	28.12%	28.34%	28.04%
MMR ($\lambda = 0.25$)	23.14%	30.09%	28.63%	29.09%
MMR ($\lambda = 0.50$)	23.81%	29.29%	27.92%	28.34%
MMR ($\lambda = 0.75$)	24.10%	28.71%	27.62%	27.90%
MMR ($\lambda = 1.00$)	23.95%	26.94%	26.10%	26.27%

Table 6.7: Performance of topic modelling with unigram inputs

Training topic models

Given that there is one test user in the SORA data set, and hundreds in the personal timelines data set, a correspondence between the two sets of experiments is formed by training one topic model for each user in the experiment, using their full historical timeline and the same hourly interval. When used in an actual summarisation setting, users who have not followed many accounts would have far less data available for training.

Training individual models will result in hundreds of topic models for each set of parameters, and is an extremely data intensive process. As such just one model is trained per user, using LSI with 200 topics, which gives the best performance for the political set as shown in Section 5.4.2.

Results

Table 6.7 shows the performance centroid and MMR with dimensionality reduction.

In this setting, LSI does not improve the Centroid score the way it did for the professional timelines. Under MMR, the use of LSI for preprocessing leads to results that are comparable to but not stronger than MMR applied to the unigram counts on which the models are based.

This failure to replicate the strongest results from the political analysis use case demonstrates once again the information needs of Twitter users are varied. A preprocessing step which significantly outperforms all baselines for a large number of queries by a professional user is no use at all to the group of general volunteer readers.

The topic modelling preprocessing step can lead to improvements in performance where exact vocabulary matching is not desired, and where differing language is used to refer to the same concepts or ‘topics’ (Dumais, 1991). For example, dimensionality reduction can help Centroid and TextRank determine central topics even if the

exact central vocabulary is not repeated often. Centroid with a topic modelling preprocessing step will still place identical tweets together in the ranking, but will score similar tweets which do not use identical terms highly too.

Nonetheless, given the weak performance of every method evaluated in this chapter to present tweets which are representative of the collection as a whole, perhaps what is needed is not a better version of centroid or TextRank ranking, but another kind of recommendation entirely, based on personalisation to the reader or the popularity of the posts.

6.3.7 Discussion

There were two purposes to evaluating the text-based algorithms in this section, the first being to replicate the promising results seen through automatic evaluation for the political use case on a more general corpus. The results from the initial experiments in political timeline ranking were not replicated, despite best attempts to reproduce the experimental setup from this earlier work. Whilst SORA present a clear and focussed information need in the social media monitoring use case, their use case and information need are not representative of Twitter as a whole.

The second objective was to determine whether in general the tweets that are of interest to a user are those which best represent the collection as a whole, or which correspond best to the filter query. Although the best of these methods, MMR, gives performance under ROUGE-F that is comparable to the strongest baseline, retweet counts, it is weaker than that baseline under MAP, indicating that although the ranking forms a summary that is comparable to sorting tweets by retweet counts up to the first 8, the ranking as a whole does not prioritise interesting content past this initial threshold. Likewise, none of the other methods considered in this section give performance that is stronger than retweet counts under either ROUGE or MAP.

Exploiting models of the text itself is ineffective when compared to simple counts of endorsements for discovering interesting posts from the home timeline. In the remainder of this chapter, a number of features will be devised which do not rely on the text of the tweets alone, incorporating also the past activity of the reader as well as other features which are specific to social media posts.

6.4 Ranking with Social Information

In summarisation of extended text, the context of a text unit is given by its situation as part of a longer section of text. The placement of a paragraph can indicate, for example, that it is the conclusion or introduction of a paper, or perhaps that it is the summary paragraph at the beginning of a news article. Sentences earlier or later in a paragraph may prove more important than those in the middle. Likewise, the rest of the document in question can contain important cues about the subject, argument and general theme and these may be compared with that sentence (Nenkova and McKeown, 2011).

In contrast, the textual context in which a tweet takes place can take a number of different forms.

- Tweets can be part of a linear series of posts by an author perhaps signalled by ‘...’
- Conversations can be formed between groups of users, in which case the context of a tweet is the post to which it directly replies.
- Posts can refer to global topics using keywords or hashtags, and as such take part in that conversation.

These contexts in which a tweet takes place are not similar to the context in which a sentence from other domains occurs. Where text summarisation techniques make use of text headings, leading and closing paragraphs, and the placement of the text within a paragraph, these are not present for tweets. In place of this textual context, other features are available for ranking, such as the authorship of the tweet, the relationship between author and reader in which the tweet may occur, and information about the reader themselves.

The availability of this additional information differentiates the summarisation of content in social media with that of summarisation of other mediums where there is a separation between author and reader such as newswire and sports broadcasts. One part of this information, the identity of the author, allows the use of their previous posts and the popularity of those in the creation of authority models, predicting the value of author tweets just based on who they are (Uysal and Croft, 2011; Feng and Wang, 2013). Having access to the writings, retweets and favourites of the intended reader allows for recommender systems which tailor posts to the reader’s own interests (Yan et al., 2012; Sun and Zhu, 2013; Hong et al., 2013).

In this section, social and personal metadata is used in place of the text of the candidate set for home timeline ranking. The introduction of additional types of information here foregoes the assumption that relevant tweets are those that are representative of the collection and instead allows the scoring of tweets along dimensions which are personal or social in nature.

6.4.1 Features

A number of features which capture author and preference information specific to social media will be presented in this section. These features broadly fall into three groups: those which measure the popularity of the author, those which quantify the relationship between author and reader, and those which compare the interests of the reader with the candidate content.

The utility of counts of retweet and favourites was previously evaluated as a baseline in Table 6.1. These features directly represent the popularity of the tweet in the wider social network. Limitations of retweet counts are discussed in Section 3.5.

The retweet and favourite count features capture the global popularity of the tweet on the whole of the network. However, since the tweets in home timelines that are considered interesting depends heavily on the individual's own interests, effective ranking needs to be personalised (Ren et al., 2013; Chen et al., 2012; Feng and Wang, 2013) through user-specific features. These include connections of the tweet author within the social network; reciprocated connection between the tweet author and the timeline owner; and the number of shared followers and friends.

Author Popularity

The popularity of the tweet author could be used as a proxy indicator of the interestingness of their tweets, which is particularly relevant for tweets by celebrities or Twitterati. The following features are evaluated for ranking; they gauge the status of the tweet author as a potential celebrity user. These features are similar to those used in the measurement of account authority, which has been considered as an indicator of tweet quality (Huang et al., 2011; Duan et al., 2010). The values are per-user, rather than per-tweet. When ranking tweets, each tweet is assigned the score for its author. Where multiple tweets are present by the same author, they are ordered at random within that author's posts.

- friends/followers of author

- Number of followers of the author
- Number of Twitter ‘lists’ (curated collections of authors) on which the author appears

Author-Reader Relationship

Since there is little to no barrier to posting on social media services, users of social media very often read content by authors with whom they have some relationship, and they may communicate directly with those authors via the same platform. Knowledge about this relationship may help in ranking social media posts, for instance if users prefer posts by authors with whom they are close personal friends, or if they prefer to read professionally authored tweets from news agencies.

Social connections have been incorporated into models for Twitter recommender systems using collaborative filtering and graph co-factoring (Ren et al., 2013; Yan et al., 2012; Sun and Zhu, 2013; Feng and Wang, 2013; Hong et al., 2013), though in this work the method by which the features are applied and the variant of the task at hand are both dissimilar to recommender systems. Recommender systems prioritise content from the entire collection which is preferred by users which are similar to the target reader (Shani and Gunawardana, 2011). The input set here is restricted to the home timeline, and the assumption is avoided that a tweet must be interesting to anyone other than the reader in order to be interesting to them, for example in the case of personal tweets which directly address the reader.

Relationships between users have also previously been characterised and applied to the task of geo-locating Twitter users automatically in our own work (Rout et al., 2013b). Here it was noted that friendships represented on Twitter can either be a mirror of a correspondence in a wider context such as a university, workplace or pub, or a relationship between a reader and some primary content producer such as a journalist, artist or celebrity. Features designed to indicate the strength of relationships were incorporated into models of geographical collocation with some success for geolocation.

During the gold standard annotation process a mapping was gathered containing the ID numbers of accounts followed by each participant in the study, and by each of the accounts that those followee accounts followed. This represents a subset of the Twitter graph containing the outgoing connections of the participants and their friends. The graph was then used to generate scalar and binary features to represent both the strength and reciprocity of the connection between author and

reader. Relationship strength is given by the number of mutual friends (followees), since close real-world relationships tend to lead to mutual connections (Easley and Kleinberg, 2010), and by counts of past interactions.

The author-reader relationship features as follows:

- Does author follow reader?
- Does the tweet mention the reader?
- Number of friends (followees) in common with author
- Number of times the author has mentioned the reader
- Number of times the reader has mentioned the author
- Number of times the reader has retweeted the author
- Number of times the reader has replied to the author

To preserve the privacy of the participants, private messages were not included in previous interactions when generating relationship features.

Reader Profile

This group of features will be used to model the reader's interests using the following sources of text:

1. Posts that are written by the reader.
2. Posts which are endorsed by the reader through retweets and favourites.
3. The personal profile of the reader, which contains a short self-description.

Authored Tweets The majority of users in the study created posts of their own, as well as consuming those of others on Twitter; 83.2% of volunteers had posted more than 50 tweets, and only one user had never posted at all. A rank plot of user post counts is shown in Table 6.1, showing that around 50% of participants in the study had authored at least 1000 posts. The maximum number of posts which could be retrieved for a user was 3000, due to the limitations of the Twitter API.

Based on the intuition that a user would not post content that they are not themselves interested in reading, previous posts are used as a source of information

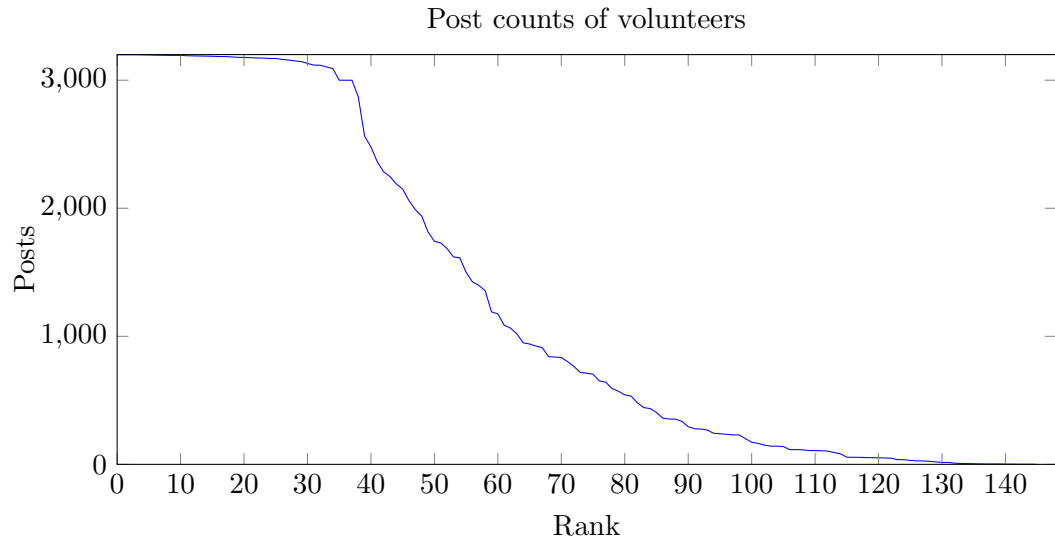


Figure 6.1: Number tweets authored by study participants, ranked by count - around half of the sampled users had written more than 1000 tweets

about reader preferences. These posts are compared to incoming tweets using unigram cosine similarity (IDF weighted with stoplist), in the same way as posts are compared to the centroid in the Centroid method (Radev et al., 2004) and to one another for TextRank (Mihalcea and Tarau, 2004).

This method differs from Centroid in terms of the source from which vocabulary is drawn for comparison. The Centroid method ranks tweets according to their similarity to the tweet set in which they occur, whereas this personalisation method ranks tweets according to their similarity to the reader’s previous tweets. These are very different quantities of tweets - the centroid method uses 50 tweets, and the personalisation method uses up to 3000 posts.

In the Centroid method, identical sets of posts in the timeline will boost the terms that they contain into the centroid for comparison, meaning that they are all more likely to be included in the summary, and due to their identical similarity to the centroid, they will all receive the same ranking score. The repetition of identical posts can lead to redundancy in the summary. This is not possible for the personalisation method, as the ranking is derived from the author’s past tweets and cannot be influenced by sets of identical posts in the collection to summarise.

Retweets and Favourites The retweets made by a user offer a considerable source of data on what a user considers their followers may want to read. Retweets are useful as a predictor of relevance, 73.32% of tweets that are relevant in the study have been retweeted at least once, though a negligible portion of the tweets marked interesting were retweeted by the volunteers themselves.

Favourites fill a similar role to retweets in that they are a form of engagement which demonstrates that the user is interested in the content in question. Unlike retweets, favourite tweets are not rebroadcast on the user’s timeline, though the author is notified of the favourite, so it is not a completely transparent act. Favourites are used for a number of reasons, including as a means to store important tweets for later, a way to express that the reader ‘likes’ the tweets, and as a conversational act in lieu of a response (Gorrell and Bontcheva, 2014).

As with authored tweets, the retweeted and favourited posts of the reader are collected to create a bag-of-words representation, which is compared with incoming tweets using cosine similarity.

User Biography Twitter users are given the ability to create a short biography entry for themselves, which may include a self description or a listing of entries. The user profile also contains information about the twitter users location, usually down to the nearest city, though users often leave this field blank or fill it with nonsensical information (Rout et al., 2013a).

Given that users may choose to express their interests through the text of the Twitter user bio, it is treated as a bag-of-words and compared to the candidate tweets using cosine similarity.

Cosine similarity compares the exact words contained in the tweet with those in the biography. Since both are limited in length, there is a potential that each will use different vocabulary but related to the same topics. Instead of cosine similarity, the sets of synonyms for the words in the biography can be compared to those in the candidate tweet to give a semantic distance measure, using WordNet.

For each term in a candidate tweet, the maximum WordNet similarity (Miller, 1990) between any of its senses and any sense of any term in the user profile is calculated, as defined by Wu and Palmer (1994). Wu-palmer similarity is given by the following definition - the $\text{depth}(C)$ of a concept is its depth in the WordNet taxonomy, and the concept C_{lcs} is the least-common superconcept of C_1 and C_2 - the deepest concept in the hierarchy which is a superconcept of both subconcepts.

Feature	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Followers of author	22.19%	26.38%	25.43%	25.73%
Lists containing author	22.63%	26.95%	25.86%	26.22%
Friends of author/Followers of author	25.88%	29.59%	29.41%	29.26%

Table 6.8: Performance of author popularity features

$$WP(C_1, C_2) = \frac{2 \times \text{depth}(C_{lcs})}{\text{depth}(C_1) + \text{depth}(C_2)} \quad (6.2)$$

The mean of these values for all terms is then used as the ranking score for the tweet. A stop list was used prior to comparison. The normalised maximum WordNet similarity was also used for measuring text similarity in answer grading by Mohler and Mihalcea (2009).

This WordNet semantic distance score used here compares all of the possible synonym sets (word senses) for all of the terms in the biography against all of the terms in the Tweet. For text units longer than Twitter biography fields, this quickly becomes an extremely large number of comparisons. The value of such comparison is also lessened when the documents in question are shorter.

6.4.2 Results

The results for author popularity features on the development portion of the gold standard are shown in Table 6.8. The number of followers of an author was not a useful feature, nor was the number of lists onto which an author had been placed; the scores for these features was given by sorting posts by the value of this feature, highest to lowest. Volunteers did not prefer tweets from very popular authors.

The friends to followers ratio was much stronger as a ranking feature, indicating a preference within the gold standard for authors who follow many over those who are followed by many. Celebrity accounts are the most followed users on Twitter, but they do not generally follow many accounts (Wu et al., 2011). These results therefore show that interesting or relevant results are not those which come solely from celebrity twitter accounts.

Given that users do not strongly prefer posts by celebrity authors, other kinds of relationship are useful in ranking, as shown in Table 6.9. The count of mentions of the author by the reader is the strongest indicator of relevance of this type. Compared to the strongest baseline retweet counts, this feature alone gives a higher score under MAP and a comparable score under ROUGE. This category of interaction en-

Feature	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Mentions of author by reader	28.22%	31.76%	32.58%	31.93%
Mentions of reader by author	26.56%	31.32%	31.27%	31.08%
Replies to author by reader	26.07%	29.34%	29.78%	29.35%
Retweets of author by reader	26.56%	31.11%	31.05%	30.85%
Tweet mentions reader?	23.07%	30.03%	29.42%	29.53%
Friends in common	27.41%	31.45%	30.82%	30.91%
Friendship is reciprocated?	22.89%	29.93%	29.20%	29.38%

Table 6.9: Performance of author reader relationship features

Similarity to:	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All posts	27.85%	31.44%	31.83%	31.43%
Past favourite posts	24.27%	30.47%	30.24%	30.14%
Past retweeted posts	23.97%	29.57%	28.91%	29.02%
Twitter biography	24.42%	30.21%	29.35%	29.61%
Twitter biography (WordNet)	23.55%	27.75%	28.67%	27.96%

Table 6.10: Performance of ranking using reader profiles

compasses both old-style retweets which contain the author’s name and replies to the author’s messages.

The two Boolean features, ‘Tweet mentions reader?’ and ‘Friendship is reciprocated?’ gave promising performance under ROUGE-F but not under MAP. The ranking within each class is random, so this is to be expected. The friends in common feature is slightly weaker than the count of interactions, but is still a useful indicator of relevance.

Both friends in common and the total count of interactions give performance which is stronger than random ranking, indicating that users in general prefer tweets from authors with whom they share a stronger social connection.

Table 6.10 shows results for features derived from the text, retweets, favourites and Twitter biography strings of the volunteer reader. Each feature other than the WordNet comparison gives results which are stronger than random under ROUGE-F and MAP for the development set, and the strongest ‘All posts’, which includes both tweets written by the user and tweets that they have retweeted outperforms the strongest baseline under MAP whilst giving comparable results for ROUGE.

The author popularity and author-reader relationship features do not incorporate the text of the tweet whatsoever; they are entirely based on social network metadata. The scores for ‘All posts’ show that text can be useful in tweet ranking for Twitter home timelines, but that the tweets of interest are not those which best reflect the

content of the tweet set but rather those which best match the reader's general topics of interest. Unlike the author popularity and relationship features, reader similarity scores allow tweets by a single author to be assigned different scores.

6.4.3 Discussion

A number of numerical and binary properties of social network posts have been shown to be useful in the ranking of home timelines by relevance. Where tweets are ranked according to their content, it has been shown to be effective not to prioritise them according to how representative they are of the set to which they belong, but rather according to how closely they match the topics on which the reader often posts. This type of feature assumes that readers are also authors, which is generally the case for social media, but not for other domains.

Given that text information in collections of tweets is fragmented and limited, it is useful that social features can help to overcome these limitations. Nonetheless, the simplest measures of popularity, counts of retweets and of favourites, still show as much utility in personal social media ranking as any other type of social or textual information.

In the remainder of this chapter, social and text-based features are combined automatically using machine learning to create robust models for ranking. These models are personalised for each volunteer, given that users may prefer different tweets for different reasons. The generalisability of the features, algorithms and models will be tested on a held-out test set.

6.5 Machine Learning

For the task of Twitter home timeline ranking, simply counting retweets and ordering tweets by those counts gives strong performance under both MAP and ROUGE-F (See Section 6.2.3). This performance was improved upon slightly by ordering by counts of interactions between author and reader, and nearly matched by counts of friends in common as well as similarity to the reader's own tweets (Section 6.4.2). Each of these scores reflects a different aspect of the tweet, such as its popularity overall, the relationship between its author and the reader, and the relevance of its topic, however tweets can be interesting for many different reasons (Section 3.4), and none of these features can be expected to cover all of these kinds of relevance alone.

Relying on a single element of the metadata from the tweets for ranking can lead to over fitting, where a feature which does well on the development set does not generalise to the test set or to the overall population. Given that a large number of features are considered in this work, this is even more likely.

Machine learning is used in this work to combine multiple diverse features, allowing for rankers which are not limited to just one aspect of the tweet. The trained models are not only more diverse, incorporating both text, popularity and personalisation features, but are likely to be more robust and less prone to over fitting.

6.5.1 Machine Learning Methods

The following machine learning methods are used in feature combination in this work. Models are trained on a number of different features using a training portion of the development set, and then evaluated on a second portion of the development set. A total of 145 documents are used for training, and 121 for testing, giving 7250 examples for training and 6050 for testing. The test set used here is not the overall held out test set. Further details of this division are given in Section 3.3.2.

Gaussian Naive Bayes

This method takes a Bayesian approach to classification. The basic model for classification here is given by Mitchell (1997a) as:

$$C = \operatorname{argmax}_C P(C|x_0, x_1 \dots x_n) \quad (6.3)$$

Where C is a predicted class, and the values x_n are the values of particular features in the instance to classify. In text models, the values for x may be the specific terms in the document, for example. Given the large number of possible combinations of x , it is generally impossible to tabulate every possible probability in this manner, so Bayes' theorem may be applied to split the equation as follows:

$$P(C|x_0, x_1 \dots x_n) = \frac{P(C)P(x_0, x_1 \dots x_n|C)}{P(x_0, x_1 \dots x_n)} \quad (6.4)$$

Since $P(x_0, x_1 \dots x_n)$ does not depend on C , it is the same for every possible class. The priors $P(C)$ can be estimated trivially by counting occurrences of each class. However, the probabilities of $P(x_0, x_1 \dots x_n|C)$ remain intractable, with the probability for each feature dependent on the value of all of the others. For this reason, the 'Naive' assumption is made, whereby all features are assumed to be

independent, and their probability attainable by simply multiplying them. Given this assumption, the class probability can be calculated by the computable.

$$P(C|x_0, x_1 \dots x_n) = P(C)P(x_0|C)P(x_1|C) \dots P(x_n|C) \quad (6.5)$$

This classification method assumes independence between represented features, which is not necessarily true in the general case, but thanks to its probabilistic nature it does not assume a hard decision boundary and may give stronger results for the subjective and thus noisy data in the gold standard (Russell and Norvig, 1995). With Gaussian Naive Bayes the estimated posterior probability for the classification is used as the ranking score. The SciKitLearn¹ implementation of Gaussian Naive Bayes was used in this work.

Support Vector Regression

Support vector machines are classifiers which treat the training data as a series of geometric points, with points belonging to one class or the other. A hyperplane can then be generated based on these points, to separate the data (Burges, 1998). If the data is linearly separable, that is, a single line can be chosen which perfectly dissects the points from the two classes, the hyperplane is chosen to give the greatest possible distance from the nearest points in either class. This separation is referred to as the margin.

In most applications, however, data is not completely linearly separable, either because of noise or because of the probabilistic nature of the attributes. This is dealt with by the introduction of kernels that allow SVM use non-linear decision boundaries, such as the radial basis function (Vert et al., 2004), and by the use of a hyperparameter C , which determines the extent to which points which lie on the wrong side of the decision boundary are penalised. A hyperplane is then chosen to minimise this penalty function.

Similar to support vector machines, support vector regression classifiers fit a decision gradient in many dimensions based on decision margins. The LIBSVM implementation of SVR (Chang and Lin, 2011) was used in this work, and parameter values were selected using grid search, varying C within the range $\{2^N : N \in \{-5 \dots 9\}\}$ and γ for $\{2^N : N \in \{-15 \dots 2\}\}$. Ultimately, SVR and related methods can be expected to perform best when the decision surface, or some lower dimensional function of it, can be separated linearly (Joachims, 1998).

¹<http://scikit-learn.org/dev/index.html>

Decision Tree Regression

Decision trees are processed by which one attribute at time of the data is interrogated. Depending on the value of the attribute, another attribute is then considered, or a classification is made. These decisions are described as a series of rules, arranged in a tree structure (Mitchell, 1997a).

There are many strategies for training decision trees, though most employ a greedy search through the space of possible decisions, which can unfortunately lead to overfitting. The ID3 algorithm, for example, iterates through each feature, determining the extent to which classifying on that feature alone could reduce entropy, a value known as information gain. The entropy of a collection of samples S is:

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (6.6)$$

Where c is the number of classes, and p_i is the portion of the samples belonging to class i . As an attribute is selected, two subsets of samples are created according to its value. The information gain of a split is then given by Mitchell (1997a) for samples S and an attribute A on which to split as:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in A} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (6.7)$$

Where v is a possible value of A (if A is continuous, it must be split into discrete values, using a threshold), and S_v is the subset of S for which the value of A is v . Repeatedly reducing this measure means rules are selected which capture the most information about the data. The Gini index of the data may also be reduced in place of entropy, representing the likelihood of a sample being misclassified if the probabilities apparent in the set of samples are used to randomly assign it to a class (Breiman et al., 1984).

$$\text{Gini}(S) = \sum_{i \neq j} p_i p_j \quad (6.8)$$

In this work, regression trees are used to create continuous values for ranking. Because output values are continuous, a suitable information gain function must be used (minimum squared error, in this thesis), and a continuous output value is generated from the average of the values for all the samples remaining at the terminal node of the tree.

Decision trees have a number of advantages - they can capture functions in which features are heavily dependent on one another (Breiman et al., 1984), and they can be easily inspected by humans to validate their logic. Unfortunately, because they also greedily search a very complex decision space, they can be prone to overfitting, so the depth of the tree as well as the number of features under consideration must be chosen carefully. In other work, collections of randomised decision trees are combined to form random forests (Breiman, 2001), which sacrifice the ease of inspection but help reducing overfitting because of the law of large numbers.

6.5.2 Learning objective

The gold standard data consists of binary judgements, however the function to learn, an estimate of tweet relevance, is continuous. The learning objective is defined based on the particular machine learning method in use.

A classifier can be trained directly on the target values in the gold standard. This will produce a binary classification, in which case the confidence score (for Gaussian Naive Bayes) or the size of the margin (for Support Vector Machines) must be used to create a continuous value for ranking.

For the regression models (Support vector regression and decision tree regression), binary relevance judgements can be used directly or an objective function can be created from the textual similarity between a candidate tweet and the collection of relevant tweets.

In this work, all models are trained on the binary decisions made by annotators, where a tweet in the relevant set is positive and the rest are negative. This gives a training distribution of 8 out of 50. With Gaussian Naive Bayes, the confidence value for the positive class is used as the ranking score. For SVR and Decision Tree Regression, the output regression value is used directly to rank from high to low.

Treating the training input in this way allows all of the models to be used consistently. Using similarity to the gold standard as a regression objective would have implied the assumption that the ideal posts are those which are most textually similar to the gold standard using a specific automatic text comparison algorithm. As was shown in Section 6.4.2, tweet relevance is not solely determined by the text the tweet contains, so this assumption is unlikely to be appropriate here.

6.5.3 Features

The features used for machine learning are those evaluated in Sections 6.2.3, 6.4 and 3.4, with the exception of features which had to be normalised or modified in some way to allow for learning. The features fall broadly under the categories of social media metadata, user profile features, and text based features.

Many features were generated in this work, and not all of them are guaranteed to be useful in ranking. The text based features, especially, were not found to be stronger than baselines under individual experiments (Section 6.3).

The remainder of this section will present a recap of the key features which will be supplied to the machine learning methods for social media timeline ranking. They have been separated into social, profile and text-based features, and any normalisation of the features which is necessary for learning will be described here.

Social Features

The social media metadata features include both features from Section 6.4 on ranking using social information, and several of the baseline values which are also unique to social networks. The baseline features used for learning are as follows:

- Age when accessed
- Count of favourites
- Count of retweets

The baselines for age, favourite and retweet count are included in this category. The age of the tweet is given in seconds as a learning feature. Retweet counts follow an exponential distribution (see Figure 6.2), so they are substituted with $\log(n + 1)$ to reduce their range and to aid in learning from them. An additional tweet popularity feature is introduced for machine learning for the same purpose of normalising retweet counts, wherein retweet counts are divided by the number of followers of the author:

- Retweets of Post/Followers of Author

Author popularity features, which are useful in differentiating the status of authors as celebrities, are used without modification (see Section 6.4.1).

- friends/followers of author

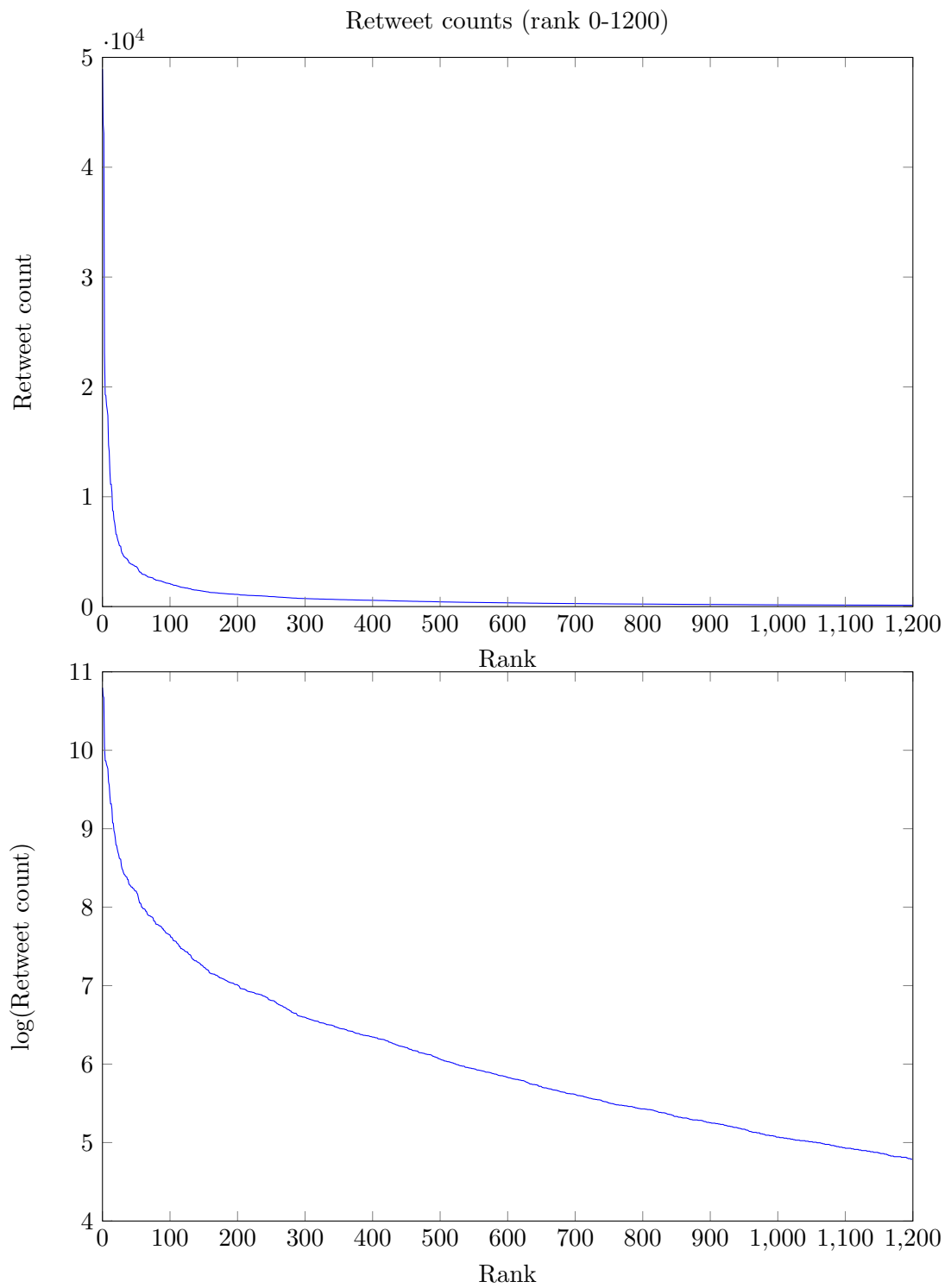


Figure 6.2: Ranked Retweet Count

- Number of followers of the author
- Number of lists on which the author appears

The remaining social features describe the strength of the relationship between the author and the reader (see Section 6.4.1). This requires that the identity of the author is known. Several of these features are very strongly dependent. The number of times the reader has mentioned the author incorporates the counts of retweets and replies, for example. These interdependent features are expected to have a negative impact on the Gaussian Naive Bayes model, which assumes independence (Russell and Norvig, 1995).

- Does author follow reader?
- Does the tweet mention the reader?
- Number of friends in common with author
- Number of times the author has mentioned the reader
- Number of times the author has retweeted the reader
- Number of times the reader has mentioned the author
- Number of times the reader has replied to the author

An evaluation of these social features in isolation can be found in Section 6.4.

Profile features

These features compare incoming posts to social media profile through the posts they have created, retweeted or favoured and incorporate the relevance of candidate tweets to these. Cosine distance is used to calculate the similarity of the tweet to the following text sources:

- Reader's own posts
- Reader's favourite posts
- Posts retweeted by the reader
- Reader's Twitter 'biography'

Additionally, WordNet similarity for the reader's biography is added, as the biography is extremely short and likely to suffer significant vocabulary mismatch, as described in Section 6.4.1.

Text Features

A subset of the textual features evaluated previously are used as features for machine learning. These features include centroid with IDF, as well as a variant of MMR, taking the ranking scores produced by these methods as features for learning.

6.5.4 Use of Features in Related Work

A full list of features used for training, along with any related work from Twitter summarisation which utilises them, is given in Table 6.11. The most common features are those which rely solely on the social nature of tweets, with the text-based features appearing less frequently in the related work. Features for which no related work is given in the table are, to our knowledge, unique to this thesis. These unique features include the use of past favourites as indicators for interests, the summary scores of the tweet given by centroid and MMR, and the specific ratios used to capture post and author popularity.

6.5.5 Feature Selection

Univariate feature selection is carried out prior to learning, heuristically selecting features which are likely to provide strong classification boundaries (Saeys et al., 2007). Univariate feature selection was chosen for its speed and simplicity, however it does not select on the basis of dependencies between pairs of features.

Given that several candidate features are clearly dependent upon one another, and Naive Bayes assumes that features are independent (Russell and Norvig, 1995), it would have been appropriate to use multivariate feature selection which captures such dependencies. Multivariate selection was not used, however, because it is much slower than univariate selection, and because we also include the decision tree regression model for learning, which can build models which utilise relationships between features and would not benefit from independence.

The number of selected features is referred to as K , and the performance for a number of different values for K is tested and reported for each model.

Feature	Related Work
Retweet count	Yan et al. (2012); Duan et al. (2010); Chen et al. (2012); Suh et al. (2010); Uysal and Croft (2011); Hong et al. (2013)
Favourites count	–
Tweet age	Ren et al. (2013); Huang et al. (2011)
Retweets of Post/Followers of Author	–
Centroid (IDF, lower case)	–
MMR (IDF, $\lambda = 0.5$)	–
Followers of author	Ren et al. (2013); Huang et al. (2011); Chen et al. (2012); Feng and Wang (2013); Hong et al. (2013)
Lists containing author	–
Friends of author/Followers of author	–
Mentions of author by reader	Chen et al. (2012); Feng and Wang (2013); Uysal and Croft (2011); Hong et al. (2013)
Mentions of reader by author	Uysal and Croft (2011)
Replies to author by reader	–
Retweets of author by reader	Feng and Wang (2013); Uysal and Croft (2011); Hong et al. (2013)
Tweet mentions reader?	Feng and Wang (2013); Uysal and Croft (2011)
Friends in common	Ren et al. (2013); Hong et al. (2013); Feng and Wang (2013)
Friendship is reciprocated?	Chen et al. (2012); Uysal and Croft (2011); Hong et al. (2013)
Biography similarity (cosine)	Mackie et al. (2014); Feng and Wang (2013)
Biography similarity (WordNet)	–
Similarity to retweets	Hong et al. (2013)
Similarity to favourites	–
Similarity to reader posts	Mackie et al. (2014); Chen et al. (2012); Feng and Wang (2013); Uysal and Croft (2011); Hong et al. (2013)

Table 6.11: Features and their use in related work

Feature	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Random Ranking	22.71%	30.21%	28.75%	29.24%
Retweet count	25.43%	32.03%	30.05%	30.83%
Favourites count	25.82%	32.23%	31.36%	31.56%
Tweet age	23.72%	28.60%	27.65%	27.87%
Retweets of Post/Followers of Author	27.00%	34.14%	31.46%	32.57%
Centroid (IDF, lower case)	22.53%	27.38%	25.68%	26.33%
MMR (IDF, $\lambda = 0.5$)	22.92%	34.69%	30.85%	32.39%
Followers of author	22.48%	26.96%	25.69%	26.15%
Lists containing author	23.45%	27.70%	26.54%	26.92%
Friends of author/Followers of author	24.76%	28.93%	28.84%	28.62%
Mentions of author by reader	28.10%	31.13%	32.12%	31.40%
Mentions of reader by author	25.91%	31.78%	31.23%	31.28%
Replies to author by reader	25.74%	30.00%	30.12%	29.85%
Retweets of author by reader	26.28%	30.54%	30.57%	30.31%
Tweet mentions reader?	22.24%	31.05%	29.99%	30.31%
Friends in common	26.41%	32.94%	32.27%	32.34%
Friendship is reciprocated?	22.13%	31.18%	29.95%	30.35%
Biography similarity (cosine)	23.32%	31.30%	30.17%	30.54%
Biography similarity (WordNet)	23.69%	33.42%	27.77%	30.12%
Similarity to retweets	24.93%	31.60%	30.07%	30.56%
Similarity to favourites	21.71%	31.18%	29.95%	30.35%
Similarity to reader posts	27.74%	32.94%	33.57%	33.06%

Table 6.12: Feature evaluation scores for learning subset

6.5.6 Results

The performance scores of individual features reported elsewhere in this work are for the entire development set; for machine learning, the development set is subdivided further into training and testing (See Section 3.3.2). Comparison between trained rankers and individual features can assist in understanding the quality of the trained models, and provide a sanity check, since a useful model should outperform the score for any single feature on which it is learned. The results for individual features on just the testing part of the development set are shown in Table 6.12. Random ranking is also included here for comparison, but is not given as a feature during learning.

The testing portion of the development set is smaller than the entire development set used in the rest of this chapter, so some variation is observed in the relative scores of the rankers under test. Whilst random ranking still outperforms many methods under ROUGE, including Centroid scores, the popularity baselines have moved in

Selected Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	28.41%	33.33%	33.11%	32.93%
2	27.60%	33.10%	31.77%	32.27%
3	27.82%	33.10%	31.95%	32.36%
4	28.27%	33.42%	32.29%	32.66%
5	27.88%	32.60%	31.38%	31.81%
7	29.46%	34.51%	33.93%	34.03%
10	28.39%	33.26%	32.82%	32.78%
15	28.72%	34.46%	34.02%	33.94%
20	28.41%	33.33%	33.11%	32.93%

Table 6.13: Performance results for Gaussian Naive Bayes machine learning

the ranking, with favourite counts more successful here than retweet counts. Since a portion of the data does not include favourite counts, the performance for favourite counts will change as the gold standard is sampled. The counts of mentions of author by reader, and reader author post similarity are both strong features here, once again. In this test set of the data, ordering by tweet age is worse than ordering randomly.

The testing portion of the development set is much smaller than the entire development set, and these tweet sets are not distinct from the development results, so on their own they should be considered less accurate than results based on the entire development set, rather than as an estimate of generalisability. The results for individual features are reported here for this smaller set, so that they may be compared with the results for the machine learning methods.

Table 6.13 shows the machine learning results for Gaussian Naive Bayes. The quality of the learned model does depend on the number of features selected prior to learning, with the best results given for 7 selected features, giving improved performance over single feature values for both MAP and ROUGE. Though the improvements seen here are relatively small, they show that machine learning can lead to strong results when the scores are viewed both as rankings and summaries.

The strongest Gaussian Model was that which used 7 features. The features selected in this instance were as follows:

- Count of Favourites
- Count of Retweets
- Friends in Common

Features	C	γ	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	2^9	2^{-13}	24.62%	31.25%	30.00%	30.41%
2	2^{-3}	2^{-13}	28.24%	33.89%	32.29%	32.90%
3	2^0	2^{-11}	28.94%	34.44%	32.63%	33.32%
4	2^{-4}	2^{-9}	29.33%	34.57%	33.30%	33.74%
5	2^1	2^{-12}	29.12%	34.46%	32.88%	33.47%
7	2^{-5}	2^{-11}	27.01%	31.37%	29.94%	30.39%
10	2^{-5}	2^{-11}	26.63%	31.51%	30.02%	30.52%
15	2^{-2}	2^{-12}	25.36%	29.89%	29.29%	29.41%
20	2^{-2}	2^{-2}	24.37%	31.25%	30.00%	30.41%

Table 6.14: Performance results for Support Vector Regression (RBF kernel)

- Mentions of Reader by Author
- Retweets of Author by Reader
- Number of lists containing Author
- Biography similarity (cosine)

Many of these features are not the strongest under individual evaluation. The criteria used in univariate feature selection is not the same as that used for evaluation of rankings. Of the features used here, only two, the retweet count and favourite counts, are not personalised to the individual author. The similarity of the user biography to the tweet is the only text-based feature used here.

The results for Support Vector Regression are shown in Table 6.14. The stopping tolerance was set at 0.1 and the maximum number of iterations at 15000. A radial basis function (RBF) was used for the kernel. Values shown are the strongest attained during grid search, which was carried out once for each feature set. The strongest SVR score, where 4 feature are used, is close to the value for Gaussian Naive Bayes. Although the results on the development set are similar, the Gaussian Naive Bayes is less prone to over fitting than Support Vector Regression here as it forms a model with more features, a much simpler parametric representation and no hyper parameters. The following 4 features were selected for this model:

- Count of Favourites
- Count of Retweets
- Friends in Common

- Tweet mentions reader?

None of the features used here depend on the text of the tweets, indicating that this model would be able to rank given very small amounts of information about the tweet.

The results for decision tree regression are shown in Table 6.15. When less than 15 and more than 2 features are used, the results do not appear to vary much for different counts of features and different tree depths. The best performance for decision trees is given with either 20 or all of the features, but a depth of just 2, though these MAP and ROUGE scores are weaker than SVM and Gaussian Naive Bayes models. Decision tree regression will select features on which to classify at each depth of the tree, and is strongest in this case when using features which would not be chosen by univariate feature selection.

Decision trees are capable of learning arbitrarily complex functions, though at a depth of just two the model learned in this instance is extremely simple.

6.5.7 Discussion

Single value features can give strong performance for home timeline ranking. The baseline value, retweet counts, was comparable to any other method used in this work. Other such values, such as counts of previous interactions between author and reader, also gave performance which was at or above the level of the strongest baseline.

Using a single feature such as retweet counts for ranking, however, does not allow the ranker to capture the notion that tweets can be relevant for a number of different reasons (Section 3.4), and the appropriate features for ranking intuitively should depend on this reason. For example, retweet counts cannot really be expected to capture whether or not a tweet is relevant to the author's interests. Since a very large number of features are evaluated, one would expect at least a few features to give higher performance figures by chance, assuming that the performance in the sample is a noisy representative of the true performance of the method.

The use of machine learning allows the combination of many and distinct types of information. The scores from algorithms which compare the user's own tweets with candidate posts can be weighed against information about the popularity of that item. The combination of text, popularity, relationship and user profile features as shown in this section will create a more diverse classifier, which, unlike single feature

Features	Depth	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	1	25.16%	32.07%	30.10%	30.87%
All	2	28.04%	33.77%	33.05%	33.21%
All	3	27.35%	33.63%	32.97%	33.07%
All	4	27.31%	33.78%	32.85%	33.09%
All	5	26.87%	33.59%	32.55%	32.78%
All	6	25.65%	31.50%	30.35%	30.66%
All	7	26.13%	31.45%	30.82%	30.89%
2	1	24.52%	31.07%	30.04%	30.34%
2	2	23.90%	31.60%	30.75%	30.97%
2	3	25.76%	33.69%	32.92%	33.07%
2	4	26.46%	33.21%	31.73%	32.22%
2	5	26.63%	31.65%	30.39%	30.81%
2	6	25.72%	31.63%	30.57%	30.84%
2	7	26.13%	32.04%	31.18%	31.38%
3	1	23.79%	31.07%	30.04%	30.34%
3	2	23.73%	30.73%	29.83%	30.07%
3	3	24.01%	31.32%	30.66%	30.79%
3	4	26.18%	32.05%	31.29%	31.47%
3	5	26.70%	32.74%	31.14%	31.72%
3	6	26.22%	31.48%	30.22%	30.61%
3	7	26.68%	32.52%	30.96%	31.50%
4	1	22.70%	31.07%	30.04%	30.34%
4	2	24.29%	30.73%	29.83%	30.07%
4	3	24.83%	31.32%	30.66%	30.79%
4	4	24.92%	32.15%	31.44%	31.59%
4	5	24.69%	31.21%	30.70%	30.71%
4	6	26.28%	32.19%	30.76%	31.23%
4	7	26.82%	31.99%	30.90%	31.21%
5	1	24.15%	31.07%	30.04%	30.34%
5	2	23.22%	30.73%	29.83%	30.07%
5	3	23.18%	30.81%	30.00%	30.20%
5	4	24.52%	31.11%	30.53%	30.62%
5	5	24.72%	31.11%	30.74%	30.69%
5	6	25.74%	31.46%	30.17%	30.57%
5	7	25.86%	32.17%	31.03%	31.35%
7	1	24.62%	31.07%	30.04%	30.34%
7	2	24.52%	31.06%	30.25%	30.44%
7	3	23.69%	31.23%	30.58%	30.69%
7	4	24.45%	30.69%	30.19%	30.23%
7	5	24.70%	30.47%	29.78%	29.92%
7	6	25.70%	31.95%	30.69%	31.10%
7	7	26.26%	31.86%	30.98%	31.20%
10	1	25.09%	32.07%	30.10%	30.87%
10	2	25.28%	32.33%	30.92%	31.40%
10	3	25.10%	32.27%	31.62%	31.73%
10	4	26.54%	31.64%	31.17%	31.20%
10	5	26.71%	31.73%	30.70%	30.95%
10	6	26.06%	31.82%	30.93%	31.13%
10	7	26.64%	31.98%	30.68%	31.09%
15	1	24.47%	32.07%	30.10%	30.87%
15	2	27.81%	33.77%	33.05%	33.21%
15	3	26.89%	33.63%	32.97%	33.07%
15	4	27.53%	33.78%	32.85%	33.09%
15	5	26.61%	33.70%	32.71%	32.91%
15	6	26.63%	32.20%	31.25%	31.43%
15	7	26.27%	32.57%	32.07%	32.05%
20	1	24.90%	32.07%	30.10%	30.87%
20	2	28.59%	33.77%	33.05%	33.21%
20	3	27.50%	33.63%	32.97%	33.07%
20	4	27.24%	33.78%	32.85%	33.09%
20	5	26.63%	33.59%	32.55%	32.78%
20	6	25.57%	31.50%	30.35%	30.66%
20	7	26.18%	31.45%	30.82%	30.89%

Table 6.15: Performance results for decision tree regression

values, might be expected to generalise beyond the development set, and to detect many different types of tweet relevance.

The strongest model from this section was Gaussian Naive Bayes, though there was not much difference in the strongest performance for other models. Gaussian Naive Bayes is a simple, easy to train probabilistic model which is robust to noise. The algorithm assumes that individual features are independent, though they are clearly not all independent in this case. Performance was still strongest for Naive Bayes despite this violated assumption.

The results for Support Vector Regression and Decision Tree Regression were very similar to Gaussian Naive Bayes, though they were both slightly weaker on the development set. Unlike Gaussian Naive Bayes, which incorporated 7 features in the best case, these two models incorporate fewer features at the strongest performance. Coupled with the complexity of the functions which can be learned by both models, the use of so few features may also harm their generalisability.

The models are sensitive to feature selection. For Gaussian Naive Bayes and Support Vector Regression, univariate feature selection (Saeys et al., 2007) lead to large increases in performance, though some of the original supplied features were very weak under both MAP and ROUGE. The strongest feature sets were largely social information, such as counts of friends in common and counts of favourites or retweets. The only text-based feature to be selected by a best performing model was the similarity to the user's Twitter biography.

In this section, machine learning models have been constructed which exceed the MAP and ROUGE scores for both baselines and social media feature values. These models will be personalised in Section 6.6, training one model per volunteer, where volunteers produced multiple data sets. Generalisability for the Gaussian Naive Bayes and SVR models will be evaluated on the held out test set and results will be given in Section 6.7.1.

6.6 Personalised Machine Learning Experiments

The timelines in the gold standard data set have been generated using a different list of Twitter users for each participant. Twitter users curate a list of authors to 'follow', based on their authors of interest as well as their social connections, exhibiting a degree of homophily with the users they follow (Kwak et al., 2010).

Many of the social ranking features described in Section 6.4 are personal to the reader, including their relationship with the author and the similarity of a tweet to

the reader's previous posts. Features like these have been combined with machine learning models, to give classifiers which are intended to be more robust than individual features (Section 6.5), but these models are universal because they assume that the appropriate weighting between features does not vary from user to user.

Given that personalisation is considered important in many kinds of Twitter summarisation (Kapanipathi et al., 2011; Ren et al., 2013; Feng and Wang, 2013), and that performance results on the ranking task for any single method can vary enormously from tweetset to tweetset (Section 3.3.1), the models which had previously been trained for the entire collection are additionally trained for individual users.

The experiments described in Section 6.5, including producing individual feature scores and scoring trained models, but each algorithm is trained only on prior data for the user upon which it is evaluated.

6.6.1 Data

Only timelines of volunteers that have taken part in the study several times are considered, and of those users, only the ones that had tweet sets present in both the training and testing subdivision of the development set are included in the gold standard here. Overall, this gives 76 training documents and 62 testing documents, a smaller number than the development set as a whole. The amount of data available for training for a single user is very small and is rarely more than a single document per volunteer.

6.6.2 Results

Since the test dataset has changed in this work from that in the previous machine learning experiments, individual feature scores must once again be re-evaluated to allow for comparison to learned models.

The relative performance of features is relatively unchanged, though the average scores under ROUGE-F are higher for this smaller set. As before, these performance figures (Table 6.16) are intended as a baseline for the learning task and they are not to be considered more accurate than those for the entire development set, nor do they show generalisation performance since the sets are not distinct.

In the setting of personalised machine learning models, the Gaussian naive Bayes learner which gave the best performance improvements without personalisation was not an improvement over any single feature (see Table 6.17).

Feature	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Retweet count	25.69%	31.74%	29.91%	30.66%
Favourites count	25.90%	33.22%	33.22%	33.07%
Tweet age	23.37%	27.53%	27.68%	27.45%
Retweets of Post/Followers of Author	27.29%	34.73%	32.29%	33.31%
Centroid (IDF, lower case)	23.30%	27.50%	26.27%	26.71%
MMR (IDF, $\lambda = 0.5$)	22.46%	33.02%	30.83%	31.73%
Followers of author	20.97%	25.80%	25.24%	25.39%
Lists containing author	21.73%	27.57%	26.80%	27.06%
Friends of author/Followers of author	25.84%	30.73%	31.34%	30.81%
Mentions of author by reader	25.78%	29.21%	30.76%	29.82%
Mentions of reader by author	26.22%	32.36%	32.39%	32.22%
Replies to author by reader	25.71%	29.69%	29.97%	29.67%
Retweets of author by reader	23.39%	28.82%	29.52%	28.99%
Tweet mentions reader?	20.84%	31.78%	32.02%	31.75%
Friends in common	29.72%	35.61%	35.47%	35.37%
Friendship is reciprocated?	24.88%	32.16%	32.15%	32.03%
Biography similarity (cosine)	22.85%	31.11%	30.58%	30.70%
Biography similarity (Word-Net)	23.72%	30.57%	31.64%	30.95%
Similarity to retweets	26.94%	33.01%	32.90%	32.76%
Similarity to favourites	22.43%	32.16%	32.15%	32.03%
Similarity to reader posts	27.95%	34.33%	35.51%	34.70%

Table 6.16: Performance of individual machine learning features on personalisation data set

Selected Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	24.75%	32.52%	32.85%	32.49%
2	27.73%	32.14%	32.06%	31.92%
3	28.50%	33.14%	32.87%	32.84%
4	28.41%	34.30%	33.68%	33.77%
5	28.48%	33.47%	32.82%	32.90%
7	30.72%	32.88%	32.82%	32.70%
10	30.09%	34.99%	34.75%	34.71%
15	30.40%	34.36%	34.12%	34.11%
20	24.47%	32.52%	32.85%	32.49%

Table 6.17: Performance results for personalised Gaussian Naive Bayes machine learning

Selected Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	22.80%	32.23%	32.19%	32.08%
2	25.47%	30.22%	30.58%	30.23%
3	27.90%	32.55%	32.48%	32.34%
4	25.82%	32.88%	32.81%	32.66%
5	27.28%	31.57%	31.25%	31.24%
7	27.59%	31.76%	31.66%	31.54%
10	23.36%	31.28%	31.33%	31.15%
15	24.68%	31.44%	31.52%	31.33%
20	21.12%	32.23%	32.19%	32.08%

Table 6.18: Performance results for personalised Support Vector Regression (RBF kernel)

As was the case when training a single model for the whole training set, the SVR learner (Table 6.18) was weaker than the Naive Bayes model (Table 6.17). In this case, the size of the training set for each learner was extremely small, and as a result the performance decrease in comparison to Gaussian Naive Bayes was even larger.

The strongest personalised decision tree learner in Table 6.19 was able to match the performance of the best single feature under ROUGE. Decision trees are very expressive and can easily over fit, a problem which is especially likely on small, noisy data sets such as the individual user preferences gold standard.

6.6.3 Discussion

These models were trained for each individual user in the hope that their individual preferences for certain kinds of tweets could be learned automatically. Unfortunately, the experimental setup was not originally designed for this case, and for many users the data available for training was extremely small. Given that some of the users have only marked 8 tweets as interesting, and more than 8 features are incorporated into many of the models, the learning step is very short.

Where features are noisy, which they are by nature for such a subjective task, the amount of data present is unlikely to be enough to overcome this noise in training. Indeed, the observed performance for these models was no stronger than the ranking given by any single feature.

This method of personalisation is a simple setup designed to evaluate whether or not this type of learning might be feasible. The performance of these personalised models might be improved, however, by replacing this training setup with one of domain adaptation (Daumé and Marcu, 2006), where a single model is trained for

Features	Depth	MAP	ROUGE-R	ROUGE-P	ROUGE-F
All	1	24.24%	32.88%	32.25%	32.39%
All	2	24.10%	32.81%	32.29%	32.42%
All	3	26.64%	33.26%	33.41%	33.14%
All	4	25.61%	33.02%	33.31%	32.97%
All	5	26.37%	35.68%	35.61%	35.42%
All	6	24.99%	34.87%	34.82%	34.65%
All	7	25.31%	34.53%	34.48%	34.35%
3	1	24.75%	32.37%	32.29%	32.16%
3	2	25.07%	33.46%	33.31%	33.24%
3	3	25.86%	32.23%	32.10%	32.00%
3	4	27.29%	32.72%	32.17%	32.31%
3	5	28.10%	34.27%	32.99%	33.48%
3	6	26.93%	32.57%	31.57%	31.91%
3	7	27.43%	32.62%	32.01%	32.15%
5	1	27.02%	32.84%	32.21%	32.34%
5	2	27.85%	34.37%	33.75%	33.86%
5	3	27.27%	34.03%	33.04%	33.35%
5	4	27.09%	33.42%	33.06%	33.06%
5	5	27.07%	33.43%	33.48%	33.28%
5	6	25.92%	32.69%	32.31%	32.33%
5	7	26.79%	33.58%	33.14%	33.17%
7	1	25.72%	33.55%	33.15%	33.18%
7	2	28.60%	34.63%	34.56%	34.39%
7	3	25.83%	30.99%	30.26%	30.47%
7	4	26.22%	32.45%	32.03%	32.10%
7	5	25.97%	31.84%	31.80%	31.64%
7	6	27.22%	33.25%	32.98%	32.96%
7	7	27.13%	33.96%	33.40%	33.50%
10	1	25.78%	34.15%	33.52%	33.64%
10	2	27.09%	35.00%	34.40%	34.50%
10	3	26.85%	32.69%	32.43%	32.38%
10	4	26.07%	32.08%	31.77%	31.79%
10	5	24.12%	30.43%	30.20%	30.15%
10	6	26.53%	31.83%	31.77%	31.62%
10	7	25.56%	31.43%	31.69%	31.40%
15	1	27.06%	34.11%	33.57%	33.65%
15	2	25.51%	33.98%	33.29%	33.46%
15	3	26.35%	33.05%	33.34%	32.97%
15	4	26.14%	32.50%	32.44%	32.31%
15	5	24.31%	32.16%	32.10%	31.96%
15	6	23.70%	31.66%	31.95%	31.61%
15	7	25.43%	31.34%	31.58%	31.28%

Table 6.19: Performance results for personalised decision tree regression

all users and then adapted for the individual participants. This avenue for future work should lead to performance that is better than that of training a single model, and would allow learners to train on the much more reasonably sized data whilst also reflecting individual preferences.

6.7 Discussion

The experiments in this chapter replicate the methodology of those in Chapter 5, and extend them, using features which are available for most Twitter users for ranking. Additional features include social information about the relationship between the reader and author, information about the possible celebrity status of the author, and a comparison between candidate tweets for ranking and the tweets that had been produced by the volunteer reader previously.

Machine learning models have been proposed which outperform individual features on the development set. Personalised ranking by training one ranking model for each volunteer was evaluated. The machine learning models are expected to be more robust and to generalise to other data sets more effectively than any single feature.

Many of the features and text-based algorithms that have been evaluated throughout this work do not give better performance according to ROUGE-F than ranking by the counts of retweets. Combining retweet counts with other sources of social information was shown to improve upon retweet counts for the development set using Gaussian Naive Bayes (Russell and Norvig, 1995), Support Vector Regression (Joachims, 1998) and Decision Tree Regression (Breiman et al., 1984).

6.7.1 Validation on Unseen Data

The results elsewhere in this chapter were calculated for the development portion of the gold standard, which was split into training and testing subsets. An additional test set of 121 documents was held out, and has not been used elsewhere in this chapter. This set can now be used to test selected algorithms, demonstrating their ability to generalise before the development set alone.

Given that a large number of individual features and algorithms were tested on the development set, it is impossible to tell from this data alone whether strong performance here is a result of genuine differences in performance, or just chance.

A limited number of key baselines and approaches are evaluated on the testing

Ranking Method	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Retweets Count	27.21%	32.40%	30.96%	31.50%
Favourites Count	26.86%	30.24%	30.60%	30.25%
Random	22.03%	27.62%	26.50%	26.89%
Hybrid TF.IDF	22.08%	28.24%	29.96%	28.84%
TextRank (unigram + IDF)	22.03%	27.04%	27.98%	27.26%
MMR ($\lambda = 0.25$)	23.10%	30.85%	28.15%	29.29%
Friends in Common	25.87%	28.15%	28.15%	27.95%
Similarity to reader posts	26.00%	28.55%	28.67%	28.36%
SVR (4 features, $C = 2^{-4}$, $\gamma = 2^{-10}$)	27.09%	29.63%	29.20%	29.25%
Gaussian Naive Bayes (7 features)	30.10%	31.66%	32.61%	31.91%

Table 6.20: Performance of Selected Ranking Methods on Held-out Testing Set

set, calculating MAP and ROUGE as before, but also the Wilcoxon signed-rank test significance of the differences between pairs of approaches.

Table 6.20 shows the results for selected rankers on the held out testing set. As in other experiments, retweet counts and favourites counts are both strong baseline values. Favourites are still weaker than retweet counts according to MAP and ROUGE, and the difference in performance between retweet counts and random ordering was statistically significant, according to both MAP ($p = 0.001$) and ROUGE-F ($p = 0.02$).

The strongest single feature from the development set was the count of mentions of the author by the reader. This measure is one of several which was intended to represent the strength of their social connection. While ranking according to the value of this feature gave performance comparable to the strongest baseline, retweet counts, in the development set, scores are worse under both ROUGE and MAP for the testing set.

The similarity to reader posts feature also measured worse on the testing set than the development set. There was not a large difference between the mean length of reader timelines in both sets (1397 tweets in development as opposed to 1298 in testing), indicating that the high performance seen on the development set was a potential fluke resulting from the number of features in the comparison.

Maximal Marginal Relevance was selected for comparison to the strongest performing of the methods which prioritised posts based on their textual relation to the others in the set. MMR generates summarisation that contain a diverse selection of

the units from the overall set, penalising redundancy in the generated summaries. MAP does not penalise redundancy as much as ROUGE, and as such the ROUGE scores for MMR on both the development and test set are much stronger than the MAP scores. Nonetheless, even the ROUGE score for MMR is not significantly different from ordering posts at random ($p = 0.19$).

The best performing method on the test set was Gaussian Naive Bayes, which was found to give results for ROUGE which were not significantly different to ranking by retweet count ($p = 0.72$) whilst attaining a ranking score under MAP which was much higher ($p = 0.025$). Whilst it is difficult to measure the effect size, as the results for average precision do not necessarily follow a normal distribution, it appears that the MAP with these learned features is nearly 3 percent-points higher than that for retweet counts alone. The features selected automatically during training for Gaussian Naive Bayes were:

- Count of favourites
- Count of retweets
- Number of lists containing tweet author
- Friends in common
- Does the tweet mention the reader
- Retweets of author's posts by reader
- Similarity of tweet to reader posts.

Many of these features are personalised, based on the relationship between the author or the reader's own past posts and how they relate to the candidate tweet. The remaining features (favourites, retweets and author lists) are all representations of the popularity of a post.

6.7.2 Discussion

In the specific use case driven summarisation of the social media analysis organisation presented in Chapter 5, the best performing method was Centroid combined with topic models for preprocessing. Their use case was one of information discovery, and presenting posts which were representative of the collection as a whole was an appropriate way to support the use case.

In this chapter, however, home timeline ranking was attempted on general Twitter users. Many of these users were not professionals, and none of them had revealed a specific use case. Whilst it is possible that at least some of the users engaged in a social media monitoring process like SORA, many also prioritised tweets which were simply funny or surprising (Section 3.4).

Upon re-evaluating the same methods which performed well on the use case driven summarisation task with the general gold standard, it became clear that the same methods were not appropriate. Centroid no longer outperformed the strongest baselines values and in fact achieved scores which were worse than ordering tweets at random (Section 6.3). Likewise TextRank, BM25 and MMR were all far less effective on the general data. On the test set, the strongest of the methods based purely on the text of the candidate tweets, MMR, did not perform significantly better than ordering randomly.

The counts of favourites and retweets of posts, which were easily surpassed on the SORA dataset, formed very strong baselines in the general case. Though retweets are considered as a gold standard in other work (Uysal and Croft, 2011; Hong et al., 2013; Feng and Wang, 2013), they have a number of limitations, such as that they take time to appear on new posts, as discussed in Section 3.5. Nonetheless, their strong baseline performance on the general tweets data set suggested that more such features might be created to exploit the metadata and structure of the tweets and the social graph to improve upon text based ranking.

A number of different social media features were developed to capture the popularity of the author, the relationship between the author and reader, and the thematic relationship between the tweet and the reader's own profile (Section 6.4). These features are unique to social networks in that they require the existence of relationships between pairs of users and also that the reader has a history of producing their own posts. Many of the single features in this case could be used for ranking on their own to match or outperform baselines on the development set.

Since a great number of features were evaluated in this manner, the scores for individual features could not be distinguished from random chance. Given that a tweet ranker which relies exclusively on the number of past interactions between the author and reader is unlikely to be very robust, machine learning was used to combine a number of features. On the test set it was shown that a relatively simple model, Gaussian Naive Bayes, could significantly outperform the strong retweet count baseline in MAP, without sacrificing any performance in ROUGE.

Both of the key hypothesis of this work were under test in this chapter. The

former hypothesis, that text-based summarisation and information retrieval methods can be used to summarise home timelines (Section 1.2.2), was not demonstrated in the general case, despite promising results for political timelines (Chapter 5). Even with modification, these methods do not produce summaries which outperform key baselines in the general case.

This chapter has demonstrated that personalisation features such as information about the social graph and the user profile are more appropriate for Twitter home timeline than their counterparts which are not personalised whatsoever, and which rely solely on the text of the collection of tweets. This was hypothesis 2 (Section 1.2.4), which has been demonstrated in this chapter through statistically significance improvements in home timeline ranking task. It has also been shown that classifiers which combine several such features using machine learning are more generalisable than any single feature.

The improvements over retweet counts are modest but significant, however in other work retweet counts were used as the gold standard for relevance themselves (Uysal and Croft, 2011; Hong et al., 2013; Feng and Wang, 2013). Outperforming retweets, even in a limited manner demonstrates again that they are not appropriate as a gold standard. When a number of features are used in combination within a Gaussian Naive Bayes model, statistically significant overall improvement in ranking performance is achieved.

Chapter 7

Conclusion

The task of Twitter home timeline summarisation is difficult because, like many other kinds of tweet summarisation, it is subjective (Alonso et al., 2013) and because readers find posts to be interesting for a variety of different reasons (Section 3.4). Nonetheless, it is a task which has value. When asked about their preferences, users indicated that on average just 16% of tweets from their timeline were interesting to them (Section 3.3.1). Automatically summarising collections of tweets allows users to discover content of interest without manually triaging their entire home timeline.

In this thesis, the summaries were produced through the ranking of tweets, similar to other work on extractive tweet summarisation (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013). Summaries could also have taken the form of extended paragraphs, and this would have presented a different set of challenges, including how to transform tweets to form grammatical sentences in such paragraphs. The ranking version of the task, however, simplifies evaluation greatly by allowing tests to disregard further subjective notions of grammatically, coherence, and the quality of the generated text.

Rather than evaluating timeline summaries on the number of retweets posts had garnered, as in other work (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013), this thesis argued that the limited context in which retweets are applicable makes them inappropriate for use as a gold standard in home timeline ranking.

The idea that retweets are unsuitable here was further demonstrated empirically as a number of manually annotated gold standards were created. Had retweets been used as a ground truth for relevance, 26.68% of interesting tweets would never have been considered, since they had never been retweeted by anyone, and, since 0.01% of tweets in the study had been retweeted by the participant themselves, none of the

interesting tweets would have been considered interesting for the users in this work. Many uninteresting tweets would also have been assumed interesting, since 72.50% of tweets that were not interesting to the volunteers had retweeted - the difference in retweet counts for interesting and uninteresting content was very small.

In total, three reader-annotated data sets were generated for the evaluation of Twitter home timeline ranking, including a pilot study, a use-case driven study and a large, generalised annotation task completed by volunteers from the university of Sheffield. These datasets will be shared with the research community.

Retweet counts are not suitable as a gold standard, though they did form a strong baseline in several tasks throughout the work. Investigating the first hypothesis, which is that the selected summarisation and information retrieval methods that have been evaluated here are effective on the task of tweet ranking (Section 1.2.2), a number of text-based methods from summarisation and information retrieval were compared to the retweet counts baseline. When the task was restricted to an information seeking use case, these methods outperformed all baselines (Section 5.4.1). In the general case, however, they were not stronger than simply counting retweets (Section 6.3). Due to the results in the general case, Hypothesis 1 could not be proven in this work.

Whilst text-based methods are needed when the documents to summarise are text, tweets exist as part of a social network and information other than text is available for tweet ranking. Not only is the author of the tweet known, but also the identity of the reader, the relationship between them, and the past activity of the reader. The second hypothesis of this work was that this type of contextual information could be exploited to improve ranking performance over a selection of approaches which rely only on the text of the tweets themselves (Section 1.2.4). The features combined through machine learning to give ranking models which outperform all baselines (Section 6.7.1). The features which comprise this model are largely personal to the reader, demonstrating that the quality of a home timeline summary is subjective and depends on the reader, and validating Hypothesis 2 with statistically significance performance results under MAP, insofar as the text-based methods evaluated in this work are outperformed by these personalised methods.

7.1 Contributions

7.1.1 Effective Evaluation

This work has argued the case for a gold standard for home timeline summarisation that is created by the reader to whom the timeline belongs. It was argued that retweets alone were not sufficient to provide reliable annotation. This argument was strengthened by the creation of a classification scheme for reasons that tweets may be found relevant.

An annotated set of 3100 tweets was created for a specific, political analysis and social media monitoring use case (Section 3.3.3). A further set of 19350 tweets were rated for relevance by 148 total users (Section 3.3.2). The data set further highlighted the weakness of retweets as a gold standard, with 73.32% of interesting tweets having been tweeted by anyone and practically none by the study participants (Section 3.5).

Existing work on home timeline summarisation uses retweets for evaluation (Uysal and Croft, 2011; Feng and Wang, 2013; Hong et al., 2013), these are global indicators of relevance, and are not personal to the reader. By comparison, the personal, user reported judgements in this work are more appropriate for evaluation of personalisation methods (these methods are the ones that perform best in this work). Tweets can enter the data set for reasons which do not overlap with reasons for retweets 3.4.

This data will be shared publicly with the wider community. Given the sparsity of retweets by the volunteers in the study in comparison to the data generated in this work, the task can now be viewed as one that is far less sparse, and includes many different types of tweet relevance, rather than just those which apply to tweets that may be retweeted.

7.1.2 Social Media Monitoring

The specific use case of social media monitoring was considered in Chapter 5, and tweet summarisers were developed to assist with the task. When annotators were looking to use the combination of search and ranking to discover tweets and trends which then drive further analysis, techniques from summarisation and text retrieval outperformed all baselines.

The combination of the Centroid method with dimensionality reduction through Latent Semantic Indexes trained on a corpus of historical incoming tweets for the account gave the strongest performance. It was demonstrated that the introduction

of dimensionality reduction here can have a significant impact in performance for tweet monitoring.

None of the existing work on home timeline summarisation has addressed the use case of social media monitoring, or carried out manual evaluation of the type demonstrated in this thesis.

7.1.3 Personalised Ranking

Whilst performance gains were shown with Centroid and dimensionality reduction for Tweet monitoring, these improvements could not be replicated in the general case, where the use case is not one of social media monitoring but rather one of ordinary Twitter use.

In this case, it was shown that effective ranking of Twitter home timelines requires knowledge of the reader as well as the author, and the exploitation of personalised features resulting from their relationship, the reader's own posts and profile, and the popularity of the author. These features, when combined through a Gaussian Naive Bayes model, were shown to outperform key baselines under MAP and match them under ROUGE.

The success of personalised ranking here demonstrates further the need for a personalised, user reported data set. Given that the best ranking methods were all based on information about the reader, attempts to create an objective gold standard will be unsuccessful, as was also argued by Alonso et al. (2013).

In no other work has personalised ranking of home timelines been carried out on filtered subsets of tweets, nor has this ranking been evaluated on a large scale user-reported gold standard.

7.2 Future Work

The key areas for future work are in better exploiting the user profile information, developing stronger machine learning models, introducing notions of tweet type and going beyond tweet ranking into creating extractive summaries.

7.2.1 User Profiles

The two strongest social media based features (aside from counts of retweets) were those which characterised the relationship between author and reader, and those which characterised the reader's content preferences based on their previous posts.

The strength of the social tie between two users also proved helpful in the author's work on geolocation (Rout et al., 2013b), in which the user graph alone was used as part of a classification to estimate whether two pairs of users are connected. It may be useful in future work to re-use the classification model developed in that work, not for estimating the likelihood of collocation but for the strength of the friendship directly, given that the two may well be related.

Additional features which are available in this work but which were not available during geolocation may be incorporated into the model of relationship strength, including the overlap in posts made by the two users, and the counts of past interactions between them.

When comparing reader interests with incoming tweets in this work, cosine similarity was used with IDF weighting and stopwords. However, the similar centroid method showed large improvements for political timelines when LSI was used as a preprocessing step. Creating topic model representations of Twitter user interests would not only be in line with existing work on user profiling (Ramage et al., 2010) but could lead to improvements in performance.

7.2.2 Machine Learning Models

The machine learning models for ranking were trained on the entire dataset including all users, and then applied universally. As shown in Chapter 5, however, there are different reasons for which a user may seek a summary of tweets, and the algorithms which perform best can depend upon the specific use case. As such it would be prudent to personalise the weighting of the various features and algorithm scores according to the type of user in question.

Experiments with personalising the ranking models to individual users were presented in Section 6.6, and unfortunately they were unsuccessful in producing models which could outperform simply ranking by singular feature values. There were some key limitations in this work. The dataset for training each feature was very small. Feature selection was carried out once for all users and not for each individual user, and the number of features was, in some cases, larger than the number of examples marked as relevant (the positive set).

Future work in this area could improve the machine learning models through domain adaptation, training one general model for all users, and adapting it specifically to each user in the set. Additional interesting training examples could be gained by retrieving the user's retweeted posts.

7.2.3 Tweet Classification

Not all tweets are interesting for the same reason (Section 3.4), and tweets can be written with many different intents (Naaman et al., 2010). In future work, tweets will be classified automatically according to their role (anecdote, humour, status update, self promotion, etc), and the potential link between these tweet classes and the reasons for relevance will be investigated.

Tweet classes may be useful in ranking home timelines in one of two different ways; different tweet classes give different reasons for relevance, and given that users read Twitter for diverse reasons, tweet classes might be used to further personalise ranking through the use of advanced machine learning methods, or by comparing the types of candidate tweets with the types of tweets that they themselves post.

It is also entirely possible that certain types of tweets will be universally more interesting than others. It is certainly the case that some reasons for relevance were applicable more often than others (Section 3.4). If self promotion tweets are very rarely of interest in the whole of Twitter, then the likelihood that a tweet belongs to this class could be used as another feature in the ranking models.

7.2.4 Extractive Summaries

The summaries produced in this work are all rankings of tweets. No text unit shorter than a tweet has been considered, however, tweets do contain sentences and phrases, like other mediums, albeit often in an informal manner (for example, the statement “This.” at the start of a tweet). When a collection of tweets occurs around a coherent theme, an extractive summary might be formed by choosing appropriate text units and combining them.

Other work exists which attempts to create phrase-based summaries out of collections of tweets, though at the time of writing the extent of this work is text generation based on common bigrams in trending topics, which are highly redundant (Sharifi et al., 2010; Judd and Kalita, 2013). None of the existing work attempts extractive summarisation of Twitter timelines.

Challenges in this type of summarisation include the ungrammatical nature of tweets (Bontcheva et al., 2013a) and the existence of syntactic features such as hashtags and mentions in tweets which would not make sense in continuous prose, as well as the inconsistent spelling which can vary between posters. These problems might be addressed through sentence selection on the basis of sentence quality, and text modification rules to increase the number of viable phrases.

Bibliography

- Abdel-Hafez, A., Phung, Q. V., and Xu, Y. (2014). Utilizing voting systems for ranking user tweets. In *Proceedings of the 2014 Recommender Systems Challenge, RecSysChallenge '14*, pages 23:23–23:28, New York, NY, USA. ACM.
- Agarwal, D., Chen, B., Gupta, R., Hartman, J., He, Q., Iyer, A., Kolar, S., Ma, Y., Shivaswamy, P., Singh, A., and Zhang, L. (2014). Activity ranking in linkedin feed. In Macskassy, S. A., Perlich, C., Leskovec, J., Wang, W., and Ghani, R., editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1603–1612. ACM.
- Alonso, O., Marshall, C. C., and Najork, M. (2013). Are some tweets more interesting than others? #hardquestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, HCIR '13*, pages 2:1–2:10, New York, NY, USA. ACM.
- Baxendale, P. B. (1958). Machine-made index for technical literaturean experiment. *IBM Journal of Research and Development*, 2(4):354–361.
- Beaudoin, C. (2008). Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. *Journal of Computer Mediated Communication*, 13:550–568.
- Becker, H., Naaman, M., and Gravano, L. (2011). Selecting Quality Twitter Content for Events. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*.
- Berkovsky, S., Freyne, J., Kimani, S., and Smith, G. (2011). Selecting items of relevance in social network feeds. In Konstan, J., Conejo, R., Marzo, J., and

- Oliver, N., editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 329–334. Springer Berlin Heidelberg.
- Bian, J., Yang, Y., and Chua, T. (2013). Multimedia summarization for trending topics in microblogs. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R., editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1807–1812. ACM.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.
- Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., and Aswani, N. (2013a). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.
- Bontcheva, K., Gorrell, G., and Wessels, B. (2013b). Social media and information overload: Survey results. Technical Report 1306.0813 [cs.SI], arXiv. <http://arxiv.org/abs/1306.0813>.
- Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.
- Bourke, S., O'Mahony, M. P., Rafter, R., and Smyth, B. (2013). Ranking in information streams. In Kim, J., Nichols, J., and Szekely, P. A., editors, *18th International Conference on Intelligent User Interfaces, IUI '13, Santa Monica, CA, USA, March 19-22, 2013, Companion Volume*, pages 99–100. ACM.
- Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.
- Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual (web) search engine. In *Proc. 7th International World Wide Web Conference (WWW7)*.
- Buckley, C. and Salton, G. (2009). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.
- Burger, J., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 1301–1309.
- Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.
- Carbonell, J. G. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336.
- Carenini, G. and Cheung, J. C. K. (2008). Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference, INLG '08*, pages 33–41.
- Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3).
- Chen, D., Schneider, N., Das, D., and Smith, N. A. (2010). Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267. Association for Computational Linguistics.
- Chen, J., Nairn, R., and Chi, E. (2011). Speak little and well: Recommending conversations in online social streams. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI '11*, pages 217–226.
- Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., and Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, pages 661–670, New York, NY, USA. ACM.

- Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 759–768, New York, NY, USA. ACM.
- Choi, J., Croft, W. B., and Kim, J. Y. (2012). Quality models for microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1834–1838, New York, NY, USA. ACM.
- Chuang, W. T. and Yang, J. (2000). Extracting sentence segments for text summarization: A machine learning approach. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00*, pages 152–159, New York, NY, USA. ACM.
- Conroy, J. M., Schlesinger, J. D., O'leary, D. P., and Goldstein, J. (2006). Back to basics: Classy 2006. In *Proceedings of the Document Understanding Conference (DUC) 2006*.
- Corston-Oliver, S. (2001). Text Compaction for Display on Very Small Screens. In *Proceedings of the Workshop on Automatic Summarization (NAAACL 2001)*, pages 1–8, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
- Dalli, A., Xia, Y., and Wilks, Y. (2004). Fasil email summarisation system. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING '04*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Damljanovic, D. and Bontcheva, K. (2012). Named entity disambiguation using linked data. In *9th Extended Semantic Web Conference (ESWC2012)*.
- Daumé, III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126.
- De Choudhury, M., Counts, S., and Czerwinski, M. (2011). Identifying relevant social media content: Leveraging information diversity and user cognition. In *Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia, HT '11*, pages 161–170, New York, NY, USA. ACM.
- Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

- Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Association for Computational Linguistics.
- Dörk, M., Carpendale, M. S. T., Collins, C., and Williamson, C. (2008). Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1205–1212.
- Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In Radev, D. and Teufel, S., editors, *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 1–8.
- Douglis, F. (2010). Thanks for the fish - but I’m drowning! *IEEE Internet Computing*, 14:4–6.
- Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *COLING*, pages 295–303.
- Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.
- Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Edmundson, H. P. (1969). New methods in automatic extracting. *J. ACM*, 16(2):264–285.
- Erkan, G. and Radev, D. R. (2004). Lexrank: graph-based lexical centrality as salience in text summarization. *Journal Artificial Intelligence Research*, 22(1):457–479.
- Feng, W. and Wang, J. (2013). Retweet or not?: Personalized tweet re-ranking. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM ’13*, pages 577–586, New York, NY, USA. ACM.
- Fink, C., Piatko, C., Mayfield, J., Finin, T., and Martineau, J. (2008). Geolocating blogs from their textual content. In *Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 1–2. AAAI Press.

- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. (2011). Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47. ACL.
- Go, A., Huang, L., and Bhayani, R. (2009). Twitter Sentiment Analysis. Technical report, Stanford University.
- Goldstein, J., Mittal, V. O., Carbonell, J. G., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL workshop on Automatic Summarization*, Seattle, WA.
- Gomez-Rodriguez, M., Gummadi, K., and Schölkopf, B. (2014). Quantifying information overload in social media and its impact on social contagions. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.
- Gorrell, G. and Bontcheva, K. (2014). Classifying twitter favourites: Like, bookmark or thanks? *Journal of the Association of Information Science and Technology*, tba:tba.
- Gorrell, G., Petrak, J., Bontcheva, K., Emerson, G., and Declerck, T. (2014). Multilingual resources and evaluation of knowledge modelling - v2. Technical Report D2.3.2, Trendminer Project Deliverable.
- Guy, I., Levin, R., Daniel, T., and Bolshinsky, E. (2015). Islands in the stream: A study of item recommendation within an enterprise social stream. In Baeza-Yates, R. A., Lalmas, M., Moffat, A., and Ribeiro-Neto, B. A., editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 665–674. ACM.
- Harabagiu, S. and Hickl, A. (2011). Relevance Modeling for Microblog Summarization. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*.
- Hargittai, E., Neuman, W. R., and Curry, O. (2012). Taming the information tide: Perceptions of information overload in the american home. *Inf. Soc.*, 28(3):161–173.

- Honeycutt, C. and Herring, S. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10.
- Hong, L., Bekkerman, R., Adler, J., and Davison, B. D. (2012). Learning to rank social update streams. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, pages 651–660, New York, NY, USA. ACM.
- Hong, L., Convertino, G., and Chi, E. H. (2011). Language matters in twitter: A large scale study. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.
- Hong, L., Doumith, A. S., and Davison, B. D. (2013). Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pages 557–566, New York, NY, USA. ACM.
- Huang, H., Zubiaga, A., Ji, H., Deng, H., Wang, D., Le, H. K., Abdelzaher, T. F., Han, J., Leung, A., Hancock, J. P., and Voss, C. R. (2012). Tweet ranking based on heterogeneous networks. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1239–1256.
- Huang, M., Yang, Y., and Zhu, X. (2011). Quality-biased Ranking of Short Texts in Microblogging Services. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 373–382.
- Inouye, D. and Kalita, J. K. (2011). Comparing Twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, pages 298–306.
- Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.
- Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, WebKDD/SNA-KDD '07*, pages 56–65, New York, NY, USA. ACM.

- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398 in Lecture Notes in Computer Science, pages 137–142, Chemnitz, Germany. Springer Verlag, Heidelberg.
- Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.
- Judd, J. and Kalita, J. (2013). Better twitter summaries? In Vanderwende, L., III, H. D., and Kirchhoff, K., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 445–449. The Association for Computational Linguistics.
- Kammerer, Y., Nairn, R., Pirolli, P., and Chi, E. H. (2009). Signpost from the masses: Learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 625–634, New York, NY, USA. ACM.
- Kapanipathi, P., Orlandi, F., Sheth, A., and Passant, A. (2011). Personalized Filtering of the Twitter Stream. In *2nd workshop on Semantic Personalized Information Management at ISWC 2011*.
- Kim, Y. and Shim, K. (2011). Twitobi: A recommendation system for twitter using probabilistic modeling. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 340–349, Washington, DC, USA. IEEE Computer Society.
- Knight, K. and Marcu, D. (2000). Statistics-based summarization – step one: Sentence compression. In *AAAI/IAAI*, pages 703–710, Austin, Texas.
- Koroleva, K. and Röhler, A. B. (2012). Reducing information overload: Design and evaluation of filtering & ranking algorithms for social networking sites. In *20th European Conference on Information Systems, ECIS 2012, Barcelona, Spain, June 10-13, 2012*, page 12.

- Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about Twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24. ACM.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, pages 79–86.
- Kwak, H., Chun, H., and Moon, S. (2011). Fragile online relationship: A first look at unfollow dynamics in twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 1091–1100, New York, NY, USA. ACM.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 591–600, New York, NY, USA. ACM.
- Laniado, D. and Mika, P. (2010). Making sense of twitter. In Patel-Schneider, P., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J., Horrocks, I., and Glimm, B., editors, *The Semantic Web – ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer Berlin / Heidelberg.
- Lee, K., Eoff, B. D., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.
- Lin, C.-Y. (2004a). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Lin, C.-Y. (2004b). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.
- Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

- Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30. The Association for Computer Linguistics.
- Mackie, S., McCreadie, R., Macdonald, C., and Ounis, I. (2014). Comparing algorithms for microblog summarisation. In Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., and Toms, E., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, volume 8685 of *Lecture Notes in Computer Science*, pages 153–159. Springer International Publishing.
- Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company.
- Mani, I. and Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, AAAI '98/IAAI '98*, pages 820–826, Menlo Park, CA, USA. American Association for Artificial Intelligence.
- Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press, Cambridge, MA. Supporting materials available at <http://www.sultry.arts.usyd.edu.au/fsnlp/>.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, New York, NY.
- Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, MIT Press. chapter 10.
- Marchionini, G. (2006). Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46.
- Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411.
- Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

- Mitchell, B. (1997a). Named Entity Recognition in German: the identification and classification of certain proper names. Master's thesis, Dept. of Computer Science, University of Sheffield. <http://www.dcs.shef.ac.uk/~campus/dcsd/projects/bm.pdf>.
- Mitchell, T. (1997b). *Machine Learning*. McGraw-Hill International Editions.
- Moens, M.-F., Li, J., and Chua, T.-S. (2014). *Mining user generated content*. CRC Press.
- Mohler, M. and Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09*, pages 567–575, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Naaman, M., Boase, J., and Lai, C. (2010). Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 189–192. ACM.
- Namaki, M., Bakhshinategh, B., Noorhoseini, M., and Dehghan, M. (2013). Expressmind: Recommending contents in an anonymous social network. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, pages 278–282.
- Nanba, H. and Okumura, M. (2000). Producing more readable extracts by revising them. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2, COLING '00*, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Nenkova, A. and McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.
- Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 43–76. Springer US.
- Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152.
- Neto, J., Freitas, A., and Kaestner, C. (2002). Automatic text summarization using a machine learning approach. In Bittencourt, G. and Ramalho, G., editors, *Advances*

- in Artificial Intelligence*, volume 2507 of *Lecture Notes in Computer Science*, pages 205–215. Springer Berlin Heidelberg.
- NTALIANIS, K., MASTORAKIS, N., DOULAMIS, A., and TOMARAS, P. (2013). Social media video collection summarization based on social graph user interactions. In Chen, Z. and Lopez-Neri, E., editors, *Recent Advances in Knowledge Engineering and Systems Science*, pages 162–167. World Scientific and Engineering Academy and Society.
- Ounis, I., Macdonald, C., Lin, J., and Soboroff, I. (2011). Overview of the trec-2011 microblog track. In *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*.
- Paek, T., Gamon, M., Counts, S., Chickering, D. M., and Dhesi, A. (2010). Predicting the importance of newsfeed posts and social network friends. In Fox, M. and Poole, D., editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press.
- Pennacchiotti, M. and Popescu, A. (2011). A Machine Learning Approach to Twitter User Classification. In *Proceedings of ICWSM 2011*, pages 281–288.
- Pirolli, P. (2009). Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42(3):33–40.
- Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.*, 1(1):81–106.
- Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA.
- Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938.
- Ramage, D., Dumais, S., and Liebling, D. (2010). Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*.

- Raux, S., Grünwald, N., and Prieur, C. (2011). Describing the web in less than 140 characters. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.
- Ren, Z., Liang, S., Meij, E., and de Rijke, M. (2013). Personalized time-aware tweets summarization. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 513–522, New York, NY, USA. ACM.
- Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.
- Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109.
- Rout, D. and Bontcheva, K. (2015). A human-annotated dataset for evaluating tweet ranking algorithms. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*.
- Rout, D., Bontcheva, K., and Hepple, M. (2013a). Reliably evaluating summaries of twitter timelines. In *Proceedings of the AAAI Symposium on Analyzing Microtext*.
- Rout, D., Preotiuc-Pietro, D., Bontcheva, K., and Cohn, T. (2013b). Wheres @wally? a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*.
- Russell, S. and Norvig, P. (1995). *Artificial Intelligence - A modern Approach*. Prentice Hall International Editions, Upper Saddle River, NJ.
- Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in Bioinformatics. *bioinformatics*, 23(19):2507–2517.
- Saggion, H. (2008). Summa: A robust and adaptable summarization tool. *Traitement Automatique des Langues*, 49(2).
- Saggion, H., Radev, D., Teufel, S., Wai, L., and Strassel, S. (2002). Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In *3rd International Conference on*

- Language Resources and Evaluation (LREC 2002)*, pages 747–754, Las Palmas, Gran Canaria, Spain.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295. ACM.
- Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02*, pages 253–260, New York, NY, USA. ACM.
- Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer.
- Sharifi, B., Hutton, M. A., and Kalita, J. (2010). Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688, Los Angeles, California.
- Sousa, D., Sarmiento, L., and Mendes Rodrigues, E. (2010). Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents, SMUC '10*, pages 63–70, New York, NY, USA. ACM.
- Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIAL-COM '10*, pages 177–184, Washington, DC, USA. IEEE Computer Society.
- Sun, J. and Zhu, Y. (2013). Microblogging personalized recommendation based on ego networks. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01, WI-IAT '13*, pages 165–170, Washington, DC, USA. IEEE Computer Society.
- Turner, J. and Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 290–297, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Uysal, I. and Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 2261–2264.
- Vert, J. P., Tsuda, K., and Scholkopf, B. (2004). A primer on kernel methods. *Kernel Methods in Computational Biology*, pages 35–70.
- Wang, L., McGee, J., and Huang, S. (2010). Sumtweets. Technical report, Texas A & M University.
- Wang, L., Raghavan, H., Castelli, V., Florian, R., and Cardie, C. (2013). A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1384–1394, Sofia, Bulgaria. Association for Computational Linguistics.
- Weng, J. Y., Yang, C. L., Chen, B. N., Wang, Y. K., and Lin, S. D. (2011). IMASS: An Intelligent Microblog Analysis and Summarization System. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 133–138, Portland, Oregon.
- Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 705–714, New York, NY, USA. ACM.
- Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94*, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yan, R., Lapata, M., and Li, X. (2012). Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 516–525, Jeju Island, Korea.
- Yardi, S. and Boyd, D. (2010). Tweeting from the town square: Measuring geographic local networks. In *Proceedings of ICWSM*.