User profiling with geo-located social media and demographic data

Adam Poulston

This dissertation is submitted for the degree of
Doctor of Philosophy

Department of Computer Science
Faculty of Engineering
The University of Sheffield
June, 2021

Abstract

User profiling is the task of inferring attributes, such as gender or age, of social media users based on the content they produce or their behaviours on-line. Approaches for user profiling typically use machine learning techniques to train user profiling systems capable of inferring the attributes of unseen users, having been provided with a training set of users labelled with their attributes. Classic approaches to user attribute labelling for such a training set may be manual or automated, examples include: direct solicitation through surveys, manual assignment based on outward characteristics, and extraction of attribute key-phrases from user description fields.

Social media platforms, such as Twitter, often provide users with the ability to attach their geographic location to their posts, known as geo-location. In addition, government organisations release demographic data aggregated at a variety of geographic scales. The combination of these two data sources is currently under-explored in the user profiling literature. To combine these sources, a method is proposed for geo-location-driven user attribute labelling in which a coordinate level prediction is made for a user's 'home location', which in turn is used to 'look up' corresponding demographic variables that are assigned to the user.

Strong baseline components for user profiling systems are investigated and validated in experiments on existing user profiling datasets, and a corpus of geo-located Tweets is used to derive a complementary resource. An evaluation of current methods for assigning fine-grained home location to social media users is performed, and two improved methods are proposed based on clustering and majority voting across arbitrary geographic regions. The proposed geo-location-driven user attribute labelling approach is applied across three demographic variables within the UK: Output Area Classification (OAC), Local Authority Classification (LAC), and National Statistics Socio-economic Classification (NS-SEC). User profiling systems are trained and evaluated on each of the derived datasets, and NS-SEC is additionally validated against a dataset derived through a different method. Promising results are achieved for LAC and NS-SEC, however characteristics of the underlying geographic and demographic data can lead to poor quality datasets, as displayed for OAC.

Contents

Acronyms

CBOW continuous bag-of-words. 69

CNN convolutional neural network. 42, 68

CSS computational social science. 44

DBSCAN density-based spatial clustering of applications with noise. 83–85, 90

EM expectation maximisation. 85

GMM Gaussian mixture model. 83, 85, 90, 95, 96

GP Gaussian process. 42, 70, 71, 73, 146

IMD indices of multiple deprivation. 11

IRB Institutional Review Board. 44, 154

KDE kernel density estimate. 107

LA Local Authority. 100, 103, 109, 113, 151

LAC Local Authority Classification. 1, 11, 45, 101–103, 106, 109, 110, 112, 114–117, 121, 125, 141, 146, 150–152, 155, 156

LAC-P Local Authority Classification profiles. 110, 112–114, 117

LAD Local Authority District. 88, 95–97, 150, 151

LDA latent Dirichlet allocation. 21, 31, 50, 53–57, 59, 68, 73, 75

LIWC Linguistic Enquiry and Word Count. 20, 32

LR logistic regression. 42, 45, 69–71, 128, 139, 141, 145, 146

LSA latent semantic analysis. 57

LSOA Lower Layer Super Output Area. 88, 95, 96

MAE mean absolute error. 31

ML machine learning. 44, 45, 153

MSOA Middle Layer Super Output Area. 88, 95, 96

NB naive Bayes. 29, 71

NLP natural language processing. 25, 44, 45, 49, 68, 71

NS-SEC National Statistics Socio-economic Classification. 1, 45, 121–123, 125–127, 129, 141, 142, 146, 150, 152, 155, 156

NSG Neighbourhood Statistics Geography. 88

OA Output Area. 88, 95, 96, 100, 102, 103, 109, 113, 123, 125, 126, 150, 151

OAC Output Area Classification. 1, 11, 45, 100–103, 106, 109, 110, 112–117, 121, 125, 141, 146, 150–152, 155, 156

OAC-P Output Area Classification profiles. 109, 110, 112–114, 117

ONS Office for National Statistics. 87, 101

POS parts-of-speech. 18, 49, 55–57, 76

RBF radial basis function. 73

RNN recurrent neural network. 42, 43

SES socio-economic status. 125, 126, 142, 152, 155

SOA Second Order Representation. 57

SOC Standard Occupational Classification. 123, 125

SVM support vector machine. 29, 42, 45, 47, 50, 57, 63, 70, 71, 74, 75, 101, 111, 112, 116, 145, 146

TF-IDF term frequency-inverse document frequency. 24, 53, 57, 58, 63, 70, 74, 75, 101, 111, 127, 141, 146

TL transfer learning. 153

Chapter 1

Introduction

1.1 User profiling

User profiling is the task of determining the characteristics of a set of users based on factors such as the text they produce, how they behave, and with whom they interact. A user profiling study will typically center on evaluating one or more *attributes* of one or more *users*. An attribute can represent any element of a person's self, ranging from obvious outward characteristics such as gender and age, to more inward qualities such as personality, political leaning or sexual orientation.

A range of potential applications exist for user profiling techniques, most obvious of course being the potential for personalisation, marketing and customer insight [1, 2, 3]. A company or organisation could apply a user profiling tool to identify their core user-base, and based on these insights marketers could further target advertisement to users who are determined to hold the same characteristics. Law enforcement could potentially use such a system to link on-line criminal behaviour with individuals—studies have already investigated the use of user profiling techniques in identifying on-line grooming [4], for example. Another potential application might be as a tool in diagnosing mental health issues such as personality disorders or depression in individuals out of reach of a trained medical professional, for example in populations with limited access to healthcare [5].

Alongside these potential applications, which range from the benevolent to the commercial, research into areas such as user profiling play a valuable role in high-

lighting the sorts of information we reveal about ourselves, both explicitly and implicitly, through our day to day discourse and activities online [6, 7]. The studies conducted in this thesis are all performed on consumer-grade hardware, using APIs and datasets openly available without any strong barriers to access; the insights we and others reveal shine a light on the areas large organisations are likely leveraging behind closed doors.

To date a variety of attributes have been examined with varying degrees of success. Earlier studies—presented in Section 2.2.1—focussed on the more obvious variations in humanity, such as age [8], gender [9] and native language [10]. Later, due to the emergence of blogging and social platforms such as Twitter increasing access to data, less obvious and much more personal attributes such as personality scores [11, 12], political leaning [13, 14] and sexuality [15] were able to be investigated.

The early, 'outward' characteristic studies generally relied on personal information obtained directly from individuals. For example, a researcher wishing to develop a system to determine the age of individuals based on their text would have to obtain a set of documents authored by a variety of individuals, and ask for their ages, known as acquiring *ground truth* data [9, 10, 16]. Some attributes however, may not be as easily available, for example, a researcher wishing to investigate variations in language across the Unites States, would struggle to directly contact a reasonable number of individuals across all states, and in general, studies were performed by researchers in universities, often taking data from student volunteers, who in general are not representative of the wider population. In addition, attributes that are more difficult to characterise, such as personality type and socio-economic status, are not readily known by most individuals; acquiring such ground-truth attributes from individuals through direct contact would therefore be a laborious process, such as through a proctored personality assessment, which is infeasible outside of small-scale experiments.

Digitizing the process of acquiring ground truth data has allowed datasets to be constructed with additional ease, as users can more easily be asked to provide information directly (e.g. by answering an on-line questionnaire) or by providing researchers with access to a social media profile [15, 17, 18]. Some attributes may be more private than others, might be more difficult to define on an individual level, or might be more more or less prevalent in users liable to self-select for a social media study, again introducing the risk that the datasets created are not

reflective of the overall population. Failing to cover a suitably representative portion of the population when examining socio-demographic concepts or developing user profiling systems is likely to lead to biased or incomplete results, or result in a system of limited utility when deployed to a more broad set of users.

Several studies, such as [19] and Preoţiuc-Pietro et al. [20], noted that social media users often either explicitly or implicitly declare certain attributes in the "about me" or "biography" section of their profile. Consider the fictional example "Sally Gulmore, 28 year old mother of two from Austin, TX", which contains several explicit and implicit nods towards personal attributes. Gender can be inferred from the use of the phrase "mother of" and the name "Sally", which we know from census statistics is most often a female name, age is declared directly, and a city-level geo-location is possible via "from Austin, TX". These disclosures of personal characteristics can be leveraged to create datasets without requesting information directly from individuals, enabling collection on a larger scale than was possible with previous approaches. Attribute data acquired in this automated way can be more prone to errors than approaches that require a human to supply information manually, as it relies on the user providing correct information in the first instance and keeping their profile up to date (e.g. updating their age each year) going forward.

1.2 Introducing geo-demographic data

Many countries regularly conduct surveys of their population that provide large-scale aggregate representations of demographic information. Countries such as the United Kingdom [21] and the United States [22] periodically conduct censuses and release the data aggregated at various geographic levels, with a liberal license for both research and public use. Governments are increasingly adopting a stance of openness and are making this data available for research purposes. Census style data has the advantage of providing population-level information gathered using methods with a strong statistical backing. A wealth of information is present in this data, such as education levels, social class, political views and occupation category for well-defined geographical areas. Much of these demographic variables have previously been explored only lightly (or not at all), in the user profiling literature.

In the same vein of open communication, users are increasingly engaging with social media platforms such as Twitter in a public manner, in some cases providing public measurements of their location at the time of posting, referred to as *geo-located posts*. Although only a small portion of posts for a given platform are geo-located [23], they still form a significant amount of data given the vast number that are generated each day.

Given the accessibility of these two rich sources of data, we propose that they could be combined to improve upon, supplement, or even replace data generation for traditional user profiling methods, as well as allow attributes to be addressed that have not previously been investigated.

1.3 Thesis aims

In order to leverage the large quantities of social media posts and demographic data that are under exploited for user profiling, we propose a method for combining geo-located posts with geographically linked demographic data to generate user profiling datasets, and train user profiling systems that address novel demographic variables.

Our approach, illustrated in Figure 1.1, allows demographic information to be associated with a given user by identifying a 'home location' based on their geo-located posts. Once a home location has been predicted for a user, they can be labeled with their local demographics by looking up their home location in geographically linked demographic datasets (also known as geodemographic segmentation datasets). Datasets suitable for use with this method must link some demographic variable, such as household income, occupational class, or educational attainment, to specific local areas in which users may be geo-located; examples of suitable datasets include raw data from the UK [21] and US censuses [22], and derivations thereof such as Output Area Classification (OAC) [24], Local Authority Classification (LAC) [25], and indices of multiple deprivation (IMD) [26], which are aggregate measures that attempt to distil the broader demographics of an area into easy to interpret variables.

Utilising geo-located social media posts and demographic data allows user profiling datasets to be constructed for as yet unaddressed demographic variables. In ad-

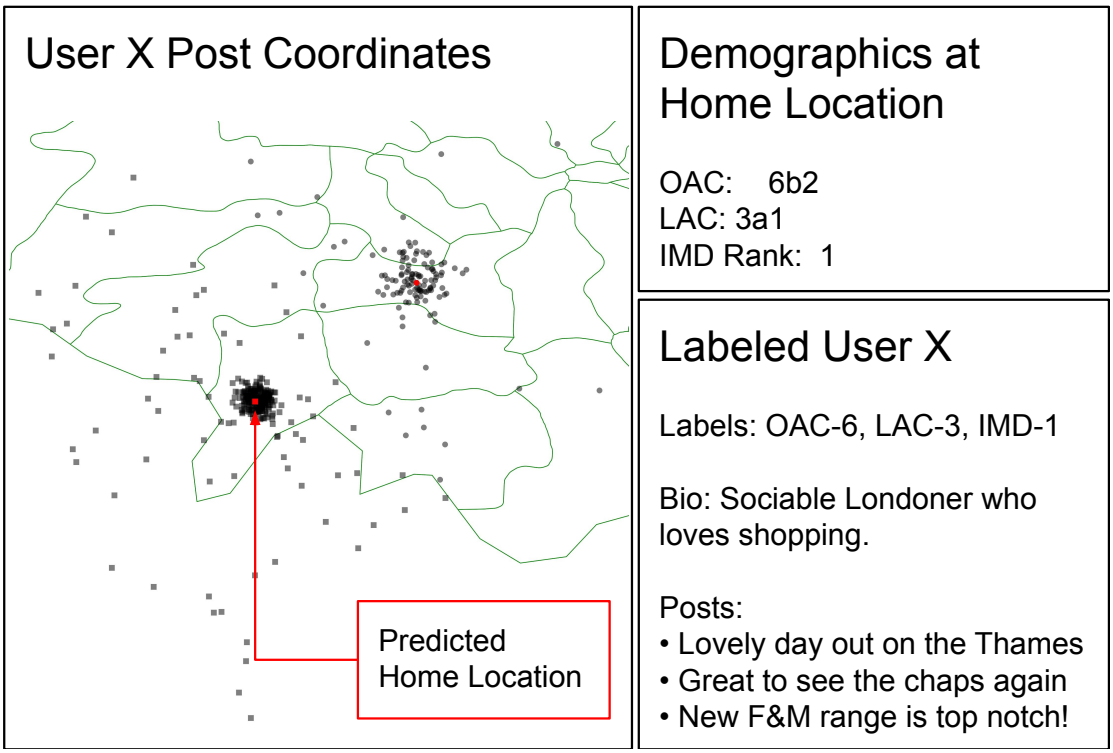

Figure 1.1: Illustration of the proposed process for labeling social media users with their local demographics. First, the user's geo-located posts are used to predict a 'home location' (*left panel*). The home location is used to lookup local demographics in a number of geodemographic datasets (*top right panel*). The resulting demographic variables are processed into a usable form and attached to the the user's profile (*bottom right panel*).

dition, we are able to generate datasets for some demographic variables that have previously been investigated at a much larger scale than presented before. The applied home location allocation method must be accurate and robust, to avoid propagating errors onwards into the derived datasets and resulting user profiling models.

Our approach relies strongly on the concept of *homophily*, the observation that individuals who live in an area, tend to be broadly representative of the demographics of that area, and that broadly similar groups of people tend to live in similar areas [27]. If this concept holds, individuals Tweeting from an area that has been identified as their 'home', are likely to match the demographics of that area. Obviously this cannot hold true for all demographic variables; for example, distributions of demographic variables such as age, gender, and sexuality do vary between areas, but rarely to such an extent that our approach would reliably assign an individual to a class of this nature. Instead, this approach is likely to be most

useful for assessing demographic variables that correspond to the socio-economic status of the user, such as household income and occupational class.

Despite the proposed method's ability to provide demographic annotations for a large number of users, it is not necessarily applicable as a wide-ranging user profiling tool in its own right, as only a small percentage of social media users consistently choose to geo-locate their posts [23]. As such, we frame the utility of our method in terms of ability to generate useful datasets for training predictive systems that infer the demographics of *all* social media users, including those who do not geo-locate their posts.

In this thesis we aim to develop and evaluate our method for generating user profiling datasets by:

- Establishing good baseline implementations for user profiling predictive systems based on datasets generated through 'classic' approaches. This is addressed in Chapters 2, 3 and 4;
- Establishing a strong grounding for our proposed method by evaluating whether simple high level datasets derived from geo-located Tweets can be used to improve performance on a user profiling task. This is addressed in Chapter 4;
- Evaluating methods for estimating user home location, determining whether the state-of-the-art is good enough for use in our proposed method, and improving on the state-of-the-art if necessary. This is addressed in Chapter 5;
- Generating novel datasets for user profiling using our proposed method, and evaluating them by developing user profiling predictive systems incorporating established strong baseline approaches. This is addressed in Chapter 6; and
- Supplementing and contrasting existing user profiling datasets with ones derived by our methods. This is addressed in Chapter 7.

1.4 Thesis outline

Chapter 2 contains a review of user profiling; attributes addressed by various approaches are explored in Section 2.2, focusing on both on and off-line media. Established approaches for acquiring "ground-truth" geo-location data for social media users are covered in Section 2.4, which we build on in Chapter 5. A review of the techniques and tools used perform a user profiling is presented in Chapter 2, which we use to establish a good baseline for our predictive experiments, starting in Chapter 3.

In Chapters 3 and 4 we address and implement classic approaches to user profiling, identifying baseline approaches for our later experiments, and also investigate introducing simple demographic data (dialect) linked to geo-tagged posts at a broad scale (country), with good results.

We finalise our approach for deriving high quality user location estimates in Chapter 5. State-of-the-art approaches for determining user locations from the literature are empirically evaluated, alongside two novel methods.

In Chapter 6, we apply our approach on a dataset of Twitter posts to address two variables not previously addressed in the literature, the *Output Area* and *Local Authority* classification schemes.

Chapter 7 focuses on the application of our approach to a variable that has been addressed before, with a publicly available dataset. We implement an analogous evaluation framework to the state-of-the-art and compare and contrast the two datasets.

The thesis is concluded with a discussion in Chapter 8.

1.5 Publications

The work in this thesis yielded four peer reviewed publications, listed below, and also helped inspire an art installation, "Happy Sheffield"[1][2], at Festival of the Mind

[1] `https://www.sheffield.ac.uk/ideasbazaar/projects/happy-sheffield`

[2] `https://www.theguardian.com/lifeandstyle/shortcuts/2016/sep/11/happy-sheffield-happiness-clock-twitter-city`

2016, which showcased a live stream of emotive Tweets across Sheffield set against a backdrop of stylised faces, inspired by literature on the psychology of human emotion [28].

- **Adam Poulston**, Mark Stevenson and Kalina Bontcheva. 2017. "Hyperlocal Home Location Identification of Twitter Profiles". Proceedings of the 28th ACM Conference on Hypertext and Social Media.
- **Adam Poulston**, Zeerak Waseem, Mark Stevenson. 2017. "Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling-Notebook for PAN at CLEF 2017". CLEF (Working Notes) 2017
- **Adam Poulston**, Mark Stevenson and Kalina Bontcheva. 2016. "User profiling with geo-located posts and demographic data". NLP+CSS at EMNLP 2016.
- **Adam Poulston**, Mark Stevenson, Kalina Bontcheva. 2015. "Topic Models and n-gram Language Models for Author Profiling - Notebook for PAN at CLEF 2015". CLEF (Working Notes) 2015

Chapter 2

Background

Studies on user profiling tend to focus on the development and evaluation of *predictive systems* capable of inferring *user attributes* (personal characteristics such as gender or age). Approaches to building these *user profiling systems* typically follow a similar process, presented in Figure 2.1.

User profiling systems begin with a collection of *user content* ((1) in Figure 2.1), such as emails, essays, blogs or social media profiles. This is then processed into *feature vectors*, numeric representations of the content, suitable for training a machine learning model. We present the types of user content used and feature vectors derived across the literature in Section 2.1.

Each user in the training set is tagged with *ground truth* user attributes that

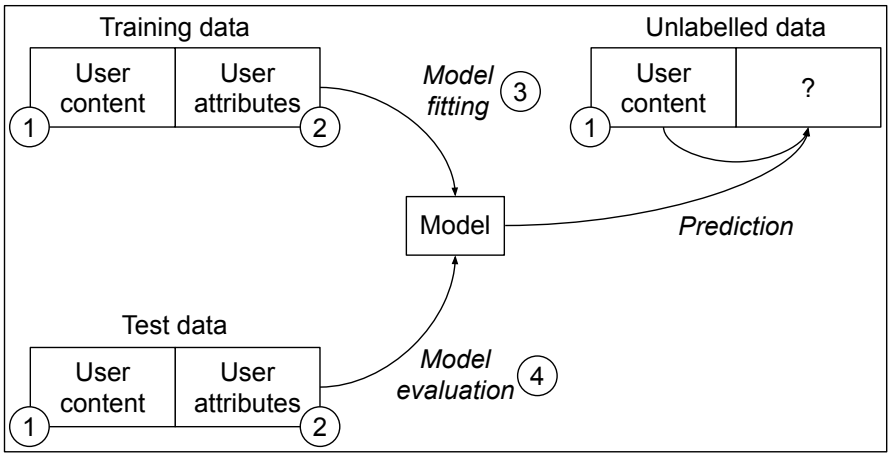

Figure 2.1: Typical process undertaken to build a user profiling system.

the system is intended to predict ((2) in Figure 2.1). We present the attributes addressed across the literature in Section 2.2, and discuss approaches for tagging these attributes Section 2.3. In Chapter 1 we introduced our novel method for user attribute labelling ((2) in Figure 2.1) in which a user is geo-located and labelled with attributes based on their local demographic data. We present methods for acquiring accurate ground truth geo-location estimates for social media users in Section 2.4.

Once the user content is processed and tagged with user attributes, the training data is used to train (fit) some machine learning model ((3) in Figure 2.1), which can then be used to predict the tagged user attributes on future unlabelled data. We discuss the process of fitting machine learning models in Section 2.5.

Typically, a portion of the labelled training data is set aside to evaluate the predictive power of the trained model when applied to unseen data ((4) in Figure 2.1). Approaches to model evaluation are presented in Section 2.6.

2.1 User content

User profiling systems attempt to make predictions of a user's attributes based on the content they output ((1) in Figure 2.1), such as the text they write, what pages they interact with, and who they follow. In this section, we present the types of data available to create user profiling systems (Section 2.1.1), and the steps taken to convert user content into feature vectors suitable for use with machine learning models (Sections 2.1.2 and 2.1.3).

2.1.1 Types of data

Earlier user profiling studies employed a mostly text based approach, focusing on the content and style of writing. Today text data remains one of the main components of a user profiling system, although it is now often supported by additional information.

Text data will typically have been authored in either a formal or an informal environment. Formal texts such as essays, professional emails or social media posts from a company are more likely to contain neutral language as well as a lack

of slang words, which are likely to be attributed to specific groups. In the case of company wide email addresses or social media profiles it is not guaranteed that one particular person has authored all of the text.

Informal texts, such as personal social media profiles or emails are more likely to contain new and non-standard language, which may be attributed to particular groups or may simply be the result of poor spelling and grammar. Systems dealing with this sort of text should not rely on adherence to an established lexicon, as valuable information may be ignored. In addition, artefacts of text authorship can be of value as features in their own right.

On-line media such as emails, blog posts and social media profiles, contain additional information such as activity times, HTML and user's social networks. From these statistics of a user's behaviour can also be calculated.

2.1.2 Pre-processing

Before feature extraction can be performed user content must be formatted in a suitable fashion. This step can vary greatly depending on the type of the input data. In the base case where raw text is passed in, this could be as simple as tokenising, typically by splitting into words and punctuation, or assigning parts-of-speech (POS) tags. Internet media, while still text based may have additional non-text content such as HTML or hyperlinks to consider. Social media also brings its own challenges, if the user's social network or behaviours are to be considered, they must be collected and calculated at this stage.

Care should also be given at this stage for user privacy. It may be prudent to anonymise references to the user and other users as well as the names of people and places.

2.1.3 Feature extraction

After the user content has been processed into a suitable format, features can be extracted to build feature vectors for each user. A feature vector is a vector of length N, where N is the total number of features across all input documents, and represents the counts of each feature present in a particular document (or

user profile in the social media case).

In machine learning a feature is some quantifiable element that can be determined from input data. In some cases such a n-gram modelling features relate directly to an obvious element, words or characters for example. In other cases, such as the "structure" or "readability" of a text, what can be used as a feature is less obvious and some measure must be used (or developed) to quantify them.

Features that can be extracted depend on the format of the input media, and an overview of the different types successfully utilised in the user profiling literature is presented and discussed below.

Stylometric

Features of this type can represent a variety of concepts. At the lowest level it may be individual words, but can be much more complicated, such as the topics present in a text, or the average sentiment score across sentences. The most prominent examples are presented and discussed below.

n-grams An n-gram is a sequence of n units of text or speech, often referred to as *tokens*; these could be any unit such as words, characters or phonemes. In the user profiling scenario, word, character, and part-of-speech n-grams are dealt with most commonly. One, two and three-grams are known as unigrams, bigrams and trigrams respectively. n-grams greater than three are usually referred to by the value of n, four-gram for example. n-grams are seen as useful features in classification tasks due to their ability to capture patterns in text and speech.

Word Word n-grams are one of the most simple features available due to their ease of acquisition. Despite this they are one of the most robust features available, forming the basis of many good systems.

Character Character n-grams allow the capture of character level language features. For example, character n-grams with n ranging from 2 to 4 allow the capture of the majority of prefixes and suffixes in the English language if taken from the beginning and end of words.

Parts-of-speech Parts-of-speech are the types of words in a language such as nouns, verbs and adjectives. They allow more general patterns to be

captured, as each element represents a word class rather than a specific word.

Word Categories Examples of dictionaries exist which can be used to transform a text into counts of different categories of words, relating to psycholinguistic concepts. Linguistic Enquiry and Word Count (LIWC)—a text analysis tool described in Pennebaker and Francis ME [29]—has been used in many user profiling studies and has been shown to produce good results. LIWC determines the degree to which a text uses up to 82 language dimensions such as positive and negative emotions, self references and causal words.

Structural The structure of a text—such as the use of paragraphs and length of sentences—is indicative of the author's writing style, and can indicate other factors too. For example, a person who writes in paragraphs consisting of sentences of reasonable lengths, where words are correctly capitalized, is likely to more educated than somebody who does not. Similarly some people may have a tendency to write in short factual sentences, whereas others may pad sentences out with emotional and descriptive words. Some mediums such as blogs or email allow presentation of text to be customised more finely, often via HTML and CSS. This is likely to indicate a higher technical ability in those who take advantage.

Errors The presence of different types of errors in text has been shown to be indicative of certain characteristics, such as native language in Koppel et al. [30]. Errors to be identified are likely to be either orthographic or syntactic. Errors in the writing of the language, such as spelling, capitalisation, punctuation or use of out of alphabet characters—perhaps from their first language—are orthographic errors. Syntactic errors deal with whether a sentence is well formed or not, highlighting problems such as false pluralisation and using the wrong tense. Text from on-line sources such as Twitter is notoriously non-standard, due to the informal setting and in some cases limits on the number of characters that can be posted. As such steps should be taken to distinguish between spelling mistakes, abbreviations and out of dictionary words such as slang.

Named Entities Named entity recognition is process of classifying elements of a text into pre-defined categories such as names of people or organisation or measures of quantity or time. For example, consider the sentence "I went

to school today". The first word. "I" indicates that a person (the author) is involved. The word "school" could refer to either a building or an organisation, in this case the most likely category is an organisation. A time can also be identified by the presence of the word "today". A relationship between these categories can also be extracted, as the words "went to" indicate travel.

Readability Readability is how easily a a text can be understood by a reader, and there exists multiple mathematical formulas that can assign a score to how readable a text is. The formulas are based of self evident concepts of readability. Short sentences, with a low quantity of long words, omitting unnecessary filler words, should score as more readable than longer rambling sentences or sentences using a lot of overly complicated wording. Studies have shown that readability does correlate with certain characteristics, with Davenport and DeLine [31] and Llorente et al. [32] showing that readability of Tweets correlates with education and unemployment respectively.

Topic Models Topic models are a group of algorithms that identify hidden themes (topics) in collections of documents. The most common topic model in use today is latent Dirichlet allocation (LDA) [33], a generative model in which documents are modelled as a finite mixture of topics. In LDA each word in a document must be generated by one of its topics. In the author profiling case, topics are typically used as a binary feature, and have been shown to produce reliable results when used alone and in conjunction with other features.

Word vectors Word2Vec is a two layer neural network for learning vector-space representations of words. Word vectors are learned by observing the context of words in documents [34]. Word vectors that are close in vector space tend to be close in meaning too, thus Word2Vec could be said to model the semantics of words. With large amounts of data available for training, the inferred meaning of these word vectors can often be highly accurate. Word2Vec's main purpose is to transform words into a format suitable for input into some deep learning system. They are also useful as a dimensionality reduction tool. One approach is to average all word vectors in a paragraph (known as creating paragraph vectors [35]), resulting in a vector of much lower dimensionality than bag-of-words. Another approach is to the cluster word vectors into groups of semantically similar words, remodelling the

problem from bag-of-words to bag-of-clusters. The latter approach is used in Preotiuc-Pietro et al. [36], Preoţiuc-Pietro et al. [20] and good results are achieved.

Behavioural

A person's attributes are know to have an influence on their behaviour. According to the big five model of personality an extroverted person is likely to be more outgoing, assertive and have a positive demeanour [37]. Another example is that of gender differences in conversation. Mulac et al. [38] aggregate results from many studies, highlighting differences between male and female speech. Male speech is identified as being more assertive, factual and self referential. Female speech on the other hand tends to be more emotionally driven, inclusive—by asking questions or opinions—and has tendency to qualify statements to seem less assertive—"kind of" or "it seems" for example. Extra examples are discussed below.

Activity Times The times a user authors their documents is likely to give information regarding their habits, and possibly their characteristics. It shown in Llorente et al. [32] that in areas with low unemployment, there is a sharp increase in Tweets around the beginning of the working day, a trend far less pronounced in areas with high unemployment. A common way to code activity is to record the fraction of a user's activity that occurs at each hour of day.

Conversational Behaviour It is shown in Mairesse and Walker [39, 40], Mairesse et al. [41] that elements of conversation can be useful features in classifying personality—and most likely other attributes. One of the corpora used in their study was a collection of telephone speech recordings and their transcripts; from this utterance and prosodic features were extracted. Utterance features recorded were the ratios of commands, prompts, questions and assertions. Prosodic features can not be applied in a purely textual domain, although in cases it may be possible to extract analogues to utterance features from textual conversations.

Social Network Specific Behaviours The nature of social networks allows various behaviours to be captured, including conversation elements, these are discussed below.

Post to reply ratio The ratio of posts to replies a user writes can be seen to be analogous to either a tendency to inform in the case of a high ratio, and a tendency to converse when the ratio is low. Care must be taken though, as a low ratio does not indicate a tendency not to inform, as the user may have initiated conversations with long reply chains.

Post to share ratio A good measure of a user's egocentricity might be in their tendency to disseminate other peoples content, such through sharing a link to an article or sharing another user's post. For example, a Twitter user who Tweets ten times a day, but rarely "retweets" or shares links is likely to be quite egocentric. Again care must be taken in interpreting the ratio, as a user who mostly posts in reply to other user's posts and doesn't share much would have a high ratio of posts to shared information, despite potentially talking very little about themselves.

Conversations It may also be possible to gather conversations a user has taken part in, depending on the platform. If it is it will be possible to trace the user's tendency to either begin or join conversations.

Network

The two most popular social networks Twitter and Facebook, can be seen as directed graphs, with vertices representing individual users and edges representing relationships between the two. On Twitter all relationships can be reciprocated, if a large organisation wanted to follow "Sarah from Alabama", it can. Facebook on the other hand requires organisations and concepts to be registered as a separate "page", which users can "like" (analogous to follow in the Twitter context) but not themselves be followed back. Therefore user to page likes in the Facebook graph are always unidirectional. Justification for various network features are discussed below.

Network Statistics Links between network structure and personal attributes are discussed in Quercia et al. [42], where four types of Twitter users are identified based on trends in personality scores. "Listeners" and "populars", those who follow and are followed a lot respectively, show significant correlations with extraversion and neuroticism. The authors highlight that this is similar to the real life trend that the two traits are predictors for number of friends in real life. "Highly read" users are those who are saved in other

user's reading lists. These users correlate significantly with openness; open people tend to be viewed as more creative than closed people. "Influential" users are identified based on the tendency for their Tweets to be read. The authors note correlations with extraversion and conscientiousness for these users.

Connections as Features The presence of a particular edge in the social graph as a feature has been shown to be of use in multiple studies, particularly when identifying classes that tend to have a favourite celebrity or product [43]. In Youyou et al. [44] a machine learning system, using only Facebook likes as features, has been shown to predict a user's big five personality scores more accurately than humans, including family and friends, and only narrowly being beaten by spouses.

2.1.4 Feature weighting

Another import element of vector space models is feature weighting. The most basic feature weighting scheme is binary, in other words, does feature x appear in document y? If so rank the feature as 1 (Yes) otherwise 0 (No). For some feature types, such as whether a Tweet is a reply or retweet, binary presence is the only way they can be represented, although others such as n-grams may appear in a document more than once, in these cases, other measures such as the the frequency each feature appears in a document may be used.

Another commonly used scheme in text classification contexts is term frequency-inverse document frequency (TF-IDF). TF-IDF is a measure used to determine how important a feature is in a document, based on how common the feature is in the corpus as a whole. It is a combination of two statistics: term frequency—the raw frequency of a term (feature) in a given document—and inverse document frequency—how much information the feature provides, based on how common it is in the corpus as a whole. TF-IDF is typically used in conjunction with the bag-of-words model, where uncommon words are ranked highly, and common words such as "the" are ranked lowly [45]. TF-IDF is a consistently popular feature weighting schemes; for example, the review in Beel et al. [46] showed that 83% of text-based recommender systems use TF-IDF.

Other values can be used as weights depending on the context. In a topic model

for example a topic has an associated probability of belonging to a document. This probability could be used as the weight.

2.2 User attributes

Work on user profiling can be sliced multiple ways, but of most relevance to this thesis is user profiling on social media. In this section we present the user attributes ((2) in Figure 2.1) addressed in previous works, split between those studies performed on classic media (Section 2.2.1), and those on social media (Section 2.2.2).

2.2.1 Profiling tasks on text

This section provides an overview of what attributes were investigated before the advent of modern social media. Although our work focusses on detecting attributes of social media users, the works here helped form the basis of approaches more relevant works, and are as such included for completeness. Many of the studies here refer to the field as "author profiling"; although some later studies will also use this term, we refer to the field as "user profiling" for consistency.

Gender

Natural language processing (NLP) and sociolinguistics were first brought together in Koppel [9], using established procedures from the text categorisation literature. Instead of classifying documents into categories such as fiction or non-fiction, the texts would be classified by the gender of the author. The study was performed on a collection of texts from the British National Corpus, including both fiction and non-fiction works. The authors report accuracies of around 80% for both genres.

Gender classification has been extensively investigated in the medium of text [10, 8, 47, 48, 49, 50, 51]. All approaches followed the established machine learning approach from text classification with varying techniques for feature selection and algorithm choice.

Native language and geographic origin

In de Vel et al. [10] it is also highlighted that the techniques used are an extension of established text categorisation techniques. They use a corpus of emails to classify gender and whether the text was written by a native speaker. They hypothesise that when a non-native speaker uses another language, elements of their mother tongue will remain present. Koppel et al. [30] expand upon this by classifying the first language of an author. They achieve this by identifying erroneous occurrences in text, such as spelling mistakes and syntax errors. Using the International Corpus of Learner English as training data, native language could be classified with an accuracy of around 80%. A similar approach is also presented in Argamon et al. [49].

A system able to classify the first language and dialect of email authors—as well as other attributes—is presented in Estival et al. [48]. The system uses established machine learning techniques to identify with around 81% accuracy which language a user has as their first, from a set of three possibilities. For users with English as a first language, the system is also able to determine whether the user had a US, UK or New Zealand dialect.

Personality

Aspects of personality were first classified in Argamon et al. [16], focusing on two aspects of personality; neuroticism—tendency to worry—and extraversion—preference for the company of others. The authors note that the work integrates knowledge from Language Psychology via Systematic Functional Grammars and highlight that—in their opinion—prior research did not contain a solid enough grounding in psychology or linguistics research, and as such the results hold less meaning. Their system classified authors of casually written texts as high or low on the two personality elements selected with an accuracy of around 58%.

The work in Argamon et al. [16] was expanded upon in Oberlander and Nowson [52] by classifying two additional personality aspects; conscientiousness—tendency towards thoroughness—and agreeableness—tendency to show concern for others. Systematic Functional Grammars were not used in this study; instead a simple word n-gram approach was used, with stringent feature selection rules. Much better results were achieved in comparison to Argamon et al. [16], although the

authors note this could be related to over-fitting from the feature selection method and the small size of the training corpus.

Openness to experience, the fifth personality element to be classified, was investigated in Mairesse and Walker [39, 40], Mairesse et al. [41]. The authors approached the problem from both a text based and a speech transcription angle, identifying that conversational cues hold information regarding personality scores.

Previous studies identified that personality classifiers trained on small datasets did not scale well the greater population, the effects of a larger dataset being used were studied in Iacobelli et al. [53]. The system, based on word bigrams, was shown to have improved results over previous studies.

Earlier author personality profiling works tended to focus on English text, although a system able to predict personality values in English and Arabic Emails is presented in Estival et al. [47, 48].

Age

The automatic prediction of author age was first investigated in Schler et al. [8]. A collection of 37,478 blogs whose author's personal information—age and gender—was public, was used to train a classifier to predict age and gender. The classifier was shown to perform well, achieving accuracies of around 80% for gender, similar to previous studies. For age the system did not predict an age value, instead it predicted from a set of predefined ranges, with this is mind accuracies of 75% were achieved. It is also highlighted that extra, subtle, clues to personal attributes may be present in blogs, such as formatting choices, but they are not covered.

Similar to gender, age has been widely investigated, with classifiers predicting exact ages (such as 27 years old) or ages in a range (20 to 30 years old for example) in conjunction with a variety of feature selection techniques [47, 48, 54, 49, 50].

Political ideology

Some studies have aimed to predict the political opinions of users, typically on some variant of the left-right spectrum or party specific support. The majority of these studies have focussed on identifying political opinions in social media, pos-

sibly due to the interest associated with election results and the wild speculation that occurs in the run up to elections.

Political ideology in more traditional text is investigated in Jelveh et al. [55], where topic models are used to show that the political leaning of economists—individuals who in general should write without bias—can be predicted with accuracies of a little less than 70% and good correlations with true values. A companion paper, Jelveh et al. [56], argues the potential applications of such predictive capabilities in decoupling ideology when considering recommendations on government and organisational policy.

Education and intelligence

The author profiling tool described in Estival et al. [47, 48] also attempts to classify level of education, distinguishing between those with tertiary education and those without. Almost no improvement over the baselines in Estival et al. [47] is achieved due to extremely skewed data. Better results are achieved in Estival et al. [48] as a more balanced dataset is used.

2.2.2 Profiling on social media

User profiling studies on social media in a modern setting usually relate to the two main providers: Facebook and Twitter, due to the openness of these platforms for data collection. Facebook has now restricted access, due to negative public and regulatory backlash surrounding inappropriate data collection by certain bodies.

Facebook allows users to create a profile, containing information such as job history and interests, and connect with other users via "friendships". Users are able to join discussion groups, post status updates and much more by way of conversation. Interest in particular products, organisations or concepts can be registered with Facebook's "like" system. When a user "likes" a page, it is shown on their profile.

Twitter is a "micro-blogging" service which allows users to post short messages, or "Tweets", to their public profile. According to ex-Twitter CEO Dick Costolo upwards of 500 million Tweets are posted every day (as of October 2012). Twitter profiles vary from many traditional blogs in that all user information is stored in

an unstructured fashion; a user does not have to provide any information regarding themselves at all, and only a short, free text, description field is provided should the user wish to give extra information. As well as this, posts are limited to 280 characters (previously 140), causing Twitter users to adapt by using slang, shortened words and"hashtags", short yet informative strings meant to summarise the topics present in the Tweet.

Both Facebook and Twitter can be viewed as "social networks". That is both are based on connections between individuals or organisations. On Facebook these connections can be mutual from person to person, and directional from person to organisation via likes. On Twitter users can "follow" other users with no guarantee of reciprocity, and organisations also have the ability to forge connections to individuals. Properties of these social networks have been shown to be of use in user profiling [42, 44].

Gender

The predictive performance of gender classification when applied to social media data is assessed in Rao et al. [13], Rao and Yarowsky [14], where a dataset of Twitter users is used to classify gender, age, political orientation and regional origin. Gender results of 72% accuracy were achieved, slightly worse than some prior studies. The approach used is mostly text based but does touch upon social media specific features such as the tendency to retweet, and preference for particular organisations.

A Bayesian model is presented in Rao et al. [57] that utilises a dictionary of known Nigerian names that map to specific genders and Nigerian ethnicities, along with topic models. The results of the proposed model are compared to the results of other machine learning models, when used with the same features. For gender prediction their proposed model is shown to operate with around 80% accuracy, whereas support vector machine (SVM) and naive Bayes (NB) are reported to have accuracies of less than 55%. The reported accuracies for SVM and NB seem particularly low in this study, as both later and earlier papers show that accuracies of above 80% can be achieved with them.

An approach from the perspective of child predator detection described in Peersman et al. [4] merges gender and age prediction into a single problem. In one

experiment users are given a gender and age category of whether they are an adolescent or not coupled with their gender—female adult or male adolescent for example. For the combined prediction task the system achieves accuracies of around 66%, by contrast, age group prediction on the same dataset without gender segregation achieves around 88% accuracy.

A comparison of the quality of human annotations, provided by Amazon Mechanical Turk workers, of gender compared to a Twitter gender prediction system based on the textual elements of a Twitter profile is presented in Burger et al. [19]. The majority of annotators presented in the study are shown to achieve a much lower accuracy rating compared to the system.

Homophily, the tendency for people to group together with and and behave the same as people with similar views is investigated in Zamal et al. [27]. The approach argues that due to homophily the posts of a user's "friends" can be included as input data to a user profiling system to improve classification accuracy. Three attributes are investigated: age, gender and political orientation. For gender inclusion of neighbour profiles was shown to give little to no improvement to accuracy.

A technique called "Differential Language Analysis (DLA)" is presented in Schwartz et al. [58]. A DLA appears to be a combination of standard text and topic models, where features are ranked based on how much they discriminate towards a particular class. When applied to a dataset of Facebook statuses accuracies of around 92% are achieved for gender.

An author profiling system described in Kosinski et al. [15] uses only Facebook likes as input to classify a wide range of attributes. For gender very high classification accuracies are reported. A simple language independent approach to gender classification is presented in Alowibdi et al. [59], where the colour scheme chosen for a person's Twitter profile is used to determine their gender. A variety of text based features for user profiling are examined in Weren et al. [60]. Gender is one of the attributes they use to demonstrate how accuracy varies across chosen features. The results of a crowd-sourcing experiment to build a gender and age classifier are presented in Meder [61], Nguyen et al. [18]. An analysis of lexical and some network properties in regards to gender identity on social media is presented in Bamman et al. [62].

Many user profiling works focus on prediction of user gender due to the relative ease with which it can be acquired as an annotation by visual inspection or comparison of names against a lexicon. Gender annotation approaches do not tend to consider the potential of diverse gender identities.

Geographic origin

Various facets of geographic origin on social media have been investigated in the literature, with different approaches attempting to predict location with varying levels of granularity. The approach in Rao et al. [13], Rao and Yarowsky [14] tackles the regional origin problem in the context of distinguishing between northern and southern Indians. The system, which uses a mostly text based approach, with some social media specific features, achieved an accuracy of around 77%.

Geographically linked topic models are investigated in O'Connor et al. [63]. Geographic topic models are presented as an extension to traditional LDA, whereby topics are somewhat hierarchical and may have geographical variants. These topic models were then used as features to perform location prediction in three tasks, a regression task centering on coordinate predication and two classification tasks attempting to predict state (49 classes) and region (4 classes). The approach beats the baselines in all tasks other than state prediction.

Other studies have attempted to profile user location based on their social network as well as their textual output [64, 65]. Pavalanathan and Eisenstein [23] surveys the field of text-based user geo-location based on geo-located Twitter posts as training data. They present an improvement based on a latent-variable model that includes prediction of gender and age in the geo-location process.

Training data for models that predict the geographic origin of a user is difficult to acquire; we present an overview of methods used to gather such data in Section 2.4, and empirically evaluate them in Chapter 5.

Personality

Personality classification with a social network focus is investigated in Golbeck et al. [11, 12]. The effect of the inclusion of network statistics on classifier accuracy was investigated, and produced promising results with mean absolute error (MAE)

scores of 10-15% being reported.

Several interesting concepts (see Section 2.1.3) regarding network statistics and personality prediction are introduced in Quercia et al. [42].

Another attribute classified using the DLA presented in Schwartz et al. [58] was personality.

In Youyou et al. [44] personality predictions from a machine learning system are compared to human judgements on the same dataset. It is shown that a machine learning system based only on Facebook "likes" performs better on average than humans, except when the human annotator is a spouse or close family member of the person being judged.

Niche personality aspects have also been covered. *Dark personality traits* are several socially aversive traits [66], that can be viewed as further aspects of personality. The dark traits that have garnered the most research are: Machiavellianism—the tendency to manipulate others; Psychopathy—a tendency towards a lack of empathy and reckless behaviour; Narcissism—self absorption; and more recently, Sadism—pleasure in the pain of others [66, 67]. These are known as the "Dark Tetrad".

Studies have shown strong correlations between dark tetrad scores and certain types of online behaviour. Most notably, self-defined "internet trolls" (users who behave in a deliberately destructive manner) tend to score highly on the dark tetrad elements sadism, psychopathy and Machiavellianism [68]. A machine learning analysis of dark personality traits was performed in Sumner et al. [17]. The approach uses LIWC classes as features, and compares the results of a variety of classifiers. As previous personality studies have shown LIWC features to be less reliable than others, a further investigation into text, behaviour and network features might be of interest.

Age

The system in Rao et al. [13], Rao and Yarowsky [14] also classifies age. Users are broken down into two categories: above 30 and below 30, and can be distinguished between with an accuracy of 75%. Peersman et al. [4] approach user profiling from a child predator detection angle. The proposed techniques are shown to predict

age categories with accuracies of around 88%. The homophily based approach described in [27] produces better results for age than gender, with an improvement noted over the baseline (single user). This improvement can possibly be attributed to tendency for people of similar age groups to stick together, especially at younger ages. Out of all continuous values estimated in Kosinski et al. [15] age achieves the best results, with $r = 0.75$ for the Pearson correlation coefficient. Age is the third attribute given as an example in Schwartz et al. [58]. Markers of age in social media are investigated in Nguyen et al. [69]. The authors also identify that the accuracy of predicting age becomes much worse for older users in the system, perhaps due to a shift in later life to more standard language. One of the attributes examined in Weren et al. [60] is age. In the crowd-sourcing experiment presented in Meder [61], Nguyen et al. [18] age is one of the attributes collected. The authors discuss future considerations for such an experiment and also identify issues with the data.

Political ideology

Detection of political leaning is also covered in Rao et al. [13], Rao and Yarowsky [14]. The problem is diluted to that of Republicans Vs. Democrats. It is identified that actual talk directly relating to politics is quite sporadic, and therefore other markers must be identified. They highlight that in their dataset possessives such as "my handgun" are quite useful. They also note that certain news networks correlated more with either party. The usefulness of hashtags in predicting the political alignment of Twitter users in investigated in Conover et al. [70]. It is shown that hashtags alone are more useful than the full the full text of a user—at least in their dataset—with accuracies of up to 91% being achieved. Social media specific and text features for political ideology—again Republic Vs. Democrat—are examined in Pennacchiotti and Popescu [43]. The most marked improvement in Zamal et al. [27] from including neighbours can be found in relation to political ideology. Political ideology is also classified in Kosinski et al. [15] and Sylwester and Purver [71]as Republican Vs. Democrat.

Education and intelligence

A weakly supervised approach to attribute detection is presented in Li et al. [72]. The authors use public profiles and information from public databases as distant

sources of supervision to infer entity relations for three attributes: education, employment and relationship status. Additional investigation is also performed into the effect of including social media statistics and relationships in the system, showing improvements in some cases.

The like-based prediction system presented in Kosinski et al. [15] also predicts intelligence. The system predicts accuracy roughly half accurately as the test-retest reliability for the intelligence test taken by users in the study.

Ethnicity and race

The Bayesian model presented in Rao et al. [57] is also used to classify Nigerian ethnicities. As with gender, the model they describe is contrasted to the performance of other machine learning models, and shown to provide a marked improvement on their dataset, achieving around 82% accuracy for Nigerian ethnicity prediction. Despite posing the problem as one of ethnicity, the ethnicities classified are not drawn from highly mixed communities and as a result the problem is closer to one of geographic origin determination.

A mixture of text and social media specific features are used in Pennacchiotti and Popescu [43] to identify between African Americans and other Americans. The approach presented in Kosinski et al. [15] also solves the problem, distinguishing between Caucasian and African Americans.

An approached based on using known county demographics to improve race prediction is presented in Mohammady and Culotta [73]. The approach presented performs similarly to a fully supervised approach, achieving 80% accuracy in predicting user race.

Miscellaneous attributes

As well as race and political leaning Pennacchiotti and Popescu [43] also classify whether users are a fan of Starbucks or not, achieving accuracies of up to 81%. This is the first study identified in this report that attempts to profile personal preference. As well as common attributes, the system presented in Kosinski et al. [15] also attempts to profile some more uncommon ones. The system is able to classify sexuality to a fairly high degree of accuracy, along with the use of the use

of drink and drugs. More surprisingly perhaps, is that whether a user's parents were separated by the time the user reached 21 years old, can be classified with reasonable accuracy of around 60%. A diagnostic application of user profiling is investigated in Resnik et al. [5], where entity relations are used in the prediction of depression in college students. Occupational class and income are predicted in Preotiuc-Pietro et al. [36], Preoţiuc-Pietro et al. [20].

Spammer, bot and fake profile detection

Unlike the authors in hand curated datasets, social media users are not guaranteed to be controlled by human agents or represent the human they claim to represent, as such identifying fake, spam and robot profiles is an important task within user profiling, with applications such as dataset noise reduction, spam filtering and online trust—a user with mostly fake followers is obviously less trustworthy than a one with real fans for example.

Such 'spammer detection' tasks tend to model users as discrete classes such as 'human', 'robot spammer' or 'human spammer'. A wide variety of strategies and tools are available online that are said to be useful in identifying fake or spam profiles; the approach in Cresci et al. [74] uses sets of fake profiles bought by the researchers to distinguish between real and fake profiles. The authors survey the related academic literature, but also concentrate on approaches, algorithms and heuristics proposed by several influential bloggers, arguing they warrant assessment with some level of academic rigour. The bloggers' approaches are compared to those presented in academic literature, and also converted into feature sets for machine learning approaches. Good results are achieved with around 99% accuracy for fake profile detection with the best feature arrangement.

The identification of Twitter profiles controlled by robots is investigated in Chu et al. [75]. Three classes of manually annotated profiles are considered: human controlled, bot controlled and cyborg (human and bot) controlled. With their best feature selection 96% accuracy was noted, unsurprisingly cyborgs were most frequently misclassified. Entropy, the complexity of a process, was shown to be a very useful feature in bot detection, achieving 82.8% accuracy on its own.

2.3 Classical user attribute labelling

There is a vast amount of public user data available on the web, with social media websites such as Twitter providing a large proportion—every day around over 500 million Tweets are Tweeted across the world according to ex-Twitter CEO Dick Costolo. The majority of this data though gives no obvious formal definition of its creators attributes, as such, training data for user profiling systems must be "enriched" by assigning user attribute labels ((2) in Figure 2.1).

There have been a variety of different methods employed to enrich data, with varying levels of success. Many employ a technique called "crowd-sourcing" where data is obtained from a large number of individuals, rather than one supplier. One such example is the myPersonality [15] dataset, which was constructed by allowing Facebook users to complete a personality quiz and then asking them if they wanted to share their data for research purposes at the end. Another example in the same vein is TweetGenie [18], a system which predicts a Twitter user's age and gender and then asks if it was right at the end.

Other studies have attempted to enrich Twitter profiles, based on the information in user's other on-line profiles, identified by links in the Twitter profile [19].

It is possible to pay human annotators to identify attributes of users, although care must be given, as some annotators might be better at identifying information than others. One method to overcome this is to employ multiple annotators to annotate the same piece of information, and only accept annotations where the annotators are in agreement [76].

Heuristic methods, such as through the use of pattern matchers (specifically *regular expressions*) have been attempted also. The approaches in Preotiuc-Pietro et al. [36] and Preoţiuc-Pietro et al. [20] for example build their dataset by crawling description fields for job titles known to correspond to particular occupational classes.

2.4 Geo-location driven user attribute labelling

Geo-location driven attribute labelling is our alternative, novel approach to user attribute labelling ((2) in Figure 2.1), which was formally introduced in Chapter 1 and is described in Figure 1.1. In our method, a user is geo-located and then labelled with attributes based on their local demographic data. This approach relies heavily on the ability to accurately geo-locate social media users.

User geo-location studies are a related field that also rely on estimates of users location, typically referred to as *home location*. In this section, we review the methods used in the user geo-location literature to assign home location to profiles. Note that none of the approaches covered here have been used previously to perform the full process of geo-location driven attribute labelling (Figure 1.1), instead focussing on the left panel of Figure 1.1, home location prediction.

Geo-location approaches typically rely on training data, consisting of profiles labelled with home location information. The definition of *home* often varies from task-to-task, but generally refers to some representation a user's local area. Some approaches use coordinate level locations, but coarse-grained regions are more widely used. Profiles are usually associated with some kind of public boundary data to generate labels and the granularity of these vary between studies. For example, Mahmud et al. [77, 78] associate users with their country of residence, Rout et al. [64], Chandra et al. [79], Chang et al. [80], Cheng et al. [81], Eisenstein et al. [82] with their local city, and Kinsella et al. [83] link users to their ZIP code.

Home location is attached to profiles in various ways. Commonly a profile's location field is resolved to a real-world location through the use of a gazetteer [64, 83], however this approach can be inaccurate. Other approaches to assigning home location make use of geo-located Tweets. Many use a profile's first geo-located Tweet, whilst others use Tweets that provide information about the location associated with the profile [77, 78, 82, 84]. However, these methods are not always reliable. Some approaches take all of a profile's geo-located Tweets into account by applying some form spatial average [85, 86, 87], which is likely to produce better judgements than a single Tweet, but may run into problems when a user is active at multiple locations (such as work and home), depending on the "average" used. The state-of-the-art methods and their limitations are now discussed in more detail.

Hecht et al. [88] report that 66% of Twitter users provide valid geographic information in the 'location' field of their profile; mostly at city level. Of the profiles that do report a location, many are accompanied with non-standard text and abbreviations, such as "kcmo-call da po po", referring to Kansas City, Missouri, and a mixture of fictional and real information, such as "Bieberville, California". In addition, less than 1% of Tweets had associated geo-location information. To overcome this lack of reliable location information, approaches have been developed to geo-locate social media users based on the content of their Tweets.

To carry out user geo-location, a training set of profiles with known home location is required. A method for creating this training set must have certain characteristics to be reliable. Geographic statistical units, such as *output areas* (a UK statistical unit), exist at various scales, with some being less than a square mile in area. As such, a method for creating a training set must be able to provide highly accurate/fine-grained location judgements. Having assigned a user a home location, some measure of 'certainty' is required (i.e. accurate to within an N mile radius), allowing profiles with low certainty to be excluded (or downgraded) from analysis.

Various approaches for generating location-labelled profiles which have been used in previous works are described below.

2.4.1 Location field information

Twitter provides users with the option to declare their location through a non-required free-text field. A number of methods have created training datasets based on the information included in this field. Kinsella et al. [83] predicted location at a number of levels, generating their location labels via the Yahoo GeoPlanet[1] tool, and applying this in a method based on ranking probable location labels by Kullback-Leibler divergence. Rout et al. [64] explore the use of social network features to geo-locate a set of UK Twitter users to the town/city level. Their training data was acquired by selecting profiles with an unambiguous reference to a UK town/city in the profile field. Rahimi et al. [89, 90] combine social network features with text based features to improve geo-location accuracy. Alex et al. [91] present an improved tool for parsing geo-located information in the location field

[1]Yahoo! GeoPlanet `https://developer.yahoo.com/geo/geoplanet/`

and other text, and apply it to geo-locate users discussing a number of topics as well as demonstrating country-wide sentiment analysis.

Due to the fluid nature of this field it can be unreliable, with users providing incorrect, out of date or non-location information, as highlighted in Hecht et al. [88]. Han et al. [92] present a pilot study identifying that despite its noisiness, self-declared user meta-data can help boost performance. In addition to accuracy issues, the task of resolving text to a point location is itself a difficult problem; there are cities called London in the UK, US, Canada and South Africa, all of which speak English as a primary or dominant language.

A limited number of profiles provide coordinate data in their location field rather than a named location. These were used to build the training dataset in Cheng et al. [81], which was used to build a model at the US region/state level based on location-specific terms within Tweets. This approach would appear favourable over named locations as it avoids the place name ambiguity problem, however, it is no longer feasible, as this information was placed by an early mobile Twitter client and thus rarely appears in recent profiles.

2.4.2 First Tweet

The simplest possible method for identifying a user's location is to take the coordinate of their earliest available geo-located Tweet, or the Tweet used to identify them in the first place. Eisenstein et al. [82] use the first geo-located Tweet by each user to label profiles in their training set at the region and city level in the US, applying this to build geographic topic models. Roller et al. [84] also make use of this dataset, as well as their own created in the same fashion, applied to a method based on adaptive grids. Mahmud et al. [77, 78] use the first Tweet coordinate to label the profiles in their dataset at time zone, state and city level in order to train a hierarchical ensemble of classifiers.

The first Tweet method is error prone, as a single Tweet provides little guarantee that a user frequents an area and is therefore unsuitable for hyperlocal home location detection. Despite this Mahmud et al. [77, 78] identify that the majority of users will rarely travel outside of their home city or state, and so if only this low granularity identification is required, first Tweet has a good chance of leading to a correct assignment. Further, steps were taken by Mahmud et al. [77, 78] to

identify and omit those users who frequently travel, avoiding incorrect labelling.

2.4.3 Geometric median

An obvious improvement to the *first Tweet* method described in Section 2.4.2 is to draw on *all* of a user's geo-located posts. Home location of each user is assigned by Jurgens [85], Compton et al. [86], Jurgens et al. [87] as a single coordinate, taken as the geometric median of their geo-located posts L (a variant of the multivariate L_1 median) given by

$$GM = \arg\min_{x \in L} \sum_{y \in L} D(x, y), \tag{2.1}$$

where the distance D is calculated using Vincenty's formula [93]. Median location was used to ensure robustness to outliers, such as those introduced by Tweets produced while the user is on holiday abroad. In Jurgens et al. [87] an additional constraint is also applied to eliminate users with overly spread coordinates. Given the geometric median, the median of distances to all posts of a user is calculated, and those of more than 30 km are discarded. This essentially removes from analysis users where over half their posts are far away from their most central location—flagging that this central location is likely uninformative.

As opposed to using first Tweet or location field information, Jurgens [85], Compton et al. [86], Jurgens et al. [87] produces judgements with the highest rigour as it takes all of a user's Tweets into account in an outlier resistant fashion, and is used in conjunction with a measure of 'certainty'.

2.4.4 Grid based

Han et al. [94, 92] use a worldwide dataset that assigns a user's home location as a coarse representation of population centres they call 'cities'. To construct these 'cities' they take all actual cities in each 'region' (such as state or province) available in the Geonames[2] dataset, and collapse the actual cities within 50km of one another into a single point, producing 3,709 very coarse location labels. To label users the world was split into a grid containing $0.5'$ by $0.5'$ cells, with a Tweet defined as being from a city if the cell containing it, or a surrounding cell,

[2]Geonames: `http://geonames.org/`

also contained a city. The most common city identified across a user's geo-located Tweets was taken as their home location.

This method is well suited for providing coarse location labels in large countries such as the US, with sporadic yet dense areas of population, but is not well suited to smaller regions with more consistent population densities such as the UK and Europe.

2.5 Model fitting

After assigning attribute labels and processing user content into a set of feature vectors, the next step is to train, or "fit", some machine learning model, capable of predicting the assigned attribute labels on unseen users ((3) in Figure 2.1). We do not attempt to undertake a thorough review of machine learning here, and instead provide a brief overview of the classes of machine learning problems relevant to user profiling.

Broadly, predictive machine learning models can be split into two classes: supervised and unsupervised. Supervised machine learning algorithms generally take features vectors and labels as input, and attempt to learn patterns in the data that maximise the chance of making a correct prediction. Unsupervised approaches such as clustering, also take feature vectors as input, but have no preconceptions of a target label to predict, and instead attempt to identify arbitrary patterns or groupings in the input data

User profiling system typically take a supervised approach, using user attributes as labels to predict. There are two main families of supervised machine learning model: *classification* and *regression*. A classification model is used in the case where discrete classes, such as male or female, are up to be predicted. When a label with a continuous value, such as age, is present a regression algorithm is often used, although in some cases attempts are made to convert to a classification problem, for example by splitting the possible values into ranges such as 20 to 30. Trained classification models are often referred to as *classifiers*, and trained regression models as *regressors*.

If more than one label is to be predicted it is referred to as a multi-label classification problem. The most common solution is to train a single classifier or

regressor for each label, although methods do exist to solve the problem with a single classifier, or regressor, that produces multiple outputs [95].

A wide range of machine learning models are used across the user profiling literature with varying degrees of success. We do not attempt to review machine learning models exhaustively, and instead focus on a few examples that appear often and are associated with consistently good performance, and are utilised in our own work.

Support vector machines (SVMs) are a class of supervised machine learning model can be used to perform classification or regression, which is consistently shown to perform well in various user profiling contexts. SVMs attempt to construct a decision boundary in feature space between two classes which is maximally far from any point in the training data [96]. The prevalence of SVMs in user profiling and other text classification problems can be attributed to their robustness and ease of use [97]. Many off-the-shelf SVM implementations exist with bindings to a wide range of programming languages. They often produce comparable or better results [98, 99] than more advanced algorithms, and are often the main algorithm used in a study or used as a baseline to compare an alternate method.

Logistic regression (LR) (also referred to as logit regression, maximum-entropy classification (MaxEnt) and the log-linear classifier) is another popular statistical model. LR is used to model the *probability* of binary classes occurring given some input data using a logistic function [96]. As with SVMs, LR with thoughtfully tuned parameters can often out-perform more advanced machine learning models, and is usually used when a probabilistic output is desired.

Gaussian processes (GPs) are a more recent approach used to handle both regression and classification problems with probabilistic output [100]. GPs work well in lower-dimensional spaces, but are of less use in high-dimensional scenarios such as bag-of-words.

In recent years, deep learning approaches have emerged as strong performers for text classification tasks. Deep learning methods for text classification [101] include: basic feed-forward networks that treat text as a bag-of-words, similar to our classification approaches throughout this thesis; recurrent neural network (RNN) models work with text as a sequence of tokens, and attempt to capture dependencies between tokens based on the structure of the text; convolutional neural

network (CNN) models extract patterns such as key phrases in text; transformers [102] are a more recent advancement that, similar to RNN, process sequences of tokens to learn relationships, and are very suited to transfer learning.

2.5.1 Alternative approaches

Despite the majority of user profiling approaches being machine learning based, other approaches have been considered and attempted that utilise domain-specific heuristics.

A pilot study is performed in Pennacchiotti and Popescu [43] employing a bespoke set of regular expressions (search patterns) that attempted to identify markers of age, gender and ethnicity. The approach looked for sentences such as "I am a 20 year old black man", and achieved inaccurate results, with only 0.1% of profiles containing markers for ethnicity. 80% of profiles did contain gender markers, although they were typically incorrect or conflicting.

Other approaches have used presence of particular entity relations as heuristics for particular attributes, rather than features in a machine learning system. Detection of triples of the entities Person, Opinion and Political Party is employed in Maynard and Funk [103], to determine voting intention of Twitter users in the UK general election.

2.6 Model evaluation

It is typical to evaluate user profiling systems, by applying them to a held out set of labelled data, and calculating performance metrics ((4) in Figure 2.1). For brevity, we do not list all performance metrics here, but define them as they are used in later chapters. System performance is also often compared to human performance on the same problem, contrasting the output of an automatic system, to result of human cognitive processes.

2.6.1 Train - validation - test

A common approach is to split the labelled training data into three sets: train, validation, and test. The train data is used to train the model as the name implies. The validation set, is used while the model/feature sets are being actively developed, and is used to fine-tune the model parameters and perform feature selection. The test set is finally used to calculate performance metrics for the model after feature selection and parameter tuning.

2.6.2 K-fold cross validation

Another approach to evaluating a model on held out data is K-fold cross validation. K-fold cross validation typically consists of splitting the training data into K pieces or "folds", and training the machine learning algorithm K times, leaving a different fold out each time. The resulting model after each iteration is used to predict the labels of the left out observations, and scored based on its ability to predict known labels accurately [104]. In cases where a continuous value is predicted, the system is scored based on a measure of how close the predicted values are on average to the observed ones.

2.7 Ethical considerations

Fields such as medicine are bound by ethical codes that stipulate any experiment must minimize the risk of harm to participants [105]. These codes recognise inherent value in experimentation on human subjects, while preventing any lines of research that could be exploitative or harmful to subjects (or the wider population). Decisions and oversight on matters of ethics are typically handled by Institutional Review Boards (IRBs). Practitioners of the data sciences (such as machine learning (ML), NLP, and computational social science (CSS)), are not (in general) bound by such strict codes (depending on their institution), and as such are likely to only seek IRB approval when working with human participants directly. In recent years however the data science community has started to acknowledge the potential societal impacts of their work [106, 107], in part due to high profile data misuse scandals such as Cambridge Analytica [108]. Several re-

spected academics have even introduced ethics education into their ML and NLP curricula [109, 110, 111]. Ethical approval was sought and granted for the works presented in thesis, and steps were taken to ensure privacy of users represented in created datasets.

2.8 Conclusion

In this chapter we have reviewed the literature surrounding the field of user profiling, covering both the attributes addressed in previous works, and the approaches used for annotation and in derived user profiling systems. In general a classic machine learning approach is used to develop user profiling systems; user content is annotated with user attributes, then transformed into feature vectors and used to train a machine learning model capable of inferring the user attributes of unseen users. Most successful approaches in the literature extract bag-of-words features from user content, and use these to train linear models such as SVM or LR; this general approach builds the foundation of the user profiling systems in our own experiments in later chapters.

The main contributions presented in thesis with regards to previous work focus on the acquisition and generation of novel user profiling datasets that stand as user profiling datasets in their own right, and also as complementary resources. Geo-location driven user attribute labelling was introduced in Chapter 1 as a new method for generating user profiling datasets. Application of geo-location driven user attribute labelling is unsuitable for demographic variables that vary fairly uniformly across regions, such as gender and age, and instead can applied to variables that exhibit more variation, such as socio-economic status, and aggregate measures that attempt to distil the broader demographics of an area into easy to interpret variables. Through this method, three novel datasets were generated covering two sets of user attributes previously unadressed in the literature, Output Area Classification (OAC) and Local Authority Classification (LAC) (Chapter 6), and one that has been investigated through a dataset generated via a different approach, National Statistics Socio-economic Classification (NS-SEC) (Chapter 7).

Geo-location driven user attribute labelling relies on the ability to accurately assign estimates of 'home location' to user profiles. Although previous works have not performed geo-location driven user attribute labelling, several works have at-

tempted to assign home location to profiles for the purpose of creating user geo-location training data. We build on four of the approaches for assigning home location to user profiles surveyed in Section 2.4, which are evaluated alongside two novel methods in Chapter 5.

Chapter 3

Topic models and n-gram language models for user profiling

In this chapter we explore and apply several techniques from Chapter 2 to an existing user profiling dataset from the *PAN 2015 Author Profiling Shared Task* to predict personal attributes. The aim of this chapter is to identify a valid baseline approach for our later experiments. Four distinct corpora, each in a different language, were used in the experiments presented here. Each corpus consisted of collections of Tweets for a number of Twitter users whose gender, age and personality scores are known. Given these corpora, the task is to construct some system capable of inferring the same attributes on as yet unseen users.

We propose and evaluate a system which utilizes two sets of text based features, *word n-grams* and *topic models*, in conjunction with support vector machines (SVMs), to predict gender, age and personality scores. We applied our system to each dataset and achieved results indicating that n-grams and topic models are effective features across a number of languages.

This chapter is an expanded version of work presented at PAN 2015 [112].

3.1 PAN 2015 Author Profiling dataset

For the Author Profiling task at PAN 2015, a set of Twitter users whose gender, age and personality are known was provided. In this task Twitter users are represented

as a collection of the text content of their Tweets. The users are further divided into four languages: Italian, English, Dutch and Spanish, yielding four distinct corpora. The task is to develop a system for each language that when given a set of unseen users, can make some judgement of age, gender and personality [113]. The system does not have to infer the native language of the user, as this is provided at both train and evaluation time.

Detailed characteristics of the dataset are presented in Table 3.1. The corpora are balanced by user gender, such that there is an equal number of male and female users present in each corpus. There is no guarantee that each user has the same number of Tweets, and as such over-fitting to particular users is a risk. Gender in this task is a classification problem; binary selection of male or female, additional gender identities are not considered in this dataset.

The task of predicting user age can be addressed either as one of predicting a continuous variable such as: a continuous variable (age in years) or a categorical variable (pre-defined binned ranges). In this case age prediction has been converted to a classification problem, where a range of ages is to be predicted rather than a continuous value. The defined ranges are: 18-24, 25-34, 35-49, and 50+; younger users are not contained in this dataset. The Italian and Dutch corpora do not include age annotations at all and as such we do not attempt to predict age for these languages. There is definite imbalance within the age groups for the English and Spanish corpora; within the English users, the two categories 18-24 and 25-34 account for most of the data; within Spanish, the 25-34 group contains more examples than the other three age groups combined.

Personality is often measured using the so-called "Big 5" or Five Factor Theory personality traits [114]. The five traits are: extroversion (E), emotional stability/neuroticism (S), agreeableness (A), conscientiousness (C), and openness to experience (O). Users in the dataset are annotated with their self-reported "Big 5" personality scores from the short-form BFI-10 online test [115]. Scores are normalised to the range of -0.5 to 0.5. All of the "Big 5" traits are included in the dataset. The bottom section of Table 3.1 shows the mean score of each personality attribute in the dataset by language; many of the mean attribute scores are towards 0.5, indicating a skew towards more positive values.

	Training				Test			
	EN	**ES**	**IT**	**DU**	**EN**	**ES**	**IT**	**DU**
Users	152	110	38	34	142	88	36	32
18-24	58	22			56	18		
25-34	60	56			58	44		
35-49	22	22			20	18		
50+	12	10			8	8		
Male	76	55	19	17	71	44	18	16
Female	76	55	19	17	71	44	18	16
E (mean)	0.16	0.18	0.17	0.24	0.17	0.16	0.15	0.24
S (mean)	0.14	0.07	0.2	0.21	0.13	0.09	0.2	0.22
A (mean)	0.12	0.14	0.22	0.13	0.14	0.14	0.19	0.15
C (mean)	0.17	0.24	0.18	0.14	0.17	0.21	0.21	0.17
O (mean)	0.24	0.18	0.23	0.29	0.26	0.19	0.25	0.28

Table 3.1: Detailed characteristics for each corpus in the PAN 2015 Author Profiling dataset, showing the number of users represented in each class for age and gender, and the mean score for each personality trait, separated by language. The Italian and Dutch corpora do not include age annotations. Adapted from Rangel et al. [113].

3.2 Approach

In Chapter 2, we noted that user profiling studies utilising word n-grams in a bag-of-words fashion to train linear models, such as logistic regression or support vector machines, frequently perform well on a variety of other NLP tasks, as such we chose to evaluate this further, and base our approach on this solid foundation. We also noted that topic models, a group of algorithms that identify hidden themes (topics) in collections of documents, have been shown to produce reliable results when used alone and in conjunction with other features [43, 58], and as such included them in our approach so that we could evaluate their utility when applied to the field of user profiling. Parts-of-speech (POS), are a common feature in natural language processing (NLP) pipelines [45]; we found that at the time these experiments were performed, POS tagger performance on social media texts was poor (or non-existent) for most languages except English. Nevertheless, we evaluated the use of a Twitter-focused tagger [116, 117, 118] for the English component of the dataset.

The architecture of our developed approach is presented in Figure 3.1. The system comprises two main components: a model generation module, and one which uses

a pre-trained model to infer the attributes it contains on unseen documents.

For model generation the training data is fed first through several pre-processing steps and then through several feature extraction modules. Firstly, a latent Dirichlet allocation (LDA) model is trained which is then used in the "Topic Extraction" module. The same data is also passed through an "n-gram Extraction" module. The resulting feature vectors are normalised, concatenated, and used to train a machine learning model. Thorough implementation details follow in Sections 3.2.1 and 3.2.2.

The machine learning algorithm used to generate the results presented in this chapter is SVMs as they have been consistently shown to produce good results in user profiling classification tasks, often achieving better or comparable results over more modern methods such as deep learning in some contexts (discussed in Section 2.5). We performed ad hoc experiments with ensemble methods and other machine learning models, but none beat the results achieved by the SVM implementation, and so are not presented here.

For age and gender a support vector classifier with a linear kernel was used. For the personality recognition element support vector regressors were used, again with a linear kernel. The linear kernel was chosen in both cases as it is well suited to high dimensionality problems such as text classification, which tend to be linearly separable [98, 97]. All implementations were provided in Scikit-learn [119].

We perform a feature ablation study through 10-fold cross-validation on the training data in Section 3.2.3 to assess the usefulness of each feature set, and report the overall effectiveness of our approach on the test set in Section 3.3.

3.2.1 Pre-processing

The text of the Tweets provided proved to be quite clean and little pre-processing was required other than tokenisation (splitting the text into semantic units such as words) using a Twitter-aware tokeniser [116], and simple text normalisation steps. A stop-list was not used to filter low information content, common tokens from the Tweets due to the multi-lingual nature of the dataset, and the tendency for more informal language to be used on social media. Instead all tokens that were used by more than 70% of the users in a given language were treated as stop-words, as

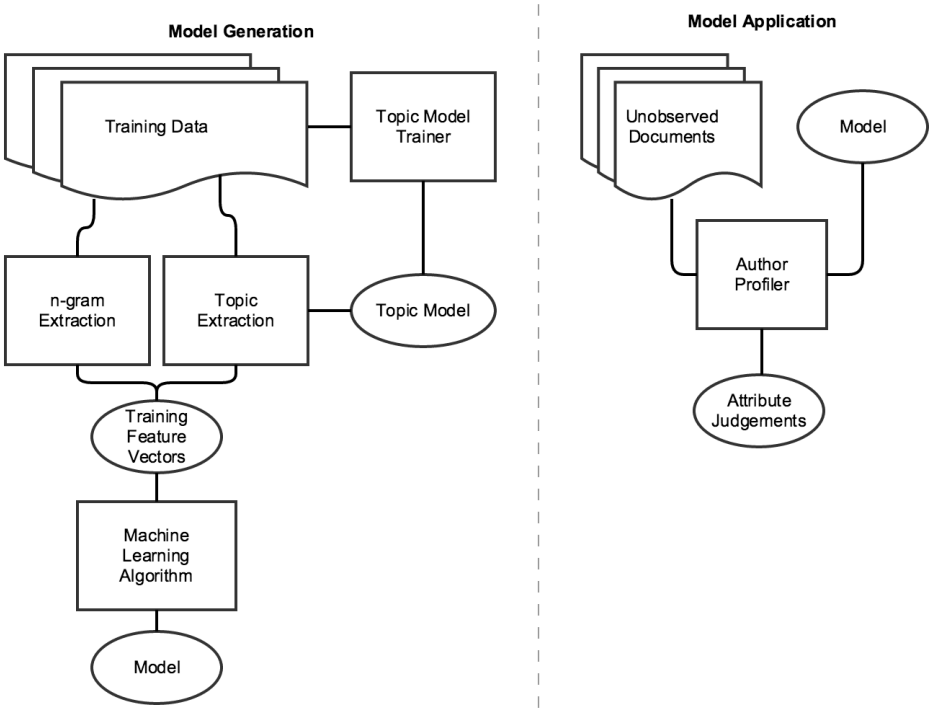

Figure 3.1: Architecture of the user profiling system presented at PAN 2015.

this is a roughly analogous, language independent technique to stop-word removal. 70% was chosen as the cut-off point for frequent term removal due to the small size of the dataset, in larger datasets more representative term frequency statistics can be calculated, and a cut-off around 90% is suitable.

In early experiments on the data, all hyperlinks present in the text were followed and converted to the domain name of the website found, as previous user profiling studies have identified website use as a potential analogue for some attributes [120, 70]. This was discarded in the final approach as no improvement could be noted with its inclusion, instead a similar experiment was performed to replace all links with a single "link present" token, but again no improvement was noted.

The Twitter-specific step of eliminating "retweets" was also performed, although as the provided data contains very few retweets, likely due to curation by the dataset authors, this step was mostly unnecessary. Another consideration is that some Tweets are in the form "shared via some app" and do not register as retweets in the data returned from the Twitter API; we do not attempt to filter for Tweets of this nature in the scope of this shared task. In most other Twitter processing tasks, especially those 'in the wild', these steps would be included to ensure any

text processed is the users own.

3.2.2 Feature extraction

n-gram extraction and weighting

Following the application of the pre-processing steps described in Section 3.2.1, each Tweet is represented as a sequence of tokens, roughly representing individual words in the Tweet. n-gram representations are generated programatically from these sequences by simply iterating across the sequence of tokens, recording the current token, and following tokens up to the value of n away.

We noted through exploratory experiments that unigrams and bigrams together produced the most reliable results and as such would form the basis of any system developed. Higher order n-grams may well prove more useful in the context of a larger dataset, but due to the limited size of the one used here, trigrams and above were commonly found to have very few occurrences, and were therefore of little practical use for classification in this case.

Tweets are aggregated to the user level in a bag-of-words fashion after n-grams are extracted, i.e., a mapping of n-gram to count is calculated for each user (illustrated in Table 3.2).

User 25	
n-gram	**Count**
fishing	21
cinema	17
big catch	12
jog	9
⋮	⋮
watching tv	3
soup	1

User 68	
n-gram	**Count**
the tube	34
matinee	27
cinema	20
watching tv	17
⋮	⋮
jog	3
quiet night	1

Table 3.2: Aggregated Tweets for two fictional users represented in a bag-of-words fashion.

When bag-of-words features are used to train a model, frequent terms can often attract higher weights than infrequent but more information-rich terms, to the

User 68		
***n*-gram**	**Count**	**Weighted**
the tube	34	1.24
matinee	27	1.16
cinema	20	0.34
watching tv	17	1.16
⋮	⋮	⋮
jog	3	0.565
quiet night	1	0.913

Table 3.3: TF-IDF weighted n-grams for fictional user 68.

overall detriment of the model. We address this by re-weighting the n-grams at the user level with the term frequency-inverse document frequency (TF-IDF) term weighting scheme. In TF-IDF, a term's rating is based not only on its frequency in a document, but also against how common the term is in the whole set of documents, reducing the weighting of very common terms and increasing the weighting for more uncommon terms [45, 46]. Table 3.3 illustrates an example re-weighting of the terms used by example user 68 in Table 3.2; note how certain terms have changed weight such that their rank would also change.

Topic model

Topic models are a class of techniques used to identify latent themes, referred to as topics, in collections of text. Here, we apply a topic modelling technique called latent Dirichlet allocation (LDA) [33]. LDA is a generative model in which documents are modelled as a finite mixture of topics, such that each word in a document must be generated by one of its topics. An example application of topic modelling to English social media texts, showing the most important words identified for each automatically derived topic, can be seen in Figure 3.2.

In our approach, during the training process, an LDA topic model is trained (using the library gensim [121]) on the Tweet texts from the training set for each language separately, with a target of 10 topics. We performed ad hoc experiments to select the number of topics for the LDA models; due to the small size of the corpora we found that increasing the number of topics above 10 resulted in thematically

Topic 0: nbsp, amp, time, il, friends, dog, just, work, like, baby, help, ll, girl, okay, want
Topic 1: people, urllink, new, bush, president, use, read, information, war, kerry, america, make, government, time, american
Topic 2: life, love, god, time, world, heart, way, know, feel, like, things, people, make, let, away
Topic 3: im, lol, dont, like, haha, oh, got, gonna, yeah, went, today, thats, cuz, didnt, ok
Topic 4: just, like, know, don, really, think, people, want, ve, time, going, say, things, ll, good
Topic 5: years, women, men, said, people, man, year, world, city, children, old, woman, white, story, young
Topic 6: urllink, com, pm, site, www, http, den, blog, link, 10, 12, 11, 2004, check, new
Topic 7: day, time, going, today, work, good, school, week, ll, got, really, just, year, ve, days
Topic 8: movie, good, love, like, music, new, great, oh, song, night, band, best, favorite, book, watch
Topic 9: went, got, like, just, home, said, night, didn, time, did, came, car, house, little, room

Figure 3.2: Example topics from a 10-component LDA model trained on social media texts.

similar terms being separated into their own topics in a non-desirable fashion.

The trained LDA models are applied to the aggregated Tweet text for each user, to infer $\mathbf{X}$, the probability distribution of each topic across their Tweets. The vector $\mathbf{X} = (X_1, \ldots, X_{10})^T$ is transformed to the binary 10 element feature vector $\mathbf{Y} = (Y_1, \ldots, Y_{10})^T$ with $Y_i = \mathbb{I}(X_i > 0.1)$, where $\mathbb{I}$ is the indicator function, representing 'presence' of each topic. This 'threshold' of 0.1 was set through experimentation. For example, the vector

$$(0.012, 0.026, 0.011, 0.015, \mathbf{0.424}, 0.013, 0.020, \mathbf{0.414}, 0.043, 0.023),$$

becomes

$$(0, 0, 0, 0, \mathbf{1}, 0, 0, \mathbf{1}, 0, 0).$$

It is entirely possible to use $\mathbf{X}$, the raw probability distribution, as a feature vector in its own right, but in this case we found that conversion to a binary vector yielded a small performance boost.

It should be noted that ideally the LDA models would be trained on large external corpora, within the same domain, to produce more robust topic models; in the scope of these experiments we aimed to assess various techniques *without* the introduction of external resources. In Chapter 4, we evaluate the use of external corpora in a similar context.

Parts–of–speech

In early experiments all Tweets were POS tagged as part of the pre–processing step using a Twitter specific part–of–speech tagger [116], which uses a reduced set of POS tags more suited to noisy social media texts. POS feature vectors were generated in the same fashion as n-gram features, tags were treated as unigrams and aggregated to the user level in a bag-of-words fashion.

We noted through our feature ablation study (Section 3.2.3) that POS tags were a useful feature for user profiling in social media texts, in line with other studies [13, 58]. Nevertheless, we chose not to include this feature set in our final approach, as the POS tagger used was English specific, and as such would not be compatible with the other three languages. Further work could be implemented to to examine

their affect on non-English results.

3.2.3 Assessing feature importance

Prior to finalising our approach we sought to assess the efficacy of several feature sets, both alone and in unison, through a feature ablation study. As no development set was provided, and due to the relatively small size of the training set, 10-fold cross validation was employed to assess the affect of different features on classifier accuracy, without removing too much of the dataset at once. Results from the feature ablation experiment are presented in Table 3.4, and represent the mean result over 10 folds. The feature(s) with the best score for each attribute for each language is highlighted in bold.

Results are presented in each language for n-gram features, LDA features, and the two in conjunction. In the English case, results for POS tagged n-grams are also included. These results show POS tagged n-grams as being the best feature for English gender and age prediction; despite this they were not used in the final approach, as a comparable POS tagger could not be found for Spanish, Dutch and Italian Tweets.

In most cases n-gram features provided the best results, but not by a significant margin, with n-grams in conjunction with LDA topics performing similarly. LDA topics on their own proved to be a very poor quality for the English and Spanish datasets, and gave the worst results in all cases.

Our finalised approach includes n-grams in conjunction with LDA topics, as these judgements proved to be more stable across folds than n-grams on their own.

3.3 Results and discussion

Our proposed approach was submitted to, and evaluated at, the PAN 2015 Author Profiling Shared Task [113]. A remote virtual environment was provided by the organisers to facilitate reproducible evaluation, whereby a deployed system could be applied to the training and test sets in a controlled environment. The results of the final system run submitted to PAN 2015 are presented in Table 3.5. The system performed best on the Italian dataset, achieving a global score above 0.8, where

scores for other submitted systems ranged from 0.8658 to 0.6024. The English and Spanish corpora scores for other approaches were in the ranges 0.7906 to 0.5217 and 0.8215 to 0.5049 respectively, with the results obtained by our system falling roughly in the middle of these ranges. The worst performance was obtained for the Dutch dataset, scoring on the bottom end of the range 0.9406 to 0.6703.

In most cases the final results are worse than those observed by applying cross-validation to the training data. However similar or better results were observed for some personality elements across languages. English age prediction and Spanish gender prediction also achieved reasonable scores compared to the cross-validation.

The results show that n-grams and topic models are a useful element in developing user profiling systems across a number of languages and provide reasonable results without any additional features, or external corpora. In order to improve the system without adding any other features the LDA topic model could be trained on a large external corpus of text, in theory leading to a more robust model.

3.3.1 Other approaches

In the rankings for the PAN 2015 Author Profiling shared task [113], our approach achieved 9th place out of 22 entries for joint prediction of gender and language variant. Of the top performing approaches submitted to the task all utilised n-grams in some form, alongside complementary or derived features.

The best performing approach, submitted by Alvarez-Carmona et al. [122], utilises an approach similar to our own. Second Order Representation (SOA) features, defined as selected discriminative content features such as content words, punctuation, and function words, are combined with TF-IDF n-gram vectors reduced in dimensionality through latent semantic analysis (LSA) to train a SVM model. LSA features, in this context, are analogous the the topics derived in our own approach through LDA, although much more better performing in this task. It is likely that the better performance of Alvarez-Carmona et al. [122] against our own similar approach is down to task-specific feature selection, and more discriminative features learned by LSA on the small training dataset than the LDA model applied in our own approach. Similarly González-Gallardo et al. [123] used combinations of character and POS n-grams, and Grivas et al. [124] combined TF-IDF n-grams with a set of style-based features such as readability and text structure.

3.3.2 Future work

Given the similarity of our approach to the best performing approaches in the task, it is clear that the TF-IDF n-grams are a reliable starting point in user profiling tasks, and task specific feature selection and auxiliary feature sets are the sensible next step to improve performance. A number of potential options for additional feature sets are discussed below.

The way an user behaves in the context of interacting with their medium (be it social media, conversation or essay writing) has, in other studies, been telling of their characteristics. For example, according to the "big five" model of personality an extroverted person is likely to be more outgoing, assertive and have a positive demeanour [37]. Conversational elements have also been shown to be useful [39, 40, 41].

It is also possible to code for behaviour in online media. Studies have identified varying patterns of social media activity times in areas of high and low unemployment, with those low employment areas seeing a sharp rise in posts around the start of the working day [32]. Other studies have attempted to detect conversational behaviours on social media, as earlier research showed them to be of use for user profiling.

An analysis of an user's social network can also give rise to interesting judgements about them. It has been shown for example, that the presence of certain "Likes" made by an author on the platform Facebook, can be indicative of wide number of characteristics. Other social network properties may also be useful, in Quercia et al. [42] four distinct groups of users, where each group has similar personality scores, were identified, based on a user's tendency to follow, be followed and favourite Tweets on Twitter.

For the purpose of this task hovever, these techniques were not further investigated, due to the format of the provided data, although in future it would be very interesting to assess their effect on system performance.

3.4 Conclusion

In this chapter we have presented our findings regarding the effect of the inclusion of LDA topics in conjunction with traditional text features. We used support vector machine classifiers and regressors in conjunction with n-gram and topic features, in order to provide judgements on age, gender and personality. We conclude the addition of LDA topics does improve system performance in most cases, and thus would form a good candidate component of a user profiling system. We propose that further performance improvements could be achieved through the inclusion of external corpora, and go on to evaluate this in Chapter 4.

Our findings indicate that an approach incorporating n-grams and topic models, in conjunction with a linear model, do indeed form a good basis of a user profiling system, and as such use similar set-ups to assess our novel data generation techniques (Chapters 6 and 7).

		Accuracy		Root Mean Squared Error				
Language	**Features**	**Gender**	**Age**	**E**	**N**	**A**	**C**	**O**
English	*n*-gram	**0.7754**	0.7245	**0.1510**	**0.1876**	0.1568	**0.1410**	**0.1281**
	LDA	0.5062	0.4683	0.1949	0.2424	0.1776	0.1686	0.1625
	n-gram + LDA	0.7500	**0.7438**	0.1559	0.2010	**0.1522**	0.1422	0.1327
	POS	*0.7758*	*0.7829*	*0.1561*	*0.2026*	*0.1700*	*0.1443*	*0.1348*
Spanish	*n*-gram	**0.8800**	**0.7300**	**0.1501**	**0.1691**	**0.1426**	**0.1468**	**0.1520**
	LDA	0.5400	0.4100	0.1715	0.2469	0.1795	0.2199	0.1967
	n-gram + LDA	0.8000	0.7200	0.1537	0.1831	0.1502	0.1617	0.1550
Dutch	*n*-gram	**0.8250**	N/A	**0.1112**	**0.1754**	**0.1374**	**0.1039**	**0.1123**
	LDA	0.7083	N/A	0.1618	0.2366	0.1873	0.1355	0.1470
	n-gram + LDA	0.7083	N/A	0.1307	0.1845	0.1476	0.1162	0.1165
Italian	*n*-gram	**0.8500**	N/A	**0.1208**	**0.1600**	**0.1283**	**0.1110**	**0.1377**
	LDA	0.6000	N/A	0.1963	0.2602	0.2150	0.1565	0.2441
	n-gram + LDA	0.7083	N/A	0.1461	0.1670	0.1492	0.1190	0.1442

Table 3.4: Feature ablation table shows average classifier accuracy and mean squared error results derived through 10-fold cross validation on training data.

			Accuracy			Root Mean Squared Error				
Language	**Global**	**RMSE**	**Gender**	**Age**	**Joint**	**E**	**N**	**A**	**C**	**O**
English	0.6743	0.1725	0.6901	0.7394	0.5211	0.1381	0.2223	0.1918	0.1749	0.1352
Spanish	0.6918	0.1619	0.8409	0.5909	0.5455	0.1669	0.2285	0.1398	0.1412	0.1329
Italian	0.8061	0.1378	0.7500	N/A	N/A	0.1279	0.1923	0.1257	0.1187	0.1243
Dutch	0.6796	0.1409	0.5000	N/A	N/A	0.1752	0.1511	0.1444	0.1344	0.0993

Table 3.5: Results of final software submission including global and individual attribute performance values.

Chapter 4

Enhancing user profiling performance with geographically derived resources

In this chapter, we expand upon the exploration of user profiling system components presented in Chapter 3, and make our first foray into deriving user profiling resources from geo-located Tweets and demographic data. We present an approach to generating representative language resources using geo-located Tweets and demonstrate their utility in a predictive setting through our approach to the 2017 edition of the PAN Author Profiling shared task [125, 126, 127]. The PAN 2017 Author Profiling shared task provided a dataset of Twitter users across four languages annotated with their language variants (at the country level) and gender (further details are described in Section 4.1.1).

We constructed a corpus of worldwide Tweets and filtered for those geo-located within the countries covered in the task's languages (except for the Arabic language variants due to low frequency). This corpus was divided into individual languages (Portuguese, English and Spanish) and used to derive Word2Vec word embeddings [34, 128] tailored for each language.

We assessed the utility of the derived word embedding models by evaluating their effect when included as part of a classification pipeline in an experiment on the PAN 2017 Author Profiling dataset. To predict gender and language variant, we applied an ensemble of probabilistic machine learning classifiers (described in de-

tail in Section 4.3). Each set of language specific word embeddings were clustered using K-means to derive a set of word to cluster mappings, which can be thought of as roughly analogous to topics in a topic model. The normalised frequency of each word cluster across a user's Tweets was used to train a Gaussian process classifier. In parallel, a logistic regression classifier was also trained using term frequency-inverse document frequency (TF-IDF) transformed unigram and bigram frequencies. The two classifiers were employed in an ensemble approach by averaging the predicted probabilities of both classifiers for each sample to determine the target user attribute.

Our ensemble of classifiers bolstered by a geographically derived resource performed well across the board, achieving good accuracy scores in both gender and native language recognition far above a random baseline. When compared to a strong baseline established from our work in Chapter 3, support vector machines (SVMs) trained on TF-IDF n-grams, our approach yielded performance increases for a number of attributes, and performed on-par for others.

Our experiments led to two main conclusions:

- Geo-located social media posts can be used to create high quality embedding models, specific to both the medium and locations the researcher wishes; and
- Resources derived through geo-located posts can be used to improve performance in a downstream task.

These observations show that there is indeed a benefit to introducing geographically derived resources in user profiling. In future chapters, we will evaluate this further, by deriving user profiling corpora entirely from geo-located social media.

This chapter is an expanded version of work presented at PAN 2017 [129].

4.1 Data

In order to evaluate the utility of language resources derived from geo-located social media posts when applied to user profiling, we gathered two datasets; a user profiling dataset—the PAN 2017 Author Profiling Shared Task dataset (Sec-

tion 4.1.1), and a collection of geo-located Tweets covering English, Spanish, and Portuguese, these are described in the following.

4.1.1 PAN 2017 Author Profiling dataset

The PAN 2017 Author Profiling Shared Task focussed on the prediction of country level language variant and gender of social media users, specifically, Twitter users. Four languages are covered in the task; English, Spanish, Portuguese, and Arabic. For each *language*, a number of country/large region level *variants* are represented (e.g. Brazilian vs Portuguese Portuguese).

A dataset was provided consisting of Twitter *users* represented as a collection of their 100 most recent Tweets and their profile meta-data at the time the Tweets were collected. Within each top-level language, users are balanced by gender and language variety. The distribution of labels in the dataset, including a full list of language variants included can be seen in Table 4.1.

To construct the dataset, unique users were selected from a collection of Tweets posted in locations identified as representative for a given variety, typically a capital city (e.g. Dublin for Ireland and Cairo for Egypt). Each unique user's historical Tweets were then collected, and users were labelled with a top level *language* and a *language variant* based on the country/region the majority of Tweets were posted from. Users whose Tweets were not in the representative place, those with too few Tweets, and those with too many Tweets in a different language were discarded.

Only Tweets authored by the user were included in the dataset (retweets were excluded), and Tweets were checked to ensure they matched the specified language for each user profile. Gender was annotated manually for each user by the dataset authors, assisted by use of a dictionary of proper nouns to produce an initial "best-guess".

The dataset was further split into *train* and *test* sets in a stratified manner, with 60% dedicated to training and 40% dedicated to testing. A second testing set was made available via the shared task's evaluation platform, but was not made available for download.

Language	Variety	Males	Females	Total
Arabic	Egypt	500	500	1000
	Gulf	500	500	1000
	Levantine	500	500	1000
	Maghrebi	500	500	1000
	All	2000	2000	4000
English	Australia	500	500	1000
	Canada	500	500	1000
	Great Britain	500	500	1000
	Ireland	500	500	1000
	New Zealand	500	500	1000
	United States	500	500	1000
	All	3000	3000	6000
Spanish	Argentina	500	500	1000
	Chile	500	500	1000
	Colombia	500	500	1000
	Mexico	500	500	1000
	Peru	500	500	1000
	Spain	500	500	1000
	Venezuela	500	500	1000
	All	3500	3500	7000
Portuguese	Brazil	500	500	1000
	Portugal	500	500	1000
	All	1000	1000	2000

Table 4.1: PAN17 dataset characteristics.

Dataset limitations

Representative places were chosen by the dataset authors as the method of filtering user location to attempt to reduce the affect of out-of-region users (e.g. tourists Tweeting from Portugal), nevertheless limitation of language variety to specific locations within a larger whole does run the risk of excluding large swathes of users who speak a language variety, but in a different part of the country/region with a different dialect. Scotland, for example, despite being notable on Twitter internationally for their use of distinctively Scottish language [130], is part of "Great Britain", but has capital city Edinburgh chosen as its representative place, over the larger city Glasgow.

Prior to commencing experiments with this dataset native speakers of each language assessed a small number of Tweets on an ad-hoc basis and found that, in general, English, Spanish, and Portuguese Tweets were representative of that language. The Arabic profiles though were found to contain numerous posts in Quranic Arabic, introduced through sharing of quotes and preacher profiles, which is not truly representative of the Arabic used casually in any of the regions addressed.

Motivation for selection

Despite its limitations, this dataset is a good fit for our initial experiments on improving user profiling with geographically derived resources. It builds on the characteristics of the dataset used in Chapter 3, so we can assume our conclusions on strong baseline user profiling components carry over. This dataset on the other hand includes a user attribute intrinsically linked to geography: regional language variant; which allows us to explore building additional corpora that might help address the same attribute. Furthermore, the dataset is much larger and each user in the dataset has the same number (100) of Tweets, which means derived models will likely have much more generalisability and avoid over-representation of more prolific users.

English (F_{en})	**Spanish** (F_{sp})	**Portuguese** (F_{pt})
Australia	Argentina	Brazil
Canada	Chile	Portugal
Great Britain	Colombia	
Ireland	Mexico	
New Zealand	Peru	
United States	Spain	
	Venezuela	

Table 4.2: Countries scraped for each language.

4.1.2 Geographically filtered Tweets

To enable our experiments related to generating resources from collections of geo-located social media posts, three corpora of geo-located Tweets were built: English (F_{en}), Spanish (F_{sp}), and Portuguese (F_{pt}). A comparable corpus was built for Arabic, but was not used in further experiments due to issues with data quantity and the quality issues discussed in Section 4.1.1.

The corpora were derived from a set of all Tweets posted worldwide throughout 2015, collected through the Twitter Firehose[1]. This collection was filtered based on country boundaries (for coordinate based geo-tags) and textual names (for text based geo-tags) to only include posts geo-located in the specific language regions covered in the PAN 2017 Author Profiling Dataset (see Table 4.2). The language property present in the metadata of the Tweets was used to exclude Tweets outside of each targeted regions' main language, but Tweets were not manually checked for correctness due to the scale of the data, therefore a proportion of Tweets is likely to contain mixed language use (code switching) or simply be misclassified. We chose not to filter the corpora further to only include Tweets from representative places as we wished to get a broader picture of language used in the regions as a whole. We expect most out-of-region user influence to be of negligible bias due to the quantity of Tweets involved.

Some language variants were less frequent in the resulting datasets than others due to differences in population, for instance we collected very few Tweets from Ireland compared to the U.S.A. Down-sampling was used to avoid over representation of

[1]Twitter Firehose has since been discontinued and can no longer be accessed.

the more prevalent language variants. Data for the language variants with the largest volume of Tweets was reduced so that it contained no more than 10 times the number of Tweets of the smallest language variant. 10 to 1 was chosen as the upper limit for ratios between languages in our datasets over a smaller ratio, such as 1 to 1, so as not to reduce the quantity of Tweets in the datasets so much that the quality of derived word embeddings would be affected, while still ensuring less frequent language variants have a significant presence in the datasets.

4.2 Tailored word embeddings

Word embedding refers to the process of mapping tokens (such as words, sub-word units, or sequences of words) into a fixed length real valued vector. Typically, word embeddings map tokens from some high dimensional vector space into a much lower dimension. Examples include:

- Bayesian methods such as latent Dirichlet allocation (LDA), which learns to map bag-of-words representations into latent clusters which resemble topics or themes;
- Word2Vec, a shallow 2-layer neural network which via learning to reconstruct contexts of words manages to capture their semantic and syntactic properties;
- GloVe [131], a matrix factorisation approach trained on word-word co-occurrence statistics which also captures semantic information; and
- Brown clustering [132], which utilises a statistical model to cluster terms in a corpus, again, based on their co-occurrence with other words in the corpus.

Embedded representations of tokens are a valuable tool in the natural language processing (NLP) practitioner's arsenal; they can be used as features in their own right, as we showed via our application of LDA features in Chapter 3, or as the input to some advanced feature extractor such as a convolutional neural network (CNN). An embedding model trained thoughtfully is easily reusable and distributable, and can greatly increase performance of downstream tasks; usage of pre-trained Word2Vec embeddings for example introduced a step-change in achievable accuracies for a wide variety of NLP tasks, with a low barrier to entry.

Word embedding models widely available at the time of writing tended to be trained on traditional media such as books and news in US English. These models are well suited to many tasks, but can fall down when applied to other variants of the same language, and in the context of noisy user generated text as is present on social media platforms. In such novel contexts, it is important to either tune an existing embedding model or train a new one to be representative of the type of language being modelled.

We found that for short-form social media platforms such as Twitter, no off-the-shelf embedding model adequately captured the sorts of language used on those platforms, and as such chose to create these resources ourselves by leveraging geo-located Twitter posts.

4.2.1 Developed resources

Word embeddings for each language dataset (F_{en}, F_{es}, and F_{pt}) were trained using the Word2Vec [34, 128] implementation in gensim [121] with continuous bag-of-words (CBOW), negative sampling, 200 dimensions, and a window size of 10.

Geo-located Tweets can be used to train a variety of language resources; we implemented procedures to train three different popular word embedding approaches: brown clusters, and GloVe and Word2Vec embeddings. In our experiments we only apply Word2Vec embeddings downstream, although the three resources would be quite interchangeable due to their similar properties, and we expect the conclusions drawn in this chapter would carry over.

4.3 Application of localised word embeddings in an ensemble approach

We assess the utility of our derived localised embeddings with regards to user profiling, by examining their affect when included in a user profiling system; our approach combines two probabilistic classifiers trained on distinct feature sets in an ensemble to predict gender and language variant.

We build on our previous results from Chapter 3, and apply a logistic regression

(LR) classifier trained on TF-IDF transformed n-grams (Section 4.3.1) as a strong baseline/starting point for our approach. LR was chosen for this component for its probabilistic output, similarity to SVM in terms of predictive performance, and ability to handle highly dimensional bag-of-words features.

To apply our localised word embeddings, we first cluster them into *word embedding clusters* using K-means. Frequencies of each cluster across a user's Tweets are aggregated into a single feature vector and used to train a Gaussian process (GP) classifier (Section 4.3.2). A GP classifier was used in conjunction with these features, as we found it was able to achieve better predictive performance than LR on the lower dimension word embedding cluster frequency feature vectors, while still being a probabilistic classifier for use in our ensemble approach.

For each unseen document, probabilities from both classifiers are taken and averaged, and the highest average probability class is taken as the prediction. Models were trained using the implementations found in scikit-learn [119] unless stated otherwise. We made no meaningful attempt to tune the classification decision boundary value in these experiments.

4.3.1 Logistic regression classifier with TF-IDF n-grams

Word unigram and bigram features were extracted for each training document, following the processes described in Section 3.2.2. The text was tokenised using a Twitter-aware tokeniser [116]; no additional steps were taken to deal with the extra complexities of Arabic text. A list of stop words was not used while deriving n-gram features, instead tokens that appeared in more than 90% of the documents were removed, as this allows for the removal of very common (low information-content) n-grams across a language's variants while also removing stop words.

TF-IDF weighting was applied to down-weight n-grams found to be common across the documents and assign a higher weight to n-grams which are rarer and more informative.

A logistic regression classifier was trained for each language using the n-gram features. Logistic regression was chosen for use with the n-gram features because it has been shown to perform well on similar high-dimensional classification tasks, and produces probabilistic outputs [133] for use in our ensemble approach.

4.3.2 Gaussian process classifier with localised word embedding clusters

To use word embeddings in a document classification approach, steps must be taken to convert them into features, before training a machine learning algorithm. 'Classic style' machine learning algorithms (e.g. naive Bayes (NB), SVM, LR) typically expect to see a one dimension vector per training example (aggregated collection of Tweets in this case), whereas our embeddings are represented as an one dimensional vector *per token* (or a two dimension sequence of embeddings). In NLP tasks such as sentence similarity or short-sequence classification a simple average of the embeddings in a document is a suitable reduction, although this is infeasible in this task, as the high token volume yields vectors that are indistinguishable from noise.

Lampos et al. [134] present a *bag-of-word embedding cluster* approach to utilising word embeddings to classify aspects socio-economic status in Twitter users. They first cluster embeddings to derive a term-cluster mapping, and use these mappings to generate feature vectors for the users in their dataset in a fashion similar to bag-of-words. The feature vectors are used to train GP classifiers with good results. Furthermore, it is noted that the derived word embedding clusters are similar in nature to topic models, in that they identify semantically similar groups of words in documents, which we showed in Chapter 3 to be a useful user profiling system component. Owing to these desirable properties, and the similar domain of application, we chose to adapt the *bag-of-word embedding cluster* approach for our own *localised word embeddings*.

Generating clusters of localised word embeddings

Within our localised word embeddings, terms that are semantically similar are expected to be closer together in embedding space than those that aren't. Using some distance metric, such as Euclidean or cosine distance, it is possible to measure the similarity between our embeddings. For example, 'dog' and 'cat' would likely be close together as they are both pets; 'puppy' would also be close, but we would expect it to be closer to 'dog', as a puppy is an infant dog; an obviously different term such as 'television' would likely be measured as much further away. The ability to measure meaningful relationships between word embeddings makes

Token	Cluster
dog	1
cat	1
anteater	1
snail	1
⋮	
table	16
chair	16
desk	16
sink	16
⋮	
played	32
went	32
ran	32
watched	32
⋮	
laptop	100
television	100
xbox	100
oven	100

Table 4.3: Example token-cluster mappings derived through clustering English localised word embeddings with K-means, highlighting conceptually similar terms within each cluster.

them highly suitable for clustering.

We applied K-means clustering [135] to the word embeddings to derive a set of 100 clusters for each language, in which each word is assigned a cluster based on its nearest cluster centroid in the embedding space. Table 4.3 illustrates the resulting mapping between tokens and embedding clusters.

Under our current clustering scheme, each term was assumed to be equally as representative of its cluster as each other term; in practise though, certain terms were closer to the centroid in embedding space than others. A potential expansion to this work would be to investigate the effect of weighting terms based on their proximity to their closest centroid.

Computing cluster frequency feature vectors

To compute feature vectors for each user, we took all of their tokenized Tweets (with frequent terms removed), looked up the relevant cluster for each token, and counted the frequency of each cluster, yielding a one dimension vector of integers of length 100. This vector was then normalised to the range $[0, 1)$, by dividing each frequency by the total number of tokens counted.

This process yields feature vectors comparable to the raw probability output of topics models such as LDA. Unlike our LDA based feature extraction approach in Section 3.2.2, we do not transform the cluster frequency feature vectors into binary feature vectors.

Gaussian process classifier

The normalised cluster frequency feature vectors for each user were used to train Gaussian process classifiers with an radial basis function (RBF) kernel [100] for each language. Gaussian process classifiers were chosen as a good fit for our ensemble approach, due to their probabilistic output and reputation for good performance on similar classification tasks [134]. The RBF kernel was chosen due to its position as a good default in the literature [100]. We confirmed through ad hoc experimentation that the GP classifiers slightly outperformed logistic regression classifiers trained on the same data.

4.4 Results and discussion

We frame our evaluation as a classification problem. All models were trained on the *train* subset of the dataset presented in Section 4.1.1. Baseline results (Section 4.4.1) are evaluated on the *test* subset, and ensemble results (Section 4.4.2) are evaluated on the PAN 2017 evaluation platform against the second *held-out test set*. For comparability to other approaches presented at PAN 2017, we measure model performance in this study in terms of *accuracy*, defined as the number of correct predictions, divided by the total number of predictions made.

We do not present results for the Arabic language due to the small number of

Tweets acquired and the data quality issues regarding non-standard use the language in both the PAN 2017 Author Profiling (Section 4.1.1) and geographically filtered Tweet (Section 4.1.2) datasets.

4.4.1 Baselines

In Chapter 3 we identified that support SVM classifiers trained on TF-IDF n-grams is an acceptable, widely used, approach for user profiling tasks. As a baseline for this evaluation, we trained an SVM classifier with a linear kernel on the TF-IDF n-gram feature vectors derived in Section 4.3.1, for each addressed attribute (gender and language variant). Table 4.4 shows the accuracy scores achieved by this baseline, as well as those achieved by a majority class baseline for comparison. The SVM baseline performed as well as expected, and is clearly able to distinguish between classes for both gender and language variant, beating the majority class baseline in all cases, especially for language variant and joint prediction. These results show that our SVM approach is good baseline to compare against our geographically derived resources.

Baseline	**Target**	**Spanish**	**English**	**Portuguese**
Majority class	Gender	0.5	0.5	0.5
	Language variant	0.1429	0.1667	0.5
	Joint	0.0714	0.0833	0.25
SVM	Gender	0.7361	0.7896	0.8263
	Language variant	0.9532	0.8617	0.9800
	Joint	0.7007	0.6838	0.8113

Table 4.4: Baseline accuracy scores for gender and language variant prediction for each language derived from a SVM classifier trained on TF-IDF n-grams.

4.4.2 Ensemble

Table 4.5 shows the results of our ensemble approach applied to the *held-out* test set. For Spanish, English and Portuguese the results were attained by applying the ensemble of logistic regression and Gaussian process classifiers described in Section 4.3. We were unable to derive baseline SVM results on this dataset due

Target	Spanish	English	Portuguese
Gender	0.7939	0.7829	0.8388
Language variant	0.9368	0.8038	0.9763
Joint	0.7471	0.6254	0.8188

Table 4.5: Accuracy scores for gender and language variant prediction for each language as submitted for the PAN: Author Profiling task 2017.

to limitations imposed by the shared task organisers, although both datasets were collected using the same methodology, and as such, results should be comparable between the two.

As with our SVM baseline, we see that the our ensemble approach beats the majority class baseline in all cases. Our ensemble performs well gender prediction, beating our SVM baseline for both the Spanish and Portuguese datasets, and achieving similar accuracy score for English. For language variant prediction, our ensemble is beaten by the SVM baseline for all three languages, more-so for English than Spanish and Portuguese. For joint prediction our ensemble beats the SVM baseline for both Spanish and Portuguese. Poor performance for English is due to errors in language variant prediction propagating through to incorrect joint predictions. Of the three languages the ensemble was applied to, the best performance was observed for Portuguese and the worst for English.

Noting improved performance on the introduction of localised word embedding cluster features for the gender prediction task, we conclude that broad topics of interest appear to be effective for the prediction of user gender. This is in line with our observations in Chapter 3, where LDA topic models were able to improve predictive performance over word TF-IDF n-grams. In the case of language variant prediction, broad topics appear to hurt predictive performance, and individual terms that are more common in (or unique to) specific language variants are more discriminating.

The differences in performance between our SVM baseline and ensemble approach are potentially attributable to random differences between the two different datasets from which results were derived; as such we also compare our approach, and it's results, against those of other authors on the same dataset. In the rankings for the PAN Author Profiling shared task [127], our approach achieved 7th place out of 22 entries for joint prediction and 6th for gender, exceeding reported

baselines. We achieved poorer results for language variant prediction at 9th place, and did not exceed the strong baseline approach proposed by the shared task organisers.

Several participants in the 2107 PAN Author Profiling also utilised n-grams in their approaches. Of the demonstrated systems utilising n-grams [136, 137, 138, 139, 140, 141, 142, 143], our ensemble approach outperformed all of these systems in the gender prediction task, except for the approach in Martinc et al. [140], which incorporated a wide range of additional manually selected feature sets including parts-of-speech (POS) tags, emoji, sentiment information, stylistic features and language variety word lists. This again shows that our localised word embedding clusters are a useful addition for improving performance at gender prediction. For language variant, our ensemble approach also out-performed most other n-gram based approaches, despite appearing to perform poorly against our own baseline. Our ensemble approach was again beaten by the approach in Martinc et al. [140], as well as by Markov et al. [139], who employed a very thorough feature selection scheme, and by Schaetti [143], who utilised deep learning methods.

A small number of approaches also experimented with word embedding models [144, 145, 146]; each approach trained a language specific word embedding model on the Tweets provided in the shared task datasets. No other approach tailored their embeddings to ensure representation for specific language variants using large-scale external corpora. For gender prediction, our ensemble approach outperformed each of the other embedding based approaches, including Kodiyan et al. [145], Sierra et al. [146], who both utilised deep learning methods, and Akhtyamova et al. [144], who used averaged word embeddings for each user to train logistic regression classifiers. For language variant prediction our ensemble out-performed Akhtyamova et al. [144], Kodiyan et al. [145], achieved similar results to Sierra et al. [146].

Deep learning models are often assumed to be the best performing, and therefore default choice, machine learning algorithm in modern contexts. A number of the submitted approaches, such as Kodiyan et al. [145] and Sierra et al. [146], utilised deep learning methods in their approaches. In fact, our ensemble approach, which utilises a purposefully designed set of features used to train a "classic" machine learning algorithm, can often match or exceed the predictive performance of the more advanced deep learning based approaches, which yield hard to explain models, and require much greater computational resource for both training and

inference.

4.5 Conclusion

In this chapter we presented a novel method for acquiring large language corpora tailored to specific regions by leveraging geo-located social media posts; we used these corpora in a down-stream task by training word embeddings on them, and demonstrated their utility in a user profiling setting. This provides us with certainty that there is at least *some* benefit to including geographically derived resources in user profiling. In future chapters, we will demonstrate that geographically derived resources have further uses, and use them to replace traditionally derived user profiling resources altogether.

We also expanded on our evaluation of state of the art tools for user profiling on established datasets; in later chapters, we will build on this strong foundation in our experiments on datasets derived through our own methods.

Chapter 5

Estimating user home location

In Section 1.3 we laid out our proposed approach for generating user profiling datasets by combining social media profiles tagged with a 'home location' and publicly available demographic datasets. This approach requires a good method for assigning home location to profiles at the fine-grained *hyperlocal* coordinate or small region level. In addition, we require the ability to assign some measure of 'uncertainty' to our estimates, such that users with sparse or highly spread activity can be excluded from analysis.

For a home location identification method to be suitable for hyperlocal user geolocation, it needs to perform accurate coordinate-level predictions or pick potentially small regions from a list of candidates. In Section 2.4 we reviewed the state-of-the-art approaches used to assign 'home location' to social media profiles in the literature, and identified four state-of-the-art methods to evaluate:

- *First Tweet* involves simply taking the the coordinates of a user's first Tweet as their home location;
- *Location field* utilises a gazetteer or geo-coding service to convert the user-declared location field into a home location;
- *Grid based* partitions the world into an arbitrary grid and counts the number of Tweets geo-located within each cell, taking the most populous as the home location; and
- *Geometric median* simply takes the geometric median of a user's collection of geo-located Tweet coordinates, which is taken to be their home location.

While the *first Tweet* method supports hyperlocal, coordinate level geo-location, it is highly error prone. Information in the *location field* can only reasonably be used to geo-locate at town/city level granularity or higher, which makes these methods less suited to hyperlocal geo-location. The *grid based* approach in Han et al. [92, 94] reliably locates profiles, but at a granularity far above what could be called hyperlocal, using grids larger than many towns (outside of the US), and as such will not be assessed in this chapter. The *geometric median* approach of Jurgens [85], Compton et al. [86], Jurgens et al. [87] is likely to be the most reliable state-of-the-art method for hyperlocal geo-location, producing single coordinate level predictions with outlier resistance and a measure of judgement certainty, but deriving home location from the median of *all* of a user's Tweets is likely to lead to incorrect judgements when a user is often active in more than one area.

None of the identified state-of-the-art approaches appear to quite satisfy our requirement of a robust and accurate method for identifying hyperlocal home location, so in this chapter we propose two novel methods that overcome some of the limitations of state-of-the-art methods for deriving ground truth user location (Section 2.4). Our methods acknowledge that social media users are often active at multiple locations, and propose that their most active location is likely to be their primary location (see Section 5.1):

1. *Majority voting* measures a user's activity in real world regions as defined by public boundary datasets. Four such resources are used in this chapter (see Section 5.2.1); and
2. *Clustering* works at the hyperlocal level by applying a clustering algorithm to each user's geo-located Tweets, revealing centres of activity. The most populous cluster is taken as the users *home*, and its geometric median is taken as the user's *home coordinate*.

Previous methods for home location identification have not been evaluated at the hyperlocal level, limiting understanding of their accuracy for this fine-grained task. A new *gold-standard* dataset containing hyperlocal geographic information was created (Section 5.2). This dataset was created by identifying a set of key phrases indicating which Tweets were sent from the user's home. These phrases were used to identify the home location of 1,042 Twitter users. Each Tweet containing a home-indicative phrase was checked to ensure it referred to 'home' in the correct context, being removed from the dataset if not. The 'true' home location was

calculated for each profile by taking the spatial average of its home-indicative Tweets.

To evaluate the accuracy of state-of-the-art and newly proposed methods for deriving user location, we implemented and applied them to all profiles in the gold-standard dataset (from Section 5.2), comparing the predicted and 'true' locations (Section 5.4). Two metrics are presented; error distance (in miles) and exact match accuracy (at four granularities of UK administrative region). The results demonstrate that our clustering based approach outperforms state-of-the-art methods on the hyperlocal geo-location task, at both coordinate and region-level granularity, and satisfies our requirement of a good method for assigning home location to Twitter profiles.

This chapter is adapted from work presented at Hypertext 2017 [129]. Section 5.1 introduces the two novel methods that we will be evaluating. The process for acquiring ground-truth (i.e. with a true home location) user profiles is described in Section 5.2, and how this will be used to assess the current and novel methods is described in Section 5.3. Results and conclusions of this investigation are given in Sections 5.4 and 5.5.

5.1 Acquiring fine grained home location estimates

The two improved methods proposed in this chapter take into account all of a user's geo-located Tweets (e.g. Figure 5.1), in order to identify the user's home location at coordinate level. Both approaches assume that each social media user commonly posts from a limited number of locations, with the highest frequency Tweeting location assumed to be the user's home (in line with human home-return patterns [147] and other user geo-location approaches [92, 94, 148]). The first method is based on majority voting across possible regions and is presented in Section 5.1.1, whereas the second is based on clustering of geo-located Tweets (see Section 5.1.2).

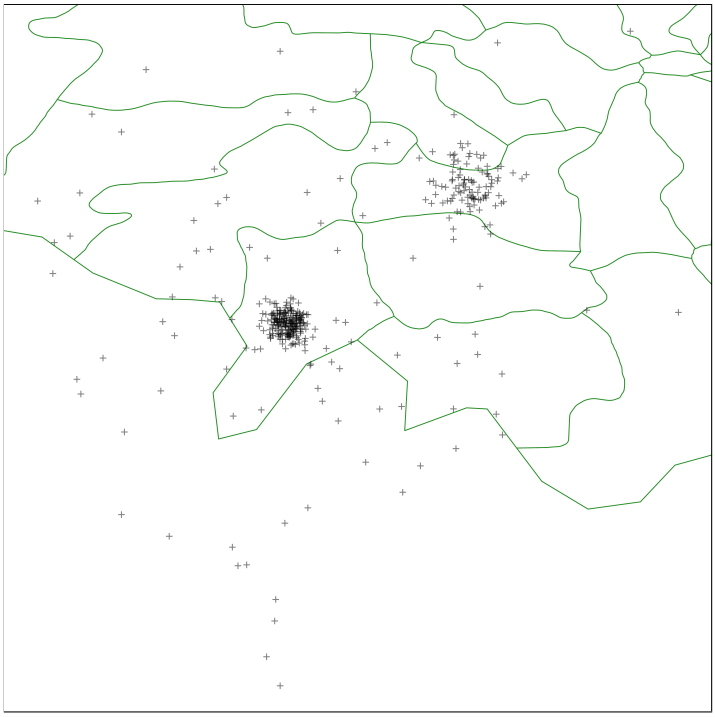

Figure 5.1: Example collection of unprocessed Tweet coordinates.

5.1.1 Majority voting

In a manner similar to how datasets are generated in a number of user geo-location studies, our method for generating user profiling datasets hinges on associating social media users with their 'home location' in the form of public boundary data, such as states or other administrative subdivisions. We propose a method that, similarly to grid-based approach presented in Han et al. [92, 94], operates in discrete space and counts points. Unlike Han et al. [92, 94] which operates on an fixed sized grid, our method can be applied to any boundary data, at any scale.

By projecting all the coordinates of a user's Tweets onto the same surface as the boundary data (county borders, for example, in the US), the boundary containing the majority of a user's Tweets can be identified. Immediately putting locations into the 'boundary space' is in contrast to most of the methods described in this Section 2.4 that first determine home location in coordinate form.

A naive implementation of this somewhat brute force method is likely to be quite slow, due to reliance on many point-in-polygon tests which can be quite computationally intensive depending on the dataset; this can be though by sensible use of methods such as spatial indexing.

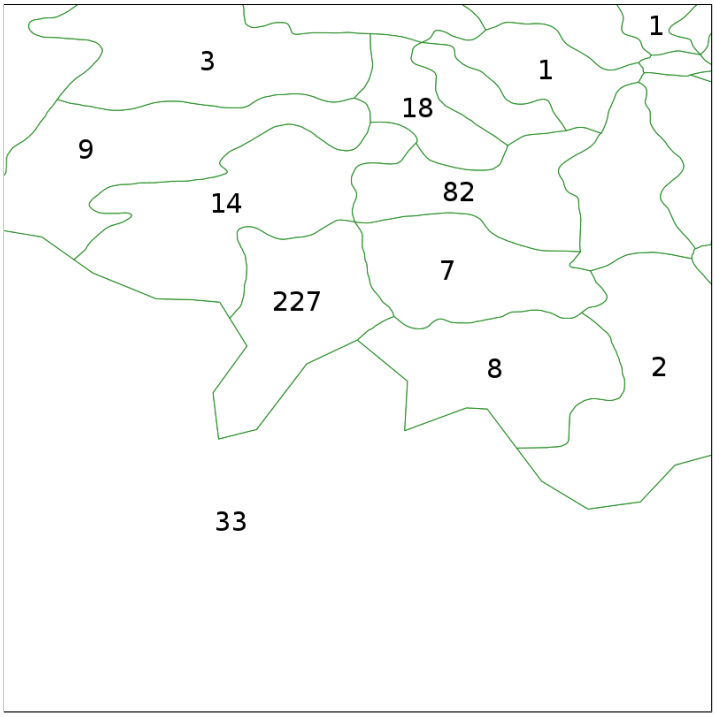

Figure 5.2: Example application of the majority voting method to the data in Figure 5.1.

To pick the home location, the boundary containing the most Tweets is chosen. The number of Tweets in other boundaries to this home estimate are known, and so some measure of uncertainty can be calculated, such as the proportion of Tweets that reside within the home region. This allows for the elimination/downgrading of profiles with low region certainty. As such, majority voting does solve the requirements of needing to pick a region/coordinate as the home location and quantify the uncertainty around such an estimate.

Figure 5.2 shows an example of the majority voting method applied to the unprocessed Tweet coordinates displayed in Figure 5.1. The region with the most Tweets contains 227 examples, and is selected as the user's home.

5.1.2 Clustering approach

Since social media users are often active from more than one location, an approach that first clusters Tweets into regions of activity prior to picking one as the home location is more promising over simply picking the geometric median of all user posts, as in Jurgens [85], Compton et al. [86], Jurgens et al. [87]. In particular, for

users with more than one commonly posted-from location (such as a home and work location), the geometric median can sometimes pick a point in between these two locations, while a clustering approach can overcome this limitation.

Our method first clusters the coordinates of each user's geo-located posts using an algorithm such as K-means, DBSCAN, or Gaussian mixture models (GMMs) (see details below). This reveals a collection of candidate clusters for the home location. The cluster with the highest number of posts is identified and taken as the 'home cluster' and the home location coordinate is taken to be the geometric median (Equation 2.1) from all locations within the home cluster.

Figure 5.3 shows an example of the clustering approach on the data in Figure 5.1; two clusters are identified with cluster membership of each point presented as a square or circle (home coordinate in red). In this case the most populous cluster is the one represented by squares and the red square is taken as the user's home coordinate.

The ability of a clustering method to estimate location can be quantified by assessing the geographic size of the estimated 'home cluster'; a geographically small cluster that is dense in points will likely lead to a better estimate than a geographically large cluster with a sparse coverage of points. A measure of the size of a cluster could therefore be to take the average distance from the estimated home location to each point in the home cluster. Explicitly these distances are determined by the Haversine [149] distance, and the mean of these is taken.

An extension of this method not evaluated here would be to use it in conjunction with the majority voting method. In an application where the goal is to assign a given user to a region, the number of point-in-polygon tests could be greatly reduced compared to majority voting by only checking the points in the user's home cluster.

Clustering algorithms

Since our clustering geo-location method relies on identifying location clusters, we experimented with partitional clustering algorithms [150]. In particular, three well established clustering algorithims were experimented with (K-means, DBSCAN and GMMs), described in the following.

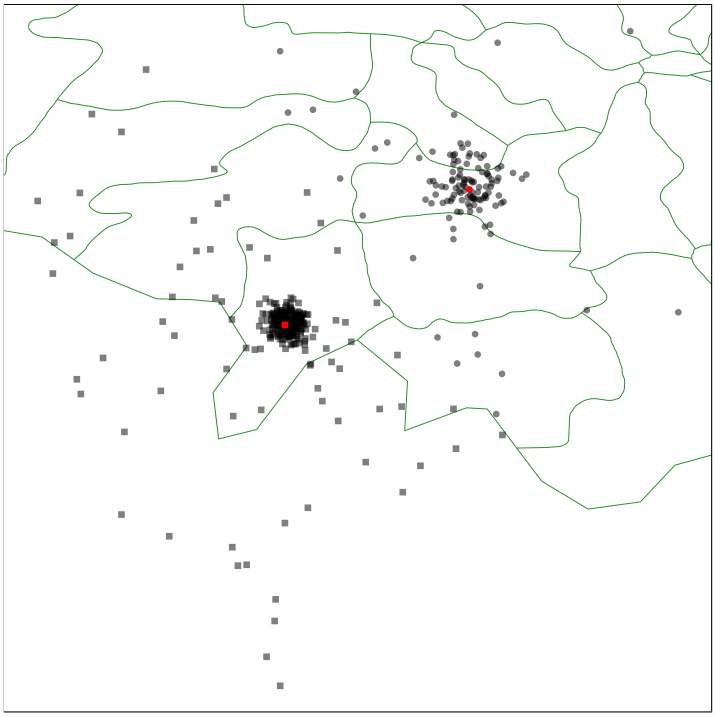

Figure 5.3: Example clustering of the Tweet coordinates in Figure 5.1.

K-means K-means [135] attempts to partition a set of observations

$$X = \{x_1, x_2, \ldots, x_N\}$$

of length N into $k \leq N$ clusters $C = \{c_1, c_2, \ldots, c_k\}$ by minimising a criterion called the 'within-cluster sum of squares' (WCSS, also known as inertia). More formally, K-means aims to find

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \text{mean}(C_i)||^2. \tag{5.1}$$

The K-means algorithm has three steps, first the k clusters have their centroids initialised (for example by taking k points at random from the dataset). Then the algorithm loops between assigning each point to its nearest centroid, and calculating new centroids based on the changes in the previous step. This continues until the difference between the new and previous centroids falls bellow a threshold (the centroids are stationary). K-means is sensitive to its initialised centroids, thus the algorithm is usually run multiple times with different initialisations to help ensure good clusters.

DBSCAN Density-based spatial clustering of applications with noise (DBSCAN)

[151] is a density-based clustering algorithm that forms clusters based on the presence of high density regions (large numbers of neighbouring points). If a point lies in a low density region (i.e. no or few nearby neighbours), the point is marked as an outlier and not included in any cluster. Unlike K-means in which clusters are always convex, DBSCAN is able to produce clusters of any shape, including rings around other clusters.

GMM A Gaussian mixture model (GMM) is a probabilistic model in which all points in the data are generated from a finite mixture of Gaussian distributions with unknown parameters [152]. Inference on the parameters of a GMM is typically approached using the expectation maximisation (EM) algorithm [153]. Having pre-specified the number of Gaussian components and initialising parameter values, the EM algorithm consists of two steps; the 'expectation' step, in which the likelihood is evaluated at the current parameter estimates, and the 'maximization' step, in which the expected likelihood from the previous step is maximised.

5.2 Data

We were unable to compare different approaches to assigning home location on an existing dataset for multiple reasons. The permanence of Twitter data is a difficult issue for studies centred on the platform; Twitter does not allow Tweets to be shared in full, instead only Tweet IDs can be provided. This leads to a problem in sharing datasets as large numbers of Tweets are time-consuming to acquire, and many Tweets and profiles are deleted over time. None of the datasets available provided high certainty gold-standard hyperlocal home location labels. Eisenstein et al. [82], Mahmud et al. [77, 78], Roller et al. [84] all used the earliest Tweet in their sample for each user to label their home location and made little effort to validate these labels. Alex et al. [91], Rout et al. [64] used the profile's location field to assign gold-standard home location at the city level, which is unsuitable for assessing hyperlocal methods. Han et al. [94] assigned user's home location taking all of a user's geo-located Tweets into account, but uses very coarse labels, not comparable to hyperlocal methods. As a result, we constructed our own gold-standard dataset of profiles labelled with home location.

The Twitter public streaming API[1] was used to identify geo-located Tweets within the UK, from November 2014 to July 2015. Both the Tweets and the users that created them were recorded. The Twitter REST API was then used to collect retrospectively a sample of 135,000 user's Tweets (up to 3,200 per user) and any public information available on their profile. A 'gold-standard' set of profiles was created from this information by assigning 'true home location' based on implicit mentions of 'home' in geo-located Tweets (discussed in Section 5.2.1).

UK users were chosen because the population- and city-density is much higher than in the USA, which makes the geo-location task much more challenging. Two major and distinct cities, Liverpool and Manchester, are close enough together that they would be considered one place in the scheme proposed in Han et al. [92, 94], for example. The UK features distinct dialects and customs between it's constituent countries, regions, and cities (and indeed towns and villages) [154], so from the context of our method for assigning local demographics to users based on user location for user profiling, the ability to accurately distinguish locations in the UK (and similar contexts worldwide) is essential.

5.2.1 Gold-standard home location dataset

To build the gold-standard dataset, we proposed that if a user emits a phrase referring to being 'at home' from the same location multiple times, it is their home location. A small number of phrases were collated by performing an analysis of Tweets containing the word 'home'. For each Tweet 3, 4 and 5-grams were calculated, and those not containing the word 'home' were discarded. The most frequent n-grams were manually inspected, and ones that seemed indicative of being 'at home' chosen. Four phrases were selected: "just got home", "glad to be home", "finally home after" and "home after a long".

On manual inspection it was noted that some indicative Tweets did not use 'home' to refer to the users' primary residence. Examples include; university students using 'home' to refer to their family home, and holiday-makers calling their hotel room 'home'. As such, the Tweets were manually checked to ensure that the text referenced residential home.

Profiles containing geo-located Tweets with one or more indicative phrase were

[1] `https://dev.twitter.com/streaming/overview`

selected from the whole collection of around 135,000 profiles, leaving 7,348 (5.44%) with at least one example and 1,498 (1.12%) with two or more. To improve certainty only those profiles with two or more 'at home' Tweets were considered. If the coordinates of their 'at home' Tweets were within a short distance of each other (taken to be 0.5 miles) we took the spatial mean of these as the 'true home location', resulting in 1,048 users. After selection each profile was represented as an anonymous table row with three fields: gold standard home location, geo-coded location field text (see Section 5.3.1), and geo-located Tweet coordinates.

It is worth noting that this approach in itself could be seen as a distinct method for assigning home location to social media users, but due to the small number of Tweets which contain phrases of this nature, would be infeasible for any large-scale application, such as our method for generating user profiling datasets by linking users to their local demographics.

Addition of boundary data

Government bodies maintain a number of geographic datasets that split countries, or even groups of countries, into subdivisions for statistical or administrative purposes. These boundaries are used in addition to, or in conjunction with, more traditional boundaries such as states, cities, electoral regions or postal codes. Many examples exist such as the European Union's 'Nomenclature of Territorial Units for Statistics' (NUTS) [2]. This is a three layer hierarchy of subdivisions that covers all European member states, which is used to tabulate EU population statistics and inform decisions on the distribution of EU funds to avoid regional disparities in wealth, income and opportunities. The US Census Bureau [3] maintains standard geographic boundaries for the entire US such as states and counties, as well as various statistical groupings/subdivisions such as 'metropolitan/micropolitan statistical areas' and 'combined statistical areas'.

Geographic data in the UK is maintained by the Office for National Statistics (ONS) [4], who provide boundary data for a wide range of geographic areas such as counties, electoral wards and civil parishes. They also provide purely statistical subdivisions, derived from distributions of population in the latest census, called

[2]NUTS: `http://ec.europa.eu/eurostat/web/nuts/overview/`
[3]US Census Bureau: `http://www.census.gov/`
[4]ONS: `http://www.ons.gov.uk/`

Geography	Total number	Min pop.	Max pop.	Average pop.
OA	181408	100	625	309
LSOA	34753	1000	3000	1500
MSOA	7201	5000	15000	7200

Table 5.1: Characteristics of the NSG shown in terms of maximum, minimum and average population.

Output Areas (OAs). OAs form the smallest building block in a hierarchy of subdivisions known as the Neighbourhood Statistics Geography (NSG), which consists of OAs, Lower Layer Super Output Areas (LSOAs), and Middle Layer Super Output Areas (MSOAs). These units are used to present census data in a consistent fashion, as well as other statistics such as the 'Indices of Deprivation', a measure of poverty in small areas across the UK.

As described earlier in this section, we have compiled a gold-standard collection of UK Twitter profiles labelled with home location in the coordinate form. Many analyses, however, do not deal with home location in the coordinate form directly, instead converting it to some region of lower granularity such as state, postal code or city. As such we additionally label our gold-standard with 'true home boundary' data, in particular we label each profile covered by the NSG at the OA, LSOA and MSOA level as well labelling each profile with their Local Authority District (LAD), a larger region type governed by a council.

In Table 5.1 we present the characteristics of the NSG which consists the OA, LSOA and MSOA region types and covers England and Wales. LSOAs and MSOAs are constructed from groupings of the previous level and both are always larger than the maximum size of the previous level; to illustrate this, Figure 5.4 shows a 0.55 square mile MSOA, one if its component LSOAs (0.13 square miles) and an OA (0.04 square miles). The groupings are consistent in terms of population but do vary in geographic size, meaning there are more divisions in areas of high population density such as cities. LADs are excluded from the table as they are not consistent in population or geographic size. 415 LADs were used to label the gold-standard, covering all of the UK, including Northern Ireland and Scotland.

Figure 5.4: Example MSOA (whole shape) with nested LSOA (shaded with horizontal line) and OA (shaded with vertical line).

5.3 Experiments

The three state-of-the-art methods for determining a user's home location: first Tweet, location field, and geometric median (described in Section 2.4), and the two new methods proposed here: majority vote and clustering (described in Section 5.1), were implemented and applied to the gold-standard dataset (Section 5.2).

5.3.1 Implementation

In order to apply the first Tweet method, the first recorded geo-located Tweet in our sample for each user is used as the home location coordinate, as with [77, 78, 82, 84].

The location field method was implemented using a gazetteer of unique place names extracted from OpenStreetMap[5] to identify profiles with a single reference to a city or town within the UK in their location field. Employing a gazetteer for this task aided in discarding those profiles whose location field were either

[5]OpenStreetMap: `https://www.openstreetmap.org/`

clearly fictitious (e.g. '221B Baker Street', '90210' and '42 Wallaby Way, Sydney'), biographical information that did not actually reference locations (e.g. 'she/they', strings of emoji), or were unclear (e.g. 'notts ladd'). Varying levels of granularity were present in the declared location fields; ranging from street level to country. A small number of actual coordinates were also included in addition to the named location profiles. As with Hecht et al. [88], city level profiles were most prevalent followed closely by town, and as such we chose to restrict our analysis to profiles who declared a location only at these two levels, similar to Rout et al. [64], rather than higher and lower granularities. A limited number ($\sim$30%) of profiles in the gold-standard had a location field resolvable to a uniquely named town or city.

The geometric median method of Jurgens [85], Compton et al. [86], Jurgens et al. [87] was applied, although the simpler Haversine distance was used instead of the more computationally intensive Vincenty for computational efficiency and consistency across approaches (although we believe this will cause no discernible change in computed distance). Median absolute deviation of the posts was not limited in order to assess this method's performance on all users; in addition we aim to limit or remove the necessity of this step entirely with our novel methods that take into account multiple locations of activity.

Majority voting was applied at the four levels of granularity introduced in Section 5.2.1. Three different clustering algorithms (K-means, DBSCAN and GMM) were experimented with, using the implementations available in scikit learn [119], using default parameters, which were found to produce sensible outputs through ad hoc experimentation. For K-means and GMM a `k` and `n_components` of 10 were used respectively.

5.3.2 Evaluation metrics

We employ two metrics that have previously been used in the user geo-location literature to evaluate the approaches.

Error distance is applied when continuous (coordinate level) predictions are made, and is the distance between the 'true home location' and the estimated home location. The mean error distance (i.e. the sum of the error distance for each profile divided by the number of profiles) is computed for each approach, and reported alongside the quantiles of the full set of error distances.

Exact match accuracy refers to the proportion of examples that are correctly geolocated within the correct region (e.g. the correct city, state or ZIP code), referred to in Section 5.2.1 as the 'true home boundary'.

5.4 Results

The results of two experiments on the state-of-the-art and novel methods are presented in the following. The first experiment was carried out in continuous-space, aiming to assign each user a coordinate home location, and compare this to the true home coordinate given by the gold-standard dataset. The ability of each method to estimate this location was evaluated via the error distance (defined in Section 5.3.2), and is presented in Section 5.4.1. The second experiment was carried out in discrete-space, aiming to assign each user a home boundary (for four boundary datasets), making a comparison with the true home boundary defined within the gold-standard. Accuracy of each method's evaluation of this was evaluated using exact match accuracy (defined in Section 5.3.2), and is presented in Section 5.4.2.

5.4.1 Error distance

The distance in miles between the 'true home location' and the estimated home location for each user was calculated (referred to in the following as the 'error distance'). The set of error distances for the whole gold-standard dataset (and around 30% in the location field case) were used to quantify each method's accuracy, with summary statistics listed in Table 5.2, and shown graphically in Figure 5.5. Note that the majority vote method is not included here because it does not involve estimating a home coordinate.

First Tweet produces an estimated home location to within 0.1849 miles of the 'true' home location in over half the users in our dataset (i.e. the 50th percentile (median) is 0.1849); this is in line with our hypothesis that Tweets have a high probability of being produced at a user's home. However, a quarter of the cases in our dataset had home location errors of over 9.8909 miles, which is high in comparison to other approaches. Figure 5.5 illustrates that up to the 25th percentile of errors, the method is almost as accurate as the more advanced methods

assessed, highlighting the nature of the errors made by the first Tweet method: it is able to make accurate predictions when the user's first Tweet *happens* to be at their home, but this is not the case for a significant proportion of the users in our dataset and leads to an estimate that is essentially noise (hence the large error at the 75th and 95th percentiles).

As location field is applied at the city/town level, error distances up to ∼5 miles are to be expected, however, a quarter of the errors calculated exceed this. On manual inspection of the profiles, it became apparent that many were poorly labelled due to unreliable information; many seemed to be residents of satellite towns or villages and declared their location as the larger nearby town or city. Examples of this include: Mansfield to Nottingham (15 miles), Cheadle to Manchester (7 miles) and many London commuter towns. The accuracy of location field based geo-location is therefore dependent on the proximity of the user to the centre of a town/city, which makes such methods ill-suited to hyperlocal geo-location. Figure 5.5 clearly shows the unsuitability of location field in comparison the other methods, as only the 5th percentile of errors could be comfortably shown on the same graph, and this value is itself higher than the 95% error percentile of K-means.

As expected, the geometric median approach of Jurgens [85], Compton et al. [86], Jurgens et al. [87] produced the best results of all state-of-the-art methods, due to the method's outlier resistant nature and usage of all geo-located Tweets. Observing Figure 5.5, we can see similar performance between the geometric median approach and our clustering approaches up to the 50th percentile, but with greater errors at the tails (the 75th and 95th percentiles). This is to be expected, because we are calculating a geometric median in our clustering approaches, but to the home cluster only. Therefore, when a user either Tweets from mainly one location, or symmetrically around it, the geometric median is essentially equivalent to our clustering approaches. The gain here is therefore in reduction of the error tail, which is due to the geometric median's inability to handle scenarios where a user's Tweets are spread across multiple locations.

Our three novel clustering methods achieved similar results to each other, being almost indistinguishable up to the 75th percentile of errors (visible in Figure 5.5). K-means has the largest mean error (Table 5.2)) of these clustering methods when applied to the full gold-standard, but is otherwise the best performing up to the 95th percentile of errors. The difference in mean and 95th percentile errors between the three algorithms is due to each approach making a small number of incorrect

Method	Percentile of the error distance					Mean
	5	25	50	75	95	
First Tweet	0.0024	0.0079	0.1849	9.8909	568.8185	165.3717
Location field	0.3871	1.0463	2.5591	5.7749	30.2101	8.7826
Geometric median	0.0025	0.0057	0.009	0.0252	1.7246	1.2769
Clustering: K-means	0.0027	**0.0051**	**0.0081**	**0.0158**	**0.3079**	10.8883
Clustering: DBSCAN	0.0025	0.0054	0.0082	0.0166	0.5211	1.4055
Clustering: GMM	**0.0024**	0.0052	0.0083	0.0176	0.6168	**0.963**

Table 5.2: Quantiles of the set of error distances (in miles) for the 'gold-standard' dataset, calculated using home locations estimated via a number of methods (best results in bold). For example under the first Tweet method, 5% of the profiles analyzed had estimated home locations to *within* 0.0024 miles of the true location. Half of the profiles analyzed were correctly estimated to *within* 0.18 miles. Finally, 5% of the profiles analyzed had estimated home locations that were incorrect by *over* 568 miles.

predictions each, with especially large errors for these outlier users in the case of K-means. Each of the clustering algorithms presented have tuning parameters that we did not attempt to optimise here, therefore we can not reasonably say that any one is 'best' given the small discrepancies between their results.

Due to the limited number of profiles with location field, we also include results for only those profiles with a valid location field so as to not penalise the method unfairly (presented in Table 5.3). Results presented in the two tables are very similar across the board, and any differences are likely attributable to the smaller number of profiles with location field information. Our clustering approaches are again shown to perform best again on this limited subset, with K-means again performing best of the three algorithms. Both location field and first Tweet exhibit the same large errors noted earlier. The lack of consistently available location field information is another reason for its unsuitability for hyperlocal geo-location.

Sources of error

Each of the clustering methods presented performed well, however a small number of the error distances were above one mile; we deem an error distance of one mile as 'high' in the context of hyperlocal home location identification. In Section 5.1.2 we highlighted that a large 'home cluster size', calculated as the mean Haversine

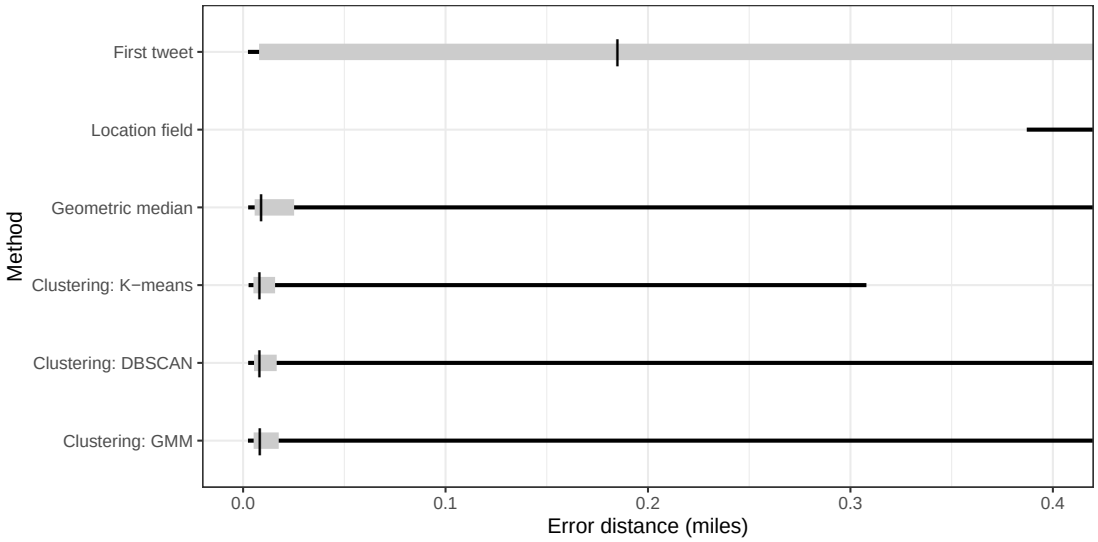

Figure 5.5: Quantiles of the error distances (in miles) for the 'gold-standard' dataset, calculated using home locations estimated via a number of methods. Lines show the central 90% range of errors, bars show the central 50% range and the vertical line shows the median (50% quantile). Note that this figure has been truncated at an error of 0.4 miles.

Method	**Percentile of the error distance**					**Mean**
	5	**25**	**50**	**75**	**95**	
First Tweet	0.0023	0.0071	0.1179	7.8129	167.2743	100.780
Location field	0.3871	1.0463	2.5591	5.7749	30.2101	8.7826
Geometric median	0.0025	0.0055	0.0089	0.0224	1.0111	1.4401
Clustering: K-means	0.0029	**0.0047**	**0.0076**	**0.0139**	**0.1585**	**0.4851**
Clustering: DBSCAN	0.0027	0.0055	0.0078	0.0149	0.2438	1.4873
Clustering: GMM	**0.0025**	0.0050	0.0080	0.0142	0.2894	1.2708

Table 5.3: Quantiles of the set of error distances (in miles) for the profiles with a valid location field from the 'gold-standard' dataset, calculated using home locations estimated via a number of methods (best results in bold).

distance to the home coordinate from each point in the home cluster, is a potential indicator of poor home location estimates.

We investigated the effect of limiting cluster size by discarding those users where their 'home cluster size' was above one mile. As noted earlier, all clustering algorithms performed similarly, to investigate limiting home cluster size we arbitrarily chose the GMM implementation. Of all (1048) error distances for GMMs, 46 (4.38%) were above 1 mile and 13 (1.24%) were above 10 miles. Upon discarding profiles with home cluster size above 1 mile, 913 error distances remained, of these 16 (1.71%) had error distances above 1 mile, and 8 (0.876%) above 10 miles. Applying this limit improved the performance of the method, thus we can conclude that cluster size is a useful indicator of the quality of home location estimates.

Limiting cluster size did not reduce all errors to zero, and therefore additional sources of error are present in the method. This is likely due to the somewhat naive assumption that the area a user posts most commonly from is their home in all cases. It is likely that some users are more active at a place other than their home, such as those who use Twitter for professional purposes or are also active while at a place of leisure or study. In future work this weakness could potentially be overcome by adding a labelling step for the identified location clusters, which would distinguish those users who post most frequently from places other than home.

5.4.2 Exact match accuracy

As stated in Section 5.2.1, the 'true' home location was used to determine four granularities (OA, LSOA, MSOA, and LAD) of 'true' home boundary, for each user in the gold-standard. The home location estimated by each of the coordinate methods is similarly converted to a boundary (the majority vote method already outputs boundaries), and the proportion of users from the dataset whose estimated home boundary was the same as the 'true' boundary was calculated. Note that the number of user profiles that could be included here is lower than the total number of users in the dataset because the boundary data does not cover the whole of the UK (1012 for LAD and 953 for the other granularities). These proportions of correct boundary classification are listed in Table 5.4, and presented graphically in Figure 5.6.

Method	Proportion correct by boundary type			
	OA	**LSOA**	**MSOA**	**LAD**
First Tweet	0.4260	0.4732	0.5273	0.6749
Location field	0.0094	0.0219	0.1003	0.6742
Geometric median	0.7534	0.8311	0.8939	0.9625
Clustering: K-means	0.7985	0.8783	0.9328	**0.9783**
Clustering: DBSCAN	0.7912	0.8688	0.9296	0.9733
Clustering: GMM	0.7954	0.8741	0.9275	0.9733
Majority vote	**0.8279**	**0.8919**	**0.9443**	0.9763

Table 5.4: Proportion of users whose estimated home boundary was equal to the true home boundary, displayed by method and boundary granularity (most accurate results in bold).

It again becomes clear that the first Tweet and location field methods are particularly inaccurate; first Tweet is only able to correctly classify a little over half the profiles at the lower two granularities (MSOA and LAD) and location field only at the lowest (LAD). The geometric median [85, 87], majority voting and clustering methods all perform well, classifying over three quarters of the profiles correctly at the highest granularity, and over 95% at the lowest granularity. As with error distance, geometric median is outperformed by clustering methods at region discrimination. Take GMM as an example, it correctly classifies an additional 4.2% of profiles at the highest granularity and 1.2% at the lowest.

Majority voting beats the clustering methods at the OA, LSOA and MSOA granularities by a small margin ($< 3\%$ at the highest granularity), but is narrowly beaten by K-means at the lowest granularity (LAD).

5.5 Discussion

The aim of this chapter was to explore and evaluate potential methods for assigning 'home location' to social media profiles, for future use in our user profiling dataset generation approach (described in Section 1.3). An ideal solution would be able to make coordinate level (hyperlocal) or small-region predictions, and allow for some method of quantifying 'uncertainty'.

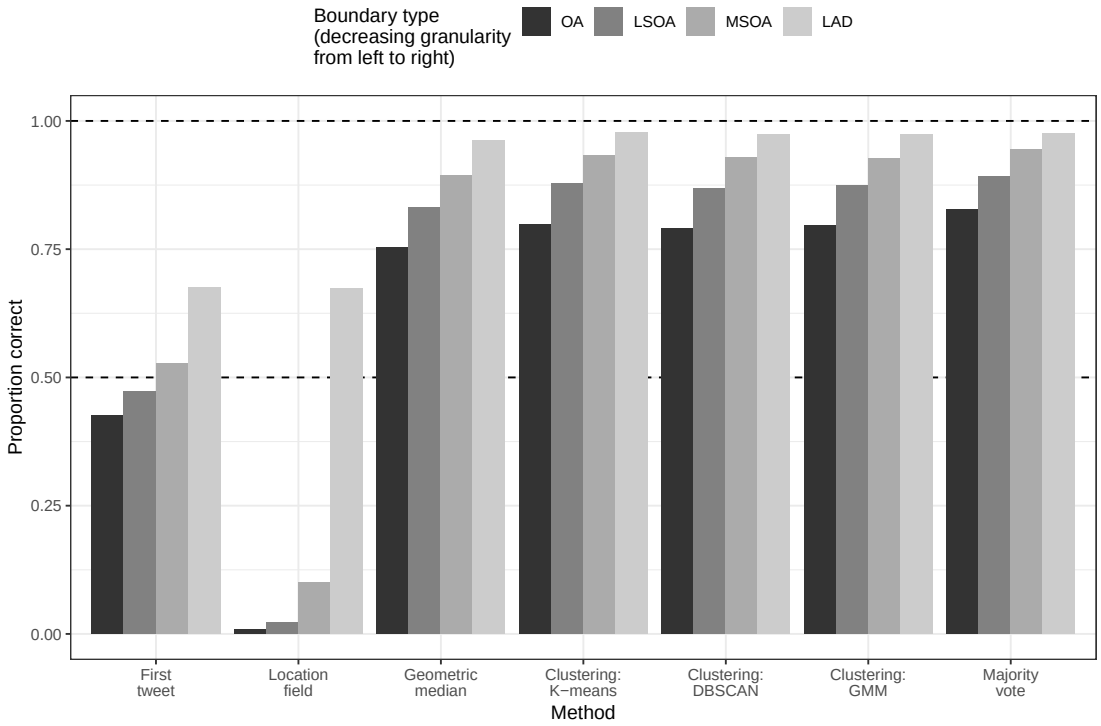

Figure 5.6: Proportion of users whose estimated home boundary was equal to the true home boundary, displayed method and boundary granularity.

Amongst the state-of-the-art methods evaluated on the new UK dataset, we found that the first Tweet method is the most prone to large errors, even though around 75% of user profiles were still correctly located to within 10 miles.

Similar to previous findings [88], we demonstrated that the textual location field is unreliable for hyperlocal user geo-location, since many users tend to specify the nearest larger city as their location, instead of the more accurate smaller village or commuter town. Consequently, user geo-location based on the location field text is ill suited to hyperlocal or fine-grain region prediction, but performed well in 67% of cases on the larger LAD regions—the largest boundaries we experimented with here.

The geometric median performed the best of all state-of-the-art methods, predicting accurately at both hyperlocal and regional level at different granularity. Nevertheless, it suffers from the limitation that it does not take into account cases where users post often from multiple, distinct locations.

Our newly proposed clustering methods outperformed all state-of-the-art methods, with a particularly strong lead over the first Tweet and location field-based

approaches. Amongst the three clustering algorithms evaluated, all achieved the comparable results at both hyperlocal and region based user geo-location, although K-means narrowly beat the other two at both. Majority voting marginally outperforms the clustering approaches at region prediction, but suffers from the drawback that it must be used in conjunction with boundary data. While such data is readily available in the UK, this may not always be the case for other countries.

Our novel approaches rely on geo-located Tweets (similar to the first Tweet and the geometric median methods), which make up only 1.24% of all Tweets [23]. This limits their applicability to only a small fraction of all Twitter users, but nevertheless, our methods are useful for unsupervised generation of high accuracy home location information, which can then be used as training data for user profiling models. In addition, the geometric median approach discards profiles with overly spread coordinates, which our method accounts for and handles, increasing the amount of users that can be geo-located accurately.

We deem that both of our novel methods satisfy our requirement of a good method for home location allocation method, as both were shown to be accurate in Section 5.4. Both methods also allow for a measure of uncertainty, an additional requirement, that allows less certain home location judgements to be excluded if required. Despite both methods being a good fit, the clustering approach will be applied over the significantly slower majority voting approach going forward; in Chapters 6 and 7, we will apply the clustering approach to generate user profiling datasets.

Alongside this chapter's contribution to selecting the home location allocation method for our user profiling dataset generation approach, several contributions were relevant to the wider user location prediction field:

- A comparative evaluation of state-of-the-art user geo-location methods, with respect to identifying the user's home location at a hyperlocal (coordinate) level, as well as at regional level (based on four UK regional classifications at multiple granularities); and
- Two new methods for user geo-location, which were shown to outperform the state-of-the-art methods;
- A new gold-standard dataset for evaluation of Twitter user geo-location at hyperlocal granularity. UK users were chosen specifically since population-

and city-density is much higher than in the USA (where geo-location studies typically focus), which makes the geo-location task much more challenging. For instance, Stockport is a large town in the UK, only 7 miles from Manchester, which is less than the error distance of some of the methods reported in Table 5.2.

Our current approaches for home location allocation rest on the assumption that a user's most posted-from location is their home, and while this assumption appears to work for most users in practice, it is highly likely that some users post more commonly from an alternate location such as their place of work, study or the homes of friends/family (leading to incorrect location labels under the current scheme). To help overcome this drawback, our method could be expanded in future work to evaluate user's active locations beyond 'home'. Identification of additional 'location types' could be carried out by incorporating the textual cues and metadata in Tweets in addition to coordinate data. For the problem of differentiating 'home' and 'work' locations the proportion of Tweets at each location in and out of canonical work hours could be investigated (building on the work in Cho et al. [155]). Additionally the nature of the geography at the predicted locations could be assessed, for example a prediction in a known residential area may more likely be correct than a predominately industrial area.

Chapter 6

Predicting user Local Authority and Output Area classification

In this chapter, we combine all of the techniques surveyed and evaluated throughout Chapters 2-5, to generate two novel user profiling systems that predict the Output Area (OA) and Local Authority (LA) classification schemes, two demographic variables that have not previously been addressed in the user profiling literature. We generate training datasets for these user profiling systems using our novel method, which links social media users to their local demographics based on a judgement of their 'home location' (first described in Section 1.3).

A set of geo-located Tweets from the United Kingdom is collected using the Twitter Streaming API, which is filtered to select users who chose to enable coordinate level geo-location of their Tweets. Historic Tweets were collected for each of these users to create a dataset containing users represented by their Tweets, and any additional associated profile information.

In Chapter 5 we proposed and evaluated several methods for assigning high quality home location estimates to Twitter users, identifying our novel clustering method as the best choice going forward. We implement and apply this clustering approach to the collection of each user's geo-located Tweets derive a 'home location' for each user.

Having derived a home location for each user, we look up the OA and LA regions the profiles lie within. These regions are then mapped onto associated UK measures of socio-economic status: the Output Area Classification (OAC) and the

Local Authority Classification (LAC), to create a labelled dataset of Twitter users (Section 6.2.2). Prior to this work, OAC and LAC were unexplored in the user profiling literature.

We leverage the labelled OAC and LAC Twitter user datasets to create two user profiling systems (Section 6.3). The developed user profiling systems incorporate the strong baseline components which were evaluated in Chapters 3 and 4; specifically, we derive a term frequency-inverse document frequency (TF-IDF) weighted feature vector of n-grams from the Tweets of each user, and use this to train a support vector machine (SVM) classifier. Both systems exceed a simple random baseline approach, and the LAC system in particular achieves promising results; detailed results and discussion are presented in Section 6.4.

This work was presented at NLP+CSS 2016 [156], and code/resources were made available at `https://github.com/adampoulston/geo-user-profiling`.

6.1 Data

The work in this chapter makes use of data from two sources: census-derived demographic data (the OAC and LAC), and public Twitter posts.

6.1.1 Demographic data

Demographic data provides information about characteristics (e.g. age, religion, ethnicity) of a population within a specified area. The UK government provides open datasets containing information about a range of demographic variables including highest qualification, job category and unemployment rates.

This chapter makes use of *geo-demographic segmentation* datasets in which an area's demographics are generalised into a single socio-economic category. These types of data sets are often used for marketing purposes [157]. The United Kingdom's Office for National Statistics (ONS) [1] provides a range of data sets including the OAC and LAC datasets. Unlike commercial datasets, such as MOSAIC[2] and

[1] `http://www.ons.gov.uk`
[2] `http://www.experian.co.uk/mosaic/`

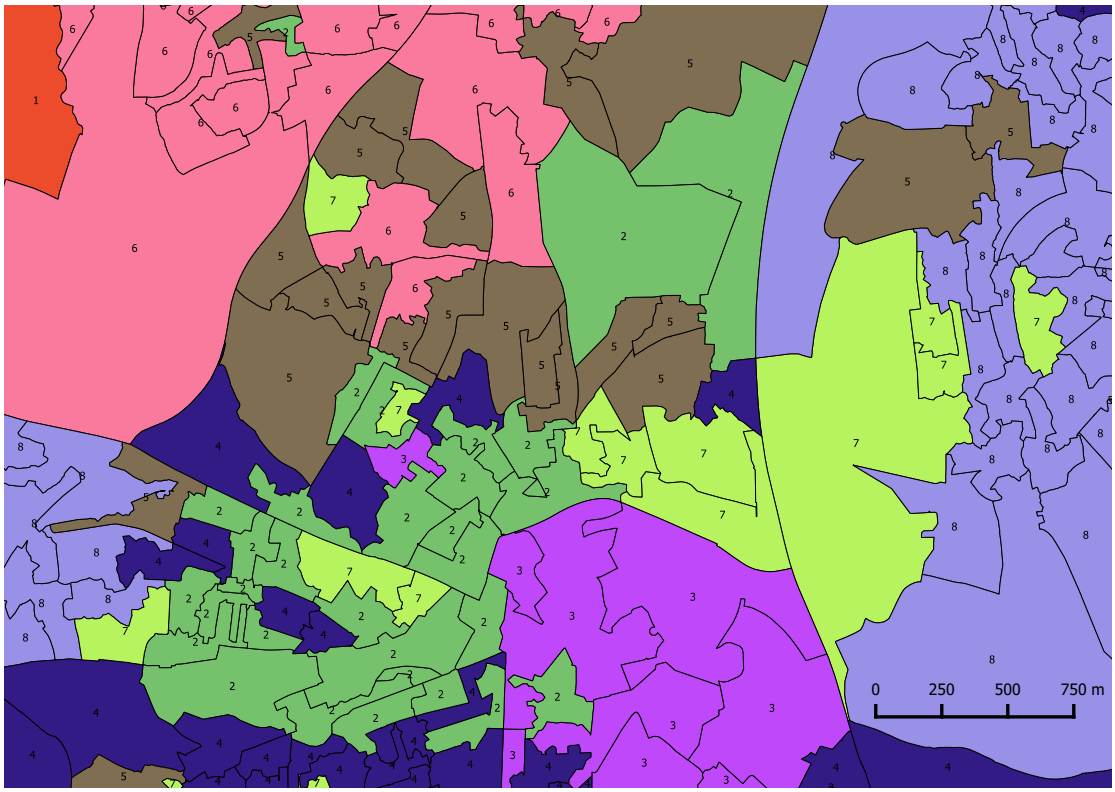

Figure 6.1: Selection of Output Areas labelled with their OAC supergroups.

Acorn[3], the methodology used to develop the OAC and LAC datasets is fully documented.

The OAC data set is organised around OAs, regions of around 200 households in England and Wales. The OAC dataset places residents of every OA into a hierarchy of socio-economic groups based on responses to the 2011 UK Census. The dataset consists of a hierarchical classification scheme with three layers: supergroups (shown in Figure 6.2), groups and subgroups. For example, a densely populated region in central Middlesbrough, North East England, is named as OA E00060869 and is associated with the '7-constrained city dwellers' supergroup , the '7b-constrained flat dwellers' group, and the '7b2-deprived neighbourhoods' subgroup. Figure 6.1 illustrates an example region in the north of England, and shows each OA with its OAC supergroup; note the level of variation in OAC between OAs, and how single examples of an OAC supergroup are often within areas that are otherwise an entirely different classification.

The LAC dataset is organised in a similar way to the OAC dataset, with eight

[3] `http://acorn.caci.co.uk/`

supergroups (shown in Figure 6.3) followed by groups and subgroups. Unlike the OAC, which contains classifications of specific small regions of around 200 households, the LAC is generalized to cover LAs, which describe areas governed by a single council across the whole of the UK, and are significantly larger than OAs; there are several hundred LAs in the UK, each of which contains hundreds of OAs. Despite some similarities in supergroup names, such as OAC 6 - 'suburbanites' and LAC 4 - 'suburban traits', the two datasets use different classification strategies leading to categories not being directly comparable.

6.1.2 Geo-located social media posts

Geo-located social media posts from the United Kingdom were identified using the Twitter public streaming API [4]. The Twitter REST API was then used to retrospectively collect each user's Tweets (up to 3,200 per user) and any public information on their profile. Users with fewer than 50 geo-located Tweets were excluded to ensure that each user profile in the dataset has enough data to derive a robust home location estimate. Excluding these users ensures that users in our the datasets have sufficient 'evidence' behind their home location estimates (and associated measures of uncertainty), and helps avoid propagating errors in home location prediction into subsequent steps. We selected the threshold of 50 Tweets through ad-hoc experimentation, having observed better estimates of home location for users with higher numbers of geo-located Tweets. Just over 135,000 profiles were initially collected, 86,262 exceeded the threshold of 50 geo-located Tweets.

A small portion of profiles (3,743) not representative of the general population (e.g. profiles of celebrities, shops, spammers) were present in the dataset. Profiles of this nature are typically not managed by or representative of an individual, and therefore are unsuitable for linking to a specific demographic. Non-representative profiles were excluded using standard approaches [74, 75], leaving 82,519 profiles used for experiments described later.

[4] `https://dev.twitter.com/streaming/overview`

Supergroups:

1 Rural Residents
2 Cosmopolitans
3 Ethnicity Central
4 Multicultural Metropolitans
5 Urbanites
6 Suburbanites
7 Constrained City Dwellers
8 Hard-Pressed Living

Full hierarchy extract:

5 Urbanites
- **5a** Urban Professionals and Families
  - **5a1** White Professionals
  - **5a2** Multi-Ethnic Professionals with Families
  - **5a3** Families in Terraces and Flats
- **5b** Ageing Urban Living
  - **5b1** Delayed Retirement
  - **5b2** Communal Retirement
  - **5b3** Self-Sufficient Retirement

6 Suburbanites
- **6a** Suburban Achievers
  - **6a1** Indian Tech Achievers
  - **6a2** Comfortable Suburbia
  - **6a3** Detached Retirement Living
  - **6a4** Ageing in Suburbia
- **6b** Semi-Detached Suburbia
  - **6b1** Multi-Ethnic Suburbia
  - **6b2** White Suburban Communities
  - **6b3** Semi-Detached Ageing
  - **6b4** Older Workers and Retirement

Figure 6.2: OAC supergroups and an extract showing the supergroup-group-subgroup hierarchy for two supergroups.

Supergroups:

1 English and Welsh Countryside
2 Scottish and Northern Irish Countryside
3 London Cosmopolitan
4 Suburban Traits
5 Business and Education Centres
6 Coast and Heritage
7 Prosperous England
8 Mining Heritage and Manufacturing

Full hierarchy extract:

3 London Cosmopolitan

- **3a** -London Cosmopolitan Suburbia
  - **3a1** Cosmopolitan North London
  - **3a2** Cosmopolitan South London
- **3b** -London Cosmopolitan Central
  - **3b1** Cosmopolitan Inner London
  - **3b2** Cosmopolitan Heart of London

4 Suburban Traits

- **4a** Growth Areas and Cities
  - **4a1** City Periphery
  - **4a2** Expanding Areas and Established Cities
- **4b** Multicultural Suburbs
  - **4b1** Multicultural Suburbs

Figure 6.3: LAC supergroups and an extract showing the supergroup-group-subgroup hierarchy for two supergroups.

6.2 Demographic dataset creation

In this section we describe the steps taken to implement and apply our novel method for user profiling dataset generation, which we described in Section 1.3. We recap our chosen clustering based method for home location allocation from Chapter 5 in Section 6.2.1, alongside a short experiment validating the method with regard to the Twitter dataset from Section 6.1.2. The end-to-end application of our user profiling dataset generation method is described in Section 6.2.2, where it is applied to derive two user profiling datasets covering the OAC and LAC variables described in Section 6.2.2.

6.2.1 Home location allocation

Geo-demographic data provides information about individuals based on their residential address, therefore it is imperative that a user is associated with that location rather than where they happened to be when sending a particular Tweet. Consequently all users in the dataset described in Section 6.1.2 were assigned a 'home location' in the form of a latitude-longitude coordinate.

We implemented the clustering approach described in Chapter 5 to assign 'home location' to each profile. It is assumed that each user posts from a limited set of locations, that the location posted from the most often is the user's home location. We do not account for the possibility that some users may Tweet more often from another location (such as place of work). Other approaches for assigning home location were considered, such as as those that consider textual [92] and social network [87] cues, but these typically only produce accurate judgements at the city level, whereas demographic datasets often operate at a finer scale. These approaches were reviewed in Chapter 5, and shown to be unsuitable for the task at hand. Specifically, the coordinates of each user's geo-located posts are clustered using k-means, with k set using the 'jump' method [158]. In Chapter 5, a range of alternative clustering algorithms were also explored and all were found to perform similarly. The most populous cluster was then identified and the point closest to the cluster centroid taken as the 'home location'. For full implementation details and a demonstration of performance, see Chapter 5.

Cluster density was calculated; defined as the average Vincenty distance [93] (a

measure of geographic distance) in miles between each data point and the cluster centroid. This density can be seen as a measure of uncertainty around the home location estimate, and provides the option to exclude users with an uncertain home location (i.e low density home cluster). Formally, to calculate the density of cluster i, let $\boldsymbol{p}_{i,j}$, for $j = 1, \ldots, N_i$, be the collection of data points in cluster i, consisting of N_i points. The cluster centroid (from k-means) is given by $\boldsymbol{\mu}_i$. The density of cluster i is defined as

$$d_i = \frac{1}{N_i} \sum_{j=1}^{N_i} |\boldsymbol{p}_{i,j} - \boldsymbol{\mu}_i|, \tag{6.1}$$

where $|\cdot|$ is standard Vincenty distance [93].

Validation of assigned home locations

In addition to the experiments validating our clustering approach for home location allocation in Chapter 5, we performed an additional experiment to verify the home location predictions made here, and ensuring that the same conclusions carry over to a new dataset.

Self-reported locations from the 'location' field were compared with those assigned by clustering. Only 728 of the 82,519 profiles include a self-reported location. Of these, 176 were discarded as being clearly fictitious; leaving 552 profiles for evaluation. These were further cleaned by manually removing extraneous information such as emoticons.

As with the evaluation in Chapter 5, varying levels of granularity were present in the declared location fields, ranging from street level to country, with the majority at town or city level, e.g. 'Edinburgh'. A number of the location fields also included a single coordinate location. The Nominatim geo-coding tool[5] was used to convert the self-reported locations to geographical coordinates. Vincenty distance[93], expressed in miles, between these coordinates and the assigned home location was calculated.

Figure 6.4 shows a histogram and kernel density estimate (KDE) (estimate of the probability density function) [159, 160] for the distances. The majority of dis-

[5] `http://openstreetmap.org/`

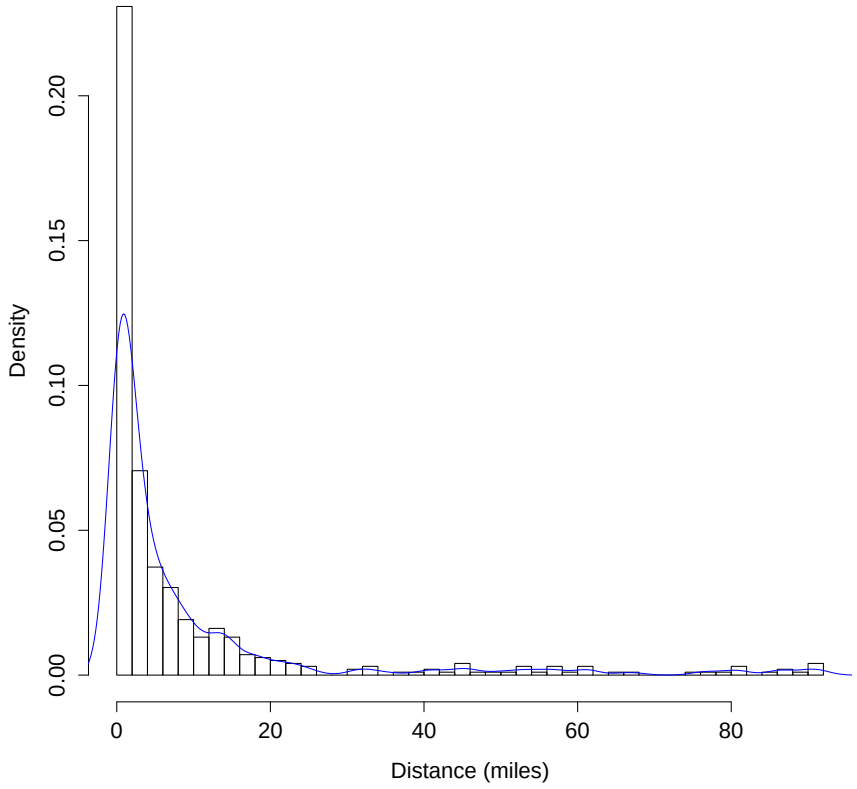

Figure 6.4: Histogram and kernel density estimate of distances. Note that this figure has been truncated at the 90% quantile of distances due to a number or large errors above the 90% quantile.

tances (69.7%) were accurate to within 10 miles, more than half (56.9%) accurate to within 5 miles and 30.8% within 1 mile. The home location gained from the location text field is only expected to be accurate to within 5 or 10 miles because the majority of self-reported locations are towns or cities, and as discussed in Chapter 5, the location field is itself unreliable, so we expect a number or errors where the user has not accurately declared their location. These results suggest that the clustering approach presented here is successfully identifying home location in the same broad locality as declared in the users location fields most of the time. Coupled with the more detailed evaluation in Chapter 5, we conclude that the clustering method is performing with an acceptable level of accuracy.

6.2.2 Demographic linking

In Section 1.3, we described our proposed method for generating social media user profiling datasets, where users are assigned a 'home location' which is used to link the user to their local demographics. In this section we describe the application of this process to a set of Twitter users, utilising two demographic datasets, the OAC and the LAC.

Using the 'clustering' home location method described in Section 6.2.1, we inferred a coordinate 'home location' and 'uncertainty' value for each of the 82,519 profiles identified of Section 6.1. Point-in-polygon tests were then used to link each user with its containing OA and LA boundary. We did not utilise the variant combined clustering/majority-vote approach proposed in Chapter 5 here, where the users coordinates are clustered, and the OA/LA of each point in the user's home cluster is looked up. This expanded approach could potentially lead to more accurate judgements of home location, and would be a good avenue to investigate in future work.

Of the initial 82,519 profiles, 69,723 had a home location within the areas covered by the OAC, and 74,749 had a home location within the areas covered by the LAC. Each profile geo-located within an OA or LA, was then annotated by looking up the associated OAC or LAC supergroup. We do not attempt to classify at the group or subgroup level in this work, and as such did not annotate the profiles to this level, although this would be trivial to implement in future work.

By way of an example, the process for an individual user might go:

1. Identify user's home coordinate.
2. Home coordinate falls in OA E00171217 (OAC supergroup 2).
3. Label the user with OAC supergroup 2.

Two user profiling datasets were created by linking users with their local demographics; users in England and Wales were labelled with one of eight OAC supergroups associated with that user's local OA, and users across the whole of the UK were labelled with one of eight LAC supergroups associated with their LA. These datasets are referred to as Output Area Classification profiles (OAC-P) and

Dataset	# Profiles	Proportion per Supergoup							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
OAC–P	69723	0.133	0.128	0.083	0.117	0.194	0.160	0.056	0.129
LAC–P	74749	0.160	0.059	0.108	0.116	0.191	0.058	0.127	0.181

Table 6.1: Dataset statistics for OAC–P and LAC–P.

Local Authority Classification profiles (LAC-P), respectively. Table 6.1 shows the proportions of each label and the number of profiles per dataset.

6.3 User demographic prediction

In Section 6.2 we constructed a corpus of Twitter users whose home location is known, and leveraged this to create two datasets of Twitter users whose OAC (OAC-P) and LAC (LAC-P) supergroup is known. This method for generating user profiling datasets is in and of itself a user profiling method, but is not suitable for widespread application as most users do not create the volume of geo-located posts required to accurately assign a home location estimate. Therefore, we wish to use these datasets to create user profiling systems capable of predicting the OAC and LAC for users whose home location *is not* known.

In this section, we combine the OAC-P and LAC-P datasets developed in Section 6.2.2, with the strong baseline users profiling system components identified is Chapters 3 and 4 to create two user profiling systems that predict OAC and LAC respectively.

We approach both user profiling systems as multi-class classification problems, and aim to use the content of a user's Tweets to predict their OAC and LAC supergroup from the eight possible classifications in each data set.

6.3.1 Classification approach

We developed a machine learning based user profiling classification approach that incorporates the components we identified as well suited to user profiling systems in Chapters 3 and 4. Specifically, a classification pipeline was created, that takes

n-gram features extracted from each user's corpus of Tweets as input to train an SVM classifier. n-grams and SVMs were chosen as they were shown in our own work in Chapters 3 and 4 to perform well on similar user profiling tasks, and have also been shown to consistently perform well at other user profiling tasks, both for social media [13, 58, 64] and other types of text [161, 162].

Prior to feature extraction, the text of each user's Tweets was preprocessed in line with our successful user profiling approaches in Chapters 3 and 4. Tokenisation was performed using a Twitter-aware tokeniser [116]. Only English Tweets were included (determined using the language property in the data returned from the Twitter API), so no steps were taken to address the complexities of other languages. Emoji were not removed, and were treated as normal characters. Tokens that appeared in more than 90% of users Tweets were removed, which results in the removal of overly common terms and stop-words without the usage of a specific dictionary, which often do not handle the non-standard nature of social media text.

Word level unigrams (1-grams) and bigrams (2-grams) were extracted for each Tweet, following the processes described in Section 3.2.2, and aggregated up to the user level. We showed in Chapters 3 and 4 that unigrams and bigrams perform well in user profiling systems, and found that inclusion higher order n-grams do not yield a significant performance increase. TF-IDF weighting was applied to down-weight n-grams found to be common across all users and assign higher weights to n-grams which are less common, and therefore more likely to be useful.

The extracted n-grams were used to SVMs with linear kernels using the implementation available in scikit-learn [119]. To handle the multi-class nature of the dataset addressed here, a one-vs-the-rest approach was utilised. Specifically, a classifier was trained for each label to discriminate between that label and all other labels. At inference time, a label is selected based on the classifier which positively predicts a label with the highest "certainty". The established utility of such an approach makes it a useful tool to establish baseline performance of models on novel datasets generated through our methods. We do not consider any models here that might achieve higher performance metrics such as neural networks, as we are more interested in establishing that the datasets can be used to train *a model* rather then eking out extra accuracy points.

Approach	OAC–P	LAC–P
Random Baseline	0.1259	0.1259
SVM classifier	0.2757	0.5047

Table 6.2: Average accuracy for OAC–P and LAC–P prediction across all folds.

6.4 Results and discussion

The accuracy of our n-gram and SVM based user profiling pipeline (described in Section 6.3.1) applied to the OAC-P and LAC-P datasets are presented in Table 6.2, alongside a random baseline approach. Due to constraints on computational resource, we were unable to train our user profiling systems on the *whole* of OAC-P and LAC-P. Subsets stratified by supergroup were extracted from the OAC-P and LAC-P datasets with 2,000 members per supergroup label, and used to train the models in both cases. Separate subsets were selected at random from both OAC-P and LAC-P. No effort was made to further stratify based on the underlying home location data used to perform the initial labelling, or to stratify based on groups and subgroups, or to limit our subsets to include profiles with lower (higher certainty) home cluster densities. Despite the underlying datasets being imbalanced, we chose to down-sample to balanced subsets as machine learning algorithms such as SVM are prone to over-fitting towards more common labels in datasets with imbalanced label distributions [163]. Cross-validation (4-fold) was used to compute average performance metrics.

The user profiling system for both OAC and LAC are able to predict one of the eight demographic variables with better than random performance. An average accuracy score across folds of 0.2757 was achieved for the OAC, which comfortably exceeds the random baseline, but is still low enough to indicate that some element of the process was not especially successful. Candidate explanations for this disappointing performance include: the automatically labelled training data contains many incorrectly labelled users; the variables being predicted do not contain especially obvious variations in language; and, our proposed pipeline is less suitable than we thought.

Results for LAC are more encouraging, achieving an average accuracy across folds of 0.5047. This indicates that it is possible to achieve promising results using our 'tried-and-tested' classification approach in conjunction with our novel data

generation approach. We can conclude therefore, that the noticeably lower performance on the OAC is likely down to some property in the underlying dataset derived through our user profiling dataset generation method.

6.4.1 Geographic properties

The large difference in performance obtained by the models trained on OAC-P and LAC-P is likely down to the geographic nature of the underlying regions used to geo-locate the users (OAs and LAs). Table 6.3 shows an analysis of the average 'size' of the two region types; size for a given OA or LA was measured as the average length of the diagonal of the minimum bounding rectangle for each region. Also shown is the proportion of users in each dataset whose 'home cluster' density (our measure for home location judgement 'certainty') was less than the average length of the region type.

Regions defined in the OAC-P dataset are much more granular than those in the LAC-P dataset; the average length region is 0.93 miles for OAs, whereas it is 34.5 miles for LAs. Only 35.2204% of the home cluster densities in OAC-P were less than 0.93 miles, in contrast to the LAC-P, where 83.5010% were within 34.5 miles.

The density of a home location cluster can be seen a proxy for how 'certain' the judgement is; a higher value indicates underlying points with a wider geographic spread, meaning the single predicted point is less likely to be correct. Given that nearly 65% of profiles in the OAC-P dataset have a certainty that overlaps the average region length, and therefore may lie in an adjoining OA, it is more likely for profiles to be mis-classified when assigned to the more fine-grained regions in the OAC-P data set, resulting in a noisier data set, and poorer predictive performance in downstream models. This is less of an issue with the LAC-P dataset since accuracy of geo-location for the majority of profiles was within the average LA region length; nevertheless almost 17% of the profiles still have a fair chance of mis-classification.

In this experiment we did not take cluster density into account in the dataset generation step and naively accepted profiles with a home–cluster of any density, it is highly likely that conditioning our novel data generation approach to include only those profiles with a good 'certainty' value would improve the performance

Region type	Average length	Proportion of profiles
OA	0.9308	35.22%
LA	34.5252	83.50%

Table 6.3: Average geographic unit length in miles for Output Areas and Local Authorities with proportion of Twitter profiles in OAC-P and LAC-P whose home cluster density is less than the average OA and LA unit length.

of resulting models.

6.4.2 Dataset characteristics

There is a difference in the core aims behind the demographic datasets that underpin the OAC-P and LAC-P user profiling datasets. The OAC scheme aims to model 'geographically independent socio-economic status' in contrast to the LAC categories which are more region dependent, including categories such as 'London Cosmopolitan'. This difference in ideology is also likely to change the characteristics of the generated user profiling datasets, which will be represented in the sorts of features deemed 'important' in derived models, and the mis-classifications a model makes.

Confusion matrices

To investigate the ways in which the models for OAC and LAC were making incorrect classifications we computed confusion matrices for a representative run of both models.

Figure 6.5 shows the normalised confusion matrix for the OAC model. The model appears to lean towards predicting supergroups 1 (Rural Residents), 3 (Ethnicity Central), 8 (Hard-Pressed Living), with 611, 803 and 672 total predictions, respectively, when only 500 of each label are present. We note that supergroups 2 (Cosmopolitans) and 4 (Multicultural Metropolitans) are frequently misclassified as 3. 2 is classified as 3 182 times, and only correctly classified as 2 133 times. Similarly for 4 it is classified as 3 129 times and only correct 105 times. These three groups all represent groups of people active in people active in large cities where more mixing of demographics happens, possibly leading to three groups

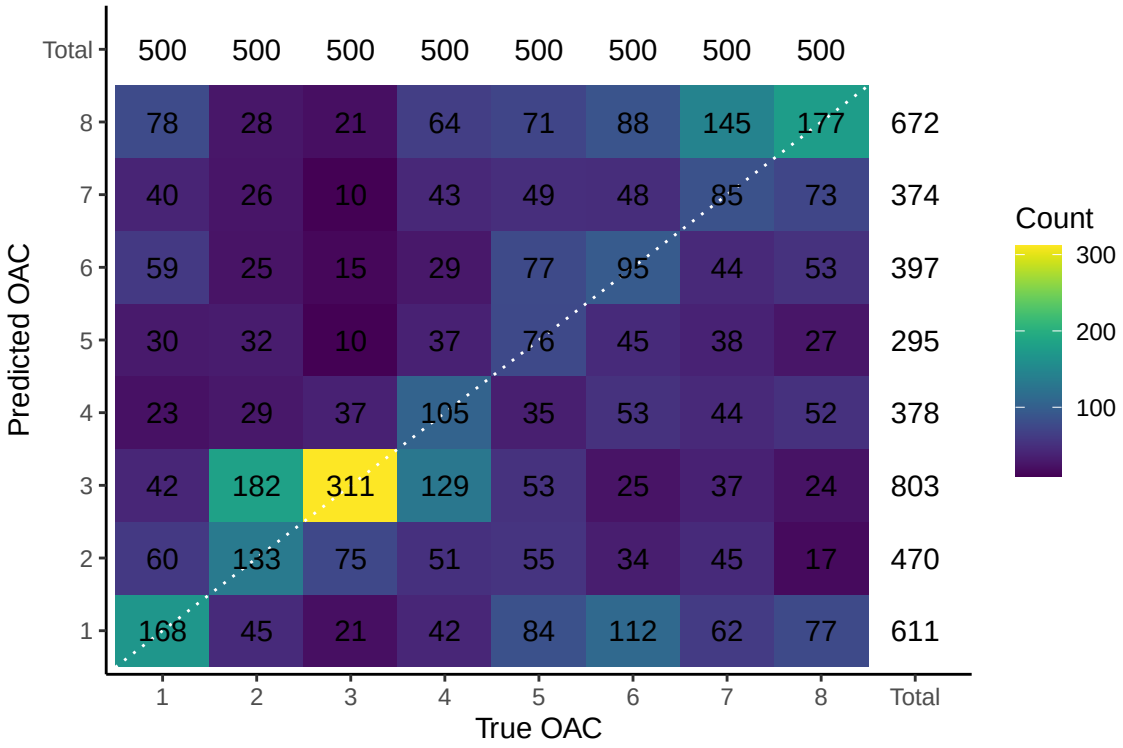

Figure 6.5: Heatmap showing the frequency at which each OAC Supergroup was classified as every other supergroup by our user profiling pipeline. Taken at random from a representative run.

with broadly similar individuals, resulting in the poor predictive performance of the OAC model. A similar pattern can be seen between supergroups 7 (Constrained City Dwellers) and 8 (Hard-Pressed Living), which both represent more deprived people living in smaller cities. Group 6 (Suburbanites) is often predicted as group 1, but not vice versa; suburban areas typically exist between urban and rural communities, which may explain this pattern.

Figure 6.6 shows the normalised confusion matrix for the LAC model. The model performs best for users in supergroups 2 (Scottish and Northern Irish Countryside), 3 (London Cosmopolitan) 6 (Coast and Heritage), and 8 (Mining Heritage and Manufacturing), and reasonably well on the other supergroups. The user profiling system got 494/500 correct for supergroup 2, and 412/500 for supergroup 3, which is likely due to strong dialect/regional features associated with Scotland and London. While less pronounced, a similar pattern of misclassification of similar groups can be seen here as well between classes 3, 4 (Suburban Traits), and 7 (Prosperous England). While the separation between these supergroups is much more pronounced than those in the OAC, all three cover a demographic that

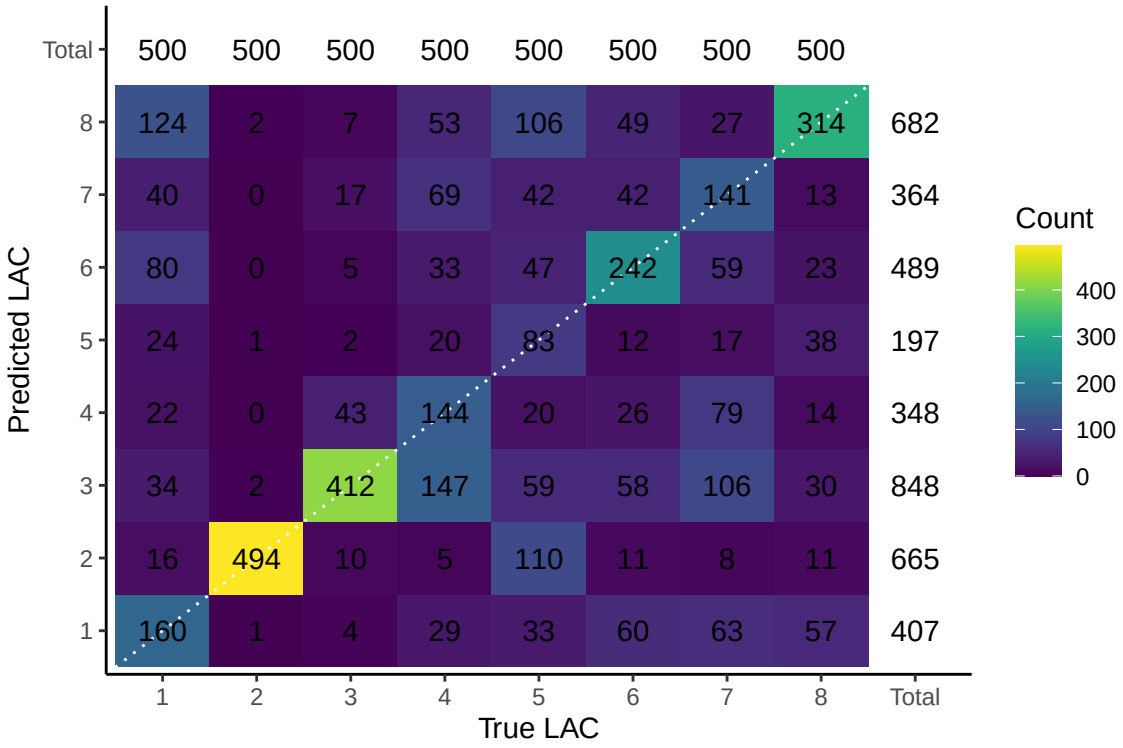

Figure 6.6: Heatmap showing the frequency at which each LAC Supergroup was classified as every other supergroup by our user profiling pipeline. Taken at random from a representative run.

could broadly be referred to as the 'middle class'. The same phenomenon can also be seen between supergroup 1 (English and Welsh Countryside) and 8, which is intuitive, as many former mining communities are in relatively rural areas.

Important features

To examine the features deemed as important by the SVM models for each OAC and LAC supergroup, the model trained for a fold was selected at random and the feature coefficients were inspected. The top 30 important features are presented in Table 6.4 for the OAC and Table 6.5 for the LAC.

Examination of the highest ranked features by SVM coefficient for each LAC supergroup revealed a connection between groups and geography. The most important features for many classes are words or phrases referencing specific areas in the UK as well as several stereotypical dialect features. The links to geography are most pronounced in the two best performing supergroups, 2 and 3. Supergroup

2's (Scottish and Northern Irish Countryside) highest ranked features cover both geographic locations, such as 'clyde', 'dunfermline' and 'glasgow', dialect features such as 'wee', 'whit', and 'tae', as well as cultural references such as 'irn bru', and 'rangers'. Supergroup 3's (London Cosmopolitan) highest ranked features relate exclusively to London, its surrounding boroughs and public transport system.

In contrast, the feature coefficients for the OAC model are not as location dependent; for example, '1-Rural Residents' contains features such as 'Severn' (a river), 'stables', 'mountain bike' and 'emmerdale' (a UK soap opera set in the countryside). Similarly, '4-Multicultural Metropolitans' is the only group identified that has non-English phrases and the Islamic holidays Eid and Ramadan as important features. Only '3-Ethnicity Central' is dominated by features specific to a particular locale, London, despite being present in cities across the UK; this behaviour is indicative that in future work, additional effort should be made to ensure even representation across the top level administrative region (UK in this case) in user profiling datasets derived through our approach.

We observe that the features deemed as important, in many cases, match nicely with the motivations behind the underlying supergroups; while not perfect (likely due to the issues with uncertain home location estimates discussed in Section 6.4.1), this indicates that the resulting models are–at least to some extent–picking up the same sorts of information modelled in the underlying demographic data, indicating value in the datasets derived through our method.

Subclasses

As noted in Section 6.2.2, the analysis in this chapter is focussed on the prediction of the top level supergroups in the OAC and LAC, and did not did not label the users in OAC-P and LAC-P with the additional subclasses present in the groups and subgroups of both hierarchies. As such we were unable to analyse the distribution of these subclasses in the datasets used to train our user profiling systems, or assess the affect the presence (or lack of) of particular subclasses might have on predictive performance. For example, we noted earlier in this section that OAC group 6 is often predicted as group 1, but not vice versa; it is possible that this pattern emerges due to properties of the underlying classification, and in fact a particular subclass of group 6 are indeed more similar to users in group 1. In future work, profiles could be labelled with these subclasses and further

characterisation of these sorts of 'errors' could be performed.

6.5 Conclusion

This chapter explored the use of our method for combining population-level demographic information with geo-located social media profiles for user profiling. Our novel approach to the generation of automatically labelled data by making use of geo-located social media posts was implemented and applied to two demographic variables previously unadressed in the literature.

The demographic datasets used to label Twitter users in this work have the advantages that they are large-scale and collected using sound methodologies. However, the information they contain is aggregated and is updated infrequently.

The 'home location' for a user is identified using clustering and then combined with publicly available information from two previously unexplored demographic datasets. A well established user profiling classification pipeline based solely on Tweet content was able to predict socio-economic status with promising results for one data set.

Analysis indicated that the properties of the demographic data are important when considering which demographic datasets to utilise. Key factors include the granularity of the associated boundary and degree to which the groupings are based on socio-economic, rather than geographic, characteristics.

In the next chapter we will apply our method to a demographic variable and associated user profiling dataset that has previously been addressed in the literature, to further explore the extent to which some of these disadvantages can be overcome, and better quantify how our approach compares to datasets derived through other methods.

1	**in newquay**, tomoz, **porth**, #education, to seeing, hollyoaks, . @, stables, min/mile, . xx, severn, mountain bike, cambs, wish you, waste of, on sat, harbour, @mention listen, —, **cambridge**, mi and, morrisons, fab !, ya tweet_end, nest, emmerdale, visitors, board, it's friday, tweet_end everyone
2	linda, #ucl, terrace, #cocktails, EMOJI_STRING, here to, **notts**, regram, @ the, (via, just over, then off, **oxford street**, display, the lecture, EMOJI_STRING, quoted, #halloween, **mayfair**, la, @hyperlink #hiring, EMOJI_STRING, the life, housemates, opening :, **central london**, via, tweet_end **#brighton**, tweet_end ., tennis
3	EMOJI_STRING, yes yes, wharf, **london and**, **woolwich**, **southbank**, carriage, **hammersmith**, ——, cab, @mention check, **finsbury park**, old woman, **old street**, **catford**, **finsbury**, **battersea**, central line, **charing cross**, uber, #startup, **highbury**, . EMOJI_STRING, vauxhall, **london is**, **the thames**, **bethnal**, @mention feat, for coming, **streatham**
4	was really, **to watford**, EMOJI_STRING, jimmy, EMOJI_STRING, tweet_end them, #wba, lool ., tweet_end pls, have me, . ok, dot, route to, one who, no more, you absolute, alot of, worship, on united, that for, :(((, line tweet_end, o2, i'm listening, virgin, loool tweet_end, the bees, ARABIC_STRING, EMOJI_STRING, one can
5	start playing, business in, clothes, so just, **worthing**, blog post, #tunbridgewells, costa, **lakeside**, the cinema, recruitment, bring on, awesome !!!, luke, !!! EMOJI_STRING, aboard, firing, thankyou, hypocrite, !! :d, photo :, sir tweet_end, EMOJI_STRING tweet_end, poults, EMOJI_STRING tweet_end, even just, only two, ah tweet_end, @mention reply, in trouble
6	tomorrow @mention, : love, dragging, year 8, 4 a, EMOJI_STRING, now ?, jess, my car, chap, do do, @mention obviously, @mention wake, go so, europa, amaze, @mention EMOJI_STRING, and half, only tuesday, asked if, coat, in biology, you monday, banter tweet_end, 10/10, town ., @mention omg, burton, aa, you ok
7	in tomorrow, your a, ;3, days off, tweet_end feeling, the weather, scouse, tweet_end had, see someone, try again, ' em, think its, tweet_end trust, a burger, , we'll, carvery, the minute, **to edinburgh**, us and, **newcastle**, thru, **exeter**, archie, EMOJI_STRING tweet_end, just wondering, ynwa, **mersey**, tweet_end feels, enjoying the
8	i seen, alive tweet_end, tom, my head, rather be, ha xx, only now, try my, housework, bed time, hubby, nana, soz tweet_end, creased at, up ya, my son, being funny, they got, dick, :) x, for is, divvy, video @hyperlink, and better, @mention bet, a twat, tweet_end creased, please please, me it's, mun tweet_end

Table 6.4: Top 30 features for each OAC supergroup. Location linked features are highlighted in bold. EMOJI_STRING represents a specific sequence of one or more emoji not separated by a space.

1	hell of, **the norwich**, **north wales**, my ass, county, holt, **in exeter**, thankyou, year 9, with mum, **birmingham**, cuz i, **drayton**, **malvern**, well that's, **plymouth**, **dorset**, **selby**, **bewdley**, **in preston**, **wisbech**, **in chorley**, quay, **in rhyl**, **beccles**, **in crewe**, abit, **anglia**, **gloucester**, **knutsford**
2	**clyde**, tweet_end sitting, nice wee, **glasgow !**, poor wee, sitting, wee day, hate when, gran, couch, **for scotland**, **glasgow tweet_end**, whit, this wee, **dunfermline**, wee man, good wee, aw, **rangers**, bru, cheeky wee, ma, c'mon, **scotland's**, aff, tae, irn, **to aberdeen**, long lie, braw
3	**in clapham**, exhibition, **to london**, **charing**, **bethnal**, street, **west london**, **ealing**, vauxhall, **in islington**, **hampstead heath**, **kensington**, east, **canary**, **peckham**, **southwark**, **walthamstow**, **#brixton**, **trafalgar**, **#london @**, **barbican**, **in greenwich**, **in brixton**, **underground**, **waterloo**, tube ., cab, uber, **london bridge**, **newington**
4	fans, **bexleyheath**, just clocked, omg the, **in barnet**, coz, atleast, in mk, g x, borough, **chatham**, : still, so pissed, towie, **albans**, a cab, geezer, **in romford**, **lakeside**, @mention lol, eastenders, birthday my, o2, cafe, piff, **peterborough**, looool tweet_end, **in uxbridge**, **north london**, **in ipswich**
5	#wawaw tweet_end, **in salford**, rd, **cardiff bay**, **hallam**, **of cardiff**, freshers, **derby**, gatecrasher, leadmill, flatmate, **gyle**, casino, town with, virgin, a fuckin, **tweet_end hull**, flatmates, **in cardiff**, #wawaw, **brum**, **dundee**, some reyt, **in #brighton**, **fallowfield**, **#bristol**, a bus, yeye, , **manchester**
6	isi, **plymouth**, **bexhill**, **#colchester**, **in hastings**, **in cheltenham**, **, york**, **plymouth .**, **somerset**, **#bath**, **of bath**, haha tweet_end, **#swansea**, **brighton**, **in worthing**, **#plymouth tweet_end**, **broadstairs**, **bristol**, down here, **at york**, **york ,,** mush tweet_end, the york, **herne bay**, **in eastbourne**, **morecambe**, **york !**, the swans, **york .**
7	**oxford .**, **hampshire)**, m40, **in solihull**, **leamington**, **thame**, @mention alright, **oxford tweet_end**, saints, **oxford ,,** **, berkshire**, cba to, the m25, blimey, **camberley**, **#harrogate**, ollie, **warwick**, **reading**, ergh, **into london**, **huntingdon**, mk, **#cambridge**, **harrogate**, **melksham**, **#essex**, **brentwood**, **eastleigh**, **marlow**, **cambridgeshire**
8	**colne**, aye, darlo, **burnley**, reyt, carnt, **in warrington**, a mint, **brighouse**, **stockton**, #safc tweet_end, **cardiff**, mint tweet_end, orate, lmfao, your mam, **west yorkshire**, boro, anorl, wey, **oldham**, toneet, tweet_end sick, peng, the minute, **barnsley**, **manchester .**, **in newport**, **southport**, **in newcastle**

Table 6.5: Top 30 features for each LAC supergroup. Location linked features are highlighted in bold. EMOJI_STRING represents a specific sequence of one or more emoji not separated by a space.

Chapter 7

Predicting user National Statistics Socio-economic Classification

The experiments with the Output Area Classification (OAC) and Local Authority Classification (LAC) schemes in Chapter 6 implemented our method for assigning demographic variables to social media profiles via geographic information, and demonstrated that models with viable performance can indeed be implemented. It is apparent, however, that the groups represented in the OAC, and especially the LAC, captured more elements of geography than they did socio-economic indicators, leading to doubt that the data/models were capturing individual differences and were instead capturing geographic variation in language.

Our work on user profiling for the OAC and LAC was novel, but therefore not directly comparable to existing work on predicting socio-economic status. As such verification against datasets derived through traditional existing approaches is not possible, which leaves our central research question of whether or not our approach is a suitable alternative/complement to existing user profiling dataset generation approaches somewhat under-explored.

In this chapter we expand on our promising results so far and attempt to form a better impression of how datasets derived through our geo-located driven user attribute labelling approach compare against others, by applying it to a different measure of socio-economic status, the National Statistics Socio-economic Classi-

fication (NS-SEC). This measure has already been investigated by Lampos et al. [134] using a different user attribute labelling approach, and as such is good fit for assessing our approach versus state-of-the-art. Lampos et al. [134] constructed a user profiling dataset annotated with NS-SEC by selecting profiles based on the presence of known job titles in their description field that have established mappings under the NS-SEC classification scheme. The materials and Twitter users used to generate their dataset were acquired, and are used to replicate their user attribute labelling approach and dataset as closely as possible, although direct access to the exact same data was not possible due to Twitter API limitations; full implementation details are provided in Section 7.2.1.

Alongside our reconstruction of the dataset from Lampos et al. [134], we constructed a dataset of Twitter users labelled with NS-SEC using our own geolocation driven user attribute labelling approach. A collection of profiles located in the UK was gathered using the Twitter API, and the clustering method chosen in Chapter 5 was applied to attain an estimate of their local area, and in turn identify the distribution of a measure socio-economic status (NS-SEC) for the identified areas from census data. Users identified to be from an area with a high proportion of a single NS-SEC category were labelled with that category, other users were excluded from the analysis (further details in Section 7.2.2).

Both datasets were used to train predictive models and achieve similar performance (Section 7.3) but both exhibit several shortfalls that limit compatibility. To attempt to overcome the weaknesses unique to each dataset, an experiment was performed to combine them (Section 7.4), which found that a linear ensemble of classifiers trained separately on each dataset was able to outperform the individual models.

In Section 7.5 and additional step is investigated to further improve reliability of the geographically derived dataset. Home location 'certainty' is used to filter down to profiles more likely to hold a correct label, and by adding this additional constraint at dataset generation time accuracy of derived models improves.

7.1 The National Statistics Socio-economic Classification

The work in this chapter makes use of similar sources to Chapter 6; we again utilise census-derived demographic data (the NS-SEC) in conjunction with public boundary data and Twitter posts, but also introduce an existing user profiling dataset to use as a comparison against datasets derived through our own geo-location driven user attribute labelling approach.

The NS-SEC is a measure of socio-economic status used in UK population statistics. Under the NS-SEC individuals are allocated a socio-economic status based on their occupation. Four groupings are available: 17 class (referred to as *operational*), 8 class, 5 class and 3 class, shown in Table 7.1. Only the 3 class version represents a 'hierarchy' of social class. In the 3 and 5 class versions, it is left up to the user to decide whether or not to include the long term unemployed as their own class.

NS-SEC is a variable covered in the UK census, available at the Output Area (OA) level, and is also linked to the Standard Occupational Classification (SOC) [164, 165] a taxonomy of occupation titles associated with a variety of socio-economic outcomes.

7.2 Twitter profile datasets

The Twitter public streaming API[1] was used to collect a sample of Tweets, S from the UK between October 6th and December 6th, 2016. Each unique user with a single UK-geo-located Tweet in S was identified, and the Twitter REST API was used to collect retrospectively their Tweets (up to 3200 per user) and any public information on their profile, giving another dataset, U_{GEO}.

Standard steps were taken to eliminate non-representative users such as brands, celebrities or those profiles under automation (malicious or otherwise) from U_{GEO}. Overly prolific users, taken to be the 1% most active users, were excluded to avoid imbalanced impact by individual users on resulting models. In practice this

[1] `https://dev.twitter.com/streaming/overview`

NS-SEC categories

3 class	5 class	8 class	Operational	Description
1	1	1.1	L1	Employers in large organisations
			L2	Higher managerial occupations
		1.2	L3	Higher professional occupations
		2	L4	Lower professional and higher technical occupations
			L5	Lower managerial occupations
			L6	Higher supervisory occupations
2	2	3	L7	Intermediate occupations
	3	4	L8	Employers in small organisations
			L9	Own account workers
3	4	5	L10	Lower supervisory occupations
			L11	Lower technical occupations
	5	6	L12	Semi-routine occupations
		7	L13	Routine occupations
*	*	8	L14	Never worked and long-term unemployed
N/C	N/C	N/C	L15	Full-time students
			L16	Occupation not stated or inadequately described
			L17	Not classifiable for other reasons

Table 7.1: Official groupings for the NS-SEC.

excluded 7129 profiles across the two month period, leaving a total of 110,367 profiles.

7.2.1 Existing NS-SEC user profiling dataset

To investigate comparability with models trained on a dataset annotated with the same labels but acquired through more 'classic' user attribute labelling approach we additionally acquired the profiles and associated NS-SEC labels used in [134], referred to from here-on as U_{SOC}. Several of the profiles in U_{SOC} were no longer accessible and our snapshot of the existing profiles covers a differing time to those used in [134], limiting exact comparison to reported results.

U_{SOC} was labelled based on the presence of job titles from the 2010 edition of the SOC in description field of Twitter profiles, which were in turn mapped to the NS-SEC. Each profile in the data-set had its job title (and by extension NS-SEC label) verified by the authors, and by extension were implicitly filtered to be profiles controlled by an individual. Despite the strengths and relative certainty of the labelling scheme, several types of user are excluded implicitly; the reliance on job titles immediately excludes users from groups with no defined occupation such as students, non-working parents or the retired. Selecting profiles which only reference a job title is also likely to introduce bias in the resulting dataset towards users who engage with Twitter mainly for professional or networking purposes, as is evident by the lower number of 'low socio-economic status (SES)' profiles in [134].

7.2.2 Labelling profiles with NS-SEC through geography

To label the users in U_{GEO} with NS-SEC, we followed our data generation established in Section 1.3, and validated throughout Chapter 6. To summarise, we first identify the so-called 'home location' of each user by applying K-means clustering to the coordinates of their geo-located posts. In turn we identify which OA (statistical geographic region of roughly 100 households) each user's home location is likely to fall in and derive an associated (3 class) NS-SEC label from it.

Unlike in our experiments with the OAC and LAC in Chapter 6, NS-SEC is not represented as a discrete variable in census data, instead an area has an

OA	NS-SEC High		Middle		Low				Students
	1	2	3	4	5	6	7	8	L15
E00039619	19.4	24.6	15.7	8.9	7.3	9.3	4.0	3.2	7.7
E00039654	9.2	26.1	14.7	10.6	5.0	16.5	8.7	3.7	5.5
E00039680	10.6	27.0	20.8	7.5	8.8	11.5	7.1	1.8	4.9
E00039761	1.8	19.6	17.3	4.8	10.1	19.0	18.5	2.4	6.5
E00040052	3.7	7.9	13.2	5.3	5.3	8.5	11.1	29.1	15.9

Table 7.2: Percentage of each NS-SEC class (8-class and full time students) for five OAs.

associated distribution of the maximum NS-SEC class for each household in the area (illustrated in Table 7.2). As such, an additional step was necessary to convert this distribution into a discrete variable for classification.

We derived discrete NS-SEC labels for any given OA by taking its NS-SEC distribution from the UK 2011 census and binning it according the official 3 class grouping shown in Tables 7.1 and 7.2, with the majority class taken to be the OA's label. Specifically we derive three classes: 'high', 'middle', and 'low'. We additionally record the value of the majority class (e.g. High - 78.2%), so that users in areas that exhibit low levels of homophily (tendency to group with similar people [166]) can be excluded when training predictive models. A limited number of OAs hold NS-SEC category L17, full-time students, as their majority class and as such we additionally record this label in U_{GEO}.

Deriving NS-SEC labels from geography helps overcome several of the limitations from Lampos et al. [134] by allowing SES labels to be estimated for users that would otherwise remain unlabelled based on job title alone, which becomes evident when comparing the distribution of labels between U_{SOC} and U_{GEO} (Table 7.3). In U_{SOC} there is a definite imbalance towards the 'high' class with roughly half as many examples for the two 'lower' classes, implying that users with more 'professional' jobs declare their specific job title more often. This imbalance is not present in U_{GEO} which contains closer proportions for each class and even a slight imbalance towards the 'middle' class.

Despite the derived labels in U_{GEO} covering more of the population than U_{SOC}, they aren't necessarily as reliable as those in U_{SOC} for several reasons; the labels are

Dataset	High	Middle	Low
U_{SOC}	0.5438	0.2309	0.2253
U_{GEO}	0.3427	0.3691	0.2882

Table 7.3: Proportions of each label in U_{SOC} and U_{GEO}.

distantly supervised from areas that aren't entirely one class, errors in deriving the users 'home' propagate through to mislabeling, and the lack of human intervention during the labelling process means profiles not representative of an individual are likely to slip through. In Section 7.5 we attempt to address these issues and assess their impact by applying additional filtering steps to the data generation approach.

7.3 Predictive modelling performance

As mentioned in earlier chapters, the methods for generating U_{GEO} and U_{SOC} are user profiling approaches in their own right, meaning that they take an unlabelled user profiled and make a judgement of a personal characteristic. Both approaches are limited in the quantity of profiles they can be applied to, as both rely on interpreting self-reported data; as such, their true utility lies in being used as input to train predictive systems capable of predicting NS-SEC on profiles without the same labelling cues.

As in previous chapters, we aim to evaluate classifiers trained on U_{GEO} and U_{SOC} on their ability to predict NS-SEC on unseen profiles. Following Lampos et al. [134], we assess the task in 3-class (High, Middle and Low) and binary (High and Middle combined with Low) configurations. We estimate model performance when trained on U_{GEO} and U_{SOC} separately using 10-fold cross validation, and also assess the ability of the model trained on U_{GEO} to predict the manually labelled data in U_{SOC}. Due to the imbalanced nature of U_{SOC} we oversample the minority classes to the same size as the majority class, as per the approach in Lampos et al. [134].

In Chapters 3, 4, and 6, we laid out and evaluated strong baseline components for user profiling systems, which we utilise again here. In the experiments that follow, our chosen model derives term frequency-inverse document frequency (TF-IDF) weighted 1,2-grams from the text of user's Tweets and uses them as features to

train logistic regression (LR) classifiers. This and similar set-ups have shown great performance on similar tasks in both our own work (Chapters 3, 4, and 6) and the literature (Chapter 2). In this case LR was chosen as the classifier due to its ability to produce probabilistic outputs, which we apply in an ensemble fashion in Section 7.4. We chose to apply a different model to the one chosen in Lampos et al. [134], for two reasons: consistency with our own previous work, and an inability to reproduce reported performance on U_{SOC} using their declared pipeline on our (admittedly different) snapshot of the data.

7.3.1 Classifier trained on U_{SOC}

Table 7.4 shows the results where the classifier was trained on U_{SOC}; the results exceed the majority baseline (0.54 in both cases due to the unbalanced nature of the dataset), but exhibit clear signs of over-fitting to the majority class in the 3-class variant, even despite the steps taken to alleviate imbalance by oversampling.

Prediction	Accuracy	Precision	Recall	F1-score
High	-	0.827	0.691	0.753
Middle	-	0.293	0.506	0.371
Low	-	0.507	0.530	0.518
Average	0.632	0.542	0.576	0.547
High	-	0.772	0.723	0.747
Low	-	0.693	0.746	0.719
Average	0.734	0.733	0.735	0.733

Table 7.4: Classifier performance metrics for U_{SOC} estimated using 10-fold cross validation

7.3.2 Classifier trained on U_{GEO}

Due the sheer quantity of data derived through our method in U_{GEO}, it is infeasible to train models on the whole dataset using the chosen predictive approach. Deep learning methods and classic machine learning approaches suited to out-of-core learning [167] could handle the full dataset, but we do not consider those here. We applied 10-fold cross validation in conjunction with the classification pipeline

described in Section 7.3, to identify a suitable number of examples to include at train time.

Models were trained with random samples of training data of increasing size, identifying that performance returns for both the 3-class and binary problems diminish around 4500 training examples per label. Reported results are as such attained using a balanced subset from U_{GEO} with 4500 examples per label, totalling 13500 and 9000 examples for the 3-class and binary tasks respectively.

Results as derived through 10-fold cross-validation for the classifier trained on U_{GEO} to predict NS-SEC in both two- and three-class configurations are shown Table 7.5. For both variants the majority class baselines (0.33 for 3 way, and 0.5 for binary) for accuracy are exceeded, achieving performance metrics broadly similar to those of the classifier trained on U_{SOC}. Despite working with a balanced dataset from the start, we still see a slight lean towards prediction of the 'high' class in the 3-class variant here, although not to the same extent as in Section 7.3.1.

Prediction	**Accuracy**	**Precision**	**Recall**	**F1-score**
High	-	0.63	0.64	0.64
Middle	-	0.51	0.50	0.51
Low	-	0.53	0.53	0.53
Average	0.56	0.56	0.56	0.56
High	-	0.75	0.70	0.72
Middle & Low	-	0.72	0.76	0.74
Average	0.73	0.73	0.73	0.73

Table 7.5: Classifier performance metrics for U_{GEO} estimated using 10-fold cross validation

7.3.3 Predictive capability between datasets

To assess comparability between the two data-sets (and resulting models) we additionally used the model trained on U_{GEO} to predict the labels in U_{SOC}, and the model trained on U_{SOC} to predict the labels in U_{GEO}. Table 7.6 shows the two- and three-class results where a model trained on U_{GEO} was used to predict the manually verified labels in U_{SOC}. Table 7.7 shows the two- and three-class results where a model trained on U_{SOC} was used to predict the manually verified labels

Prediction	Accuracy	Precision	Recall	F1-score
High	-	0.69	0.6	0.64
Middle	-	0.28	0.24	0.26
Low	-	0.37	0.54	0.44
Average	0.50	0.52	0.5	0.51
High	-	0.67	0.63	0.65
Middle & Low	-	0.59	0.64	0.61
Average	0.63	0.64	0.63	0.63

Table 7.6: Classifier performance metrics where U_{GEO} was used as train and U_{SOC} was used as test.

Prediction	Accuracy	Precision	Recall	F1-score
High	-	0.42	0.59	0.49
Middle	-	0.36	0.20	0.25
Low	-	0.40	0.41	0.41
Average	0.40	0.39	0.40	0.38
High	-	0.61	0.52	0.56
Low	-	0.58	0.66	0.62
Average	0.59	0.59	0.59	0.59

Table 7.7: Accuracy, precision, recall and F1 where U_{SOC} was used to train the model and U_{GEO} was used to test.

in U_{GEO}.

Despite promising performance of the models trained on U_{GEO} and U_{SOC} in isolation, it appears the two datasets are not entirely compatible, as indicated by the the worsening results when U_{GEO} was used to predict U_{SOC} and vice-versa. The inability for either model to accurately predict labels derived through a different method indicates that, despite ostensibly predicting the same demographic variable, the two user attribute labelling approaches are generating labels that represent different sorts of user, or that one (or both) of the approaches is resulting in incorrect labels; in Section 7.3.4 we investigate potential causes and solutions.

7.3.4 Dataset difference analysis

Although models trained on U_{GEO} were able to predict the labels from U_{SOC} with reasonable accuracy, they could not predict with equivalent accuracy to those trained on U_{SOC}. A likely cause is bias towards certain types of profile introduced in either user attribute labelling approach. To investigate differences between the two datasets we first gathered a subset of U_{GEO} at random containing the same number of profiles as U_{SOC}. The two datasets were then shuffled and split into two, yielding four subsets U_{GEO1}, U_{GEO2}, U_{SOC1} and U_{SOC2}.

Table 7.8 shows the description field vocabulary overlap for each pair of subsets with stop-words and non-alphanumeric characters removed (except for emoji). To be counted as a vocabulary item in a set a term must have appeared more than twice. The inter-set vocabulary overlap (bottom left and top right corners) and the intra-set vocabulary overlap for U_{GEO} (bottom right corner) remain consistent around 250, where as the intra-set vocabulary overlap for U_{SOC} (top right quadrant) is greater (but not so much greater to indicate a significant difference between U_{SOC} and U_{GEO}).

	U_{SOC1}	U_{SOC2}	U_{GEO1}	U_{GEO2}
U_{SOC1}	526	300	242	245
U_{SOC2}	300	527	254	245
U_{GEO1}	242	254	458	254
U_{GEO2}	245	245	254	452

Table 7.8: Vocabulary overlap (intersection of set of terms) for each pair of datasets.

Table 7.9 shows the cosine distance between each combination of the subsets of U_{SOC} and U_{GEO}. The interset cosine values (bottom left and top right quadrants) average 0.054 and the intraset values average 0.081 ; from this we can again conclude that the description field vocabularies for U_{SOC} and U_{GEO} are more similar to themselves than each other (U_{SOC} more so than U_{GEO}).

Despite the apparent overall similarity between U_{SOC} and U_{GEO} observed from Tables 7.8 and 7.9, inspecting the high frequency terms (Table 7.10) in both datasets points towards a potential difference in the 'types' of user most represented. For U_{SOC} many of the most frequent description field terms (even ignoring the specific job titles selected in dataset creation) clearly relate to the user's nature as a

	U_{SOC1}	U_{SOC2}	U_{GEO1}	U_{GEO2}
U_{SOC1}	-	0.04809	0.07779	0.08491
U_{SOC2}	0.04809	-	0.07761	0.08642
U_{GEO1}	0.07779	0.07761	-	0.06021
U_{GEO2}	0.08491	0.08642	0.06021	-

Table 7.9: Pair wise cosine distance of vocabulary item probability for the each pair of subsets. Intra-set values are highlighted in green and inter-set values in blue.

professional; terms ranked 5, 9, 12 and 13 for example all attempt to distance the views of the user from their respective employer. Additionally, emoji use within the description field for U_{GEO} is more frequent at 257 occurrences than in U_{SOC} at 98, indicating a more formal use of the platform for U_{SOC} users.

Rank	U_{SOC}	U_{GEO}
0	@ment @ment 48.0	@ment @ment 61.0
1	graphic designer 29.0	newline newline 30.0
2	personal trainer 22.0	@ment @ment @ment 25.0
3	@ment @ment @ment 17.0	personal trainer 13.0
4	newline newline 14.0	@ment heavyblackheart 11.0
5	views expressed 13.0	instagram @ment 9.0
6	manager @ment 12.0	@ment @ment @ment @ment 9.0
7	director @ment 11.0	living life 7.0
8	newline hyperlink 11.0	year old 7.0
9	@ment views 10.0	graphic designer 7.0
10	@ment newline 8.0	newline newline newline 6.0
11	web developer 8.0	craft beer 6.0
12	fan views 7.0	owner @ment 6.0
13	personal capacity 7.0	newline instagram 5.0
14	software engineer 7.0	peanut butter 5.0
15	head chef 7.0	newline @ment 5.0
16	husband father 7.0	manager @ment 5.0
17	@ment @ment @ment @ment 7.0	animal lover 5.0
18	hyperlink newline 7.0	social media 5.0
19	makeup artist 6.0	hyperlink hyperlink 5.0
20	level 3 6.0	heavyblackheart @ment 5.0
21	mental health 6.0	heavyblackheart @ment heavyblackheart 5.0
22	social media 6.0	hyperlink newline 5.0
23	newline hyperlink newline 6.0	season ticket holder 4.0
24	beauty therapist 6.0	lake district 4.0
25	freelance graphic 6.0	proud father 4.0

Table 7.10: Popular terms (2-4 grams) in the description fields of U_{SOC} and U_{GEO} profiles.

7.3.5 Highly ranked features

Tables 7.11 and 7.12 show the highest ranked terms learned by the classifier trained on each dataset for the binary task. As in the analysis of description field terms (discussed in Section 7.3.4), U_{SOC} again seems to contain a number of professional related terms such as 'construction' and 'architecture' for the upper class and 'design', 'chef' and 'photographer' for lower. The classifier trained on U_{GEO} on the other hand has obviously identified more geographic features as being most discriminating, with the upper class having many terms obviously related to London, and the lower class to less affluent cities. Tables 7.13 and 7.14 show the highest ranked terms learned for the three-way task. Again, 'professional' terms dominate U_{SOC}, and more 'geographic' terms are learned for U_{GEO}.

7.3.6 Weaknesses of U_{GEO} and U_{SOC}

Although similar results for both datasets were achieved in isolation, compatibility between them was lacking, and we find that both datasets have their own sets of weaknesses. The labels in U_{SOC}, while being of high certainty, are scarce and artificially select for certain types of Twitter profile. Geographically derived labels on the other hand are easy to acquire in bulk, but have a greater chance of being inaccurate due to geographic uncertainty, and the fact that the label is derived from a percentage. Despite some simple quality control checks, the results of models trained on U_{GEO} presented so far were derived from a random sample across the whole set. It is likely that by identifying subsets of the data where we are more certain profiles are labeled correctly, we can further improve the accuracy of models trained on U_{GEO}. In Section 7.5 we investigate this further, by filtering the dataset to only include those profiles with more granular home location estimates. Ultimately, neither approach is likely to be perfect, and therefore some combination of the two may be the best way to utilise the strengths of both; we attempt this in Section 7.4.

Upper	Lower
hai	av
#love	jersey
#avfc	design
#safc	chef
construction	#greece
rt quotetoken	spa
_pile_of_poo	devils
burton	care
article	legal aid
_tennis_racquet_and_ball	chefs
arsenal	carlisle
#f1	designers
data	_heavy_black_heart
_police_cars_revolving_light	— mentiontoken
milan	musical
108	_grinning_face_with_sweat x 2
brexit	photographer
architecture	mentiontoken via
science	#eurovision
albion	gaga
tesco	ey
text	greek
fostering	bu
so so	vat
report	#fitfam
ipswich	holly
reading	uber
god	grass
writing	trailer
_smiling_face_with_heart-shaped_eyes	sheffield
review	shooting
vegan	creative
rt :	futures
#nffc	rt if
Tweetboundary mentiontoken	guitar
#iran	#bcfc
diy	https ...
att	the inspirational
michael	#indyref
belfast	#rt

Table 7.11: Highly ranked terms for the binary model trained on U_{SOC}.

Upper	Lower
london	liverpool
#london	devon
in london	#cornwall
fulham	cornwall
_emoji_modifier_fitzpatrick_type-5	newcastle
cambridge	#liverpool
#fashion	photography
reading	ios
you are	#essex
scarf	blackpool
mentiontoken for	north
party	): hyperlinktoken
wine	hull
london's	gay
jewellery	paul
mentiontoken -	Tweetboundary how
_sports_medal	of our
chelsea	festival
oxford	bay
di	#gym
_sparkles	draw
quotetoken Tweetboundary	xxx
tube	Tweetboundary exactly
listen	medieval
tickets to	days
rt -	store
de	done
... newlinetoken	salford
soho	cornish
dope	tour
article	dancing
_pile_of_poo	#fitfam Figure
greenwich	cumbria
brunch	welsh
battersea	wedding
cheshire	Tweetboundary hyperlinktoken
she	manchester
sugar	wales
ii	well done
boris	#kent

Table 7.12: Highly ranked terms for the binary model trained on U_{GEO}.

Upper	Middle	Lower
#iran	#nufc	spa
rt quotetoken	gracias	devils
Tweetboundary mentiontoken	av	care
#safc	design	_grinning_face_with_sweat
science	hartlepool	barnsley
data	#greece	#amwriting
ipswich	_raising_hands	jersey
#lcfc	holly	guitar
_pile_of_poo	_fish	#bcfc
#avfc	#ttot	_snowman_without_snow
construction	_bear_face	barber
burton	mentiontoken via	unfollowers
_person_with_folded_hands	designer	chefs
brexit	legal aid	cunt
maths	#foodporn	la
text	#indyref	#fitfam
so so	#cpfc	#yoga
att	photographer	rt if
allah	pizza	yoga
vegan	#thewalkingdead	_grinning_face_with_sweat
fostering	toronto	girls
review	photography	_shadowed_white_star
_tennis_racquet_and_ball	— mentiontoken	bought :
photography by	photos	golf
#love	restaurant	craft

Table 7.13: Highly ranked terms for the three class model trained on U_{SOC}.

Upper	Middle	Lower
london	#cornwall	liverpool
#london	essex	#liverpool
in london	#essex	blackpool
_emoji_modifier_fitzpatrick_type-5	devon	northampton
jewellery	farm	derby
cambridge	brighton	swindon
di	cornwall	ir
hockey	bath	_beer_mug
fashion	surf	_heart
london's	#cpfc	wwe
de	weekend	hackney
#fashion	farmers	lincoln
greenwich	west ham	newlinetoken free
oxford	libraries	newcastle
london hyperlinktoken	marketing	oldham
donald	golf	midlands
richmond	http :/	_american_football
circle_symbol	west	#cbb
#fitness	website	boro
party	charlton	in liverpool
reading	sussex	academy
#gameofthrones	josh	#ynwa
mentiontoken via	british	#mcfc
mentiontoken for	river	store
photos	ar	#newcastle

Table 7.14: Highly ranked terms for the three class model trained on U_{GEO}.

7.4 Combining labels from disparate annotation schemes

It is possible that the distantly supervised dataset U_{GEO} can be used in conjunction with the directly supervised dataset U_{SOC} to improve performance over models trained on either. In this section we combine classifiers trained on U_{GEO} and U_{SOC} in an attempt to improve predictive performance.

7.4.1 Ensemble of classifiers

Our chosen ensemble consists of two probabilistic classifiers, the LR models trained in Section 7.3, which predict the probability distribution for each label for a given input; one classifier is trained on U_{SOC}, the other on a subset of U_{GEO} with 4500 examples per label. Figure 7.1 illustrates the process of our ensemble; a user is passed through a feature extraction module and passed into the two classifiers, the output distribution from each classifier is multiplied by a set weight ($W(U_{GEO})$ and $W(U_{SOC})$, derived through grid search) and the two are added together with the maximum taken as the label.

To derive weights for each classifier a grid search was performed between the values of 0.5 and 1.5 for each classifier. We found the optimum weights to be $W(U_{GEO}) = 0.5$ and $W(U_{SOC}) = 1.35$ in both the binary and 3-way problem configurations.

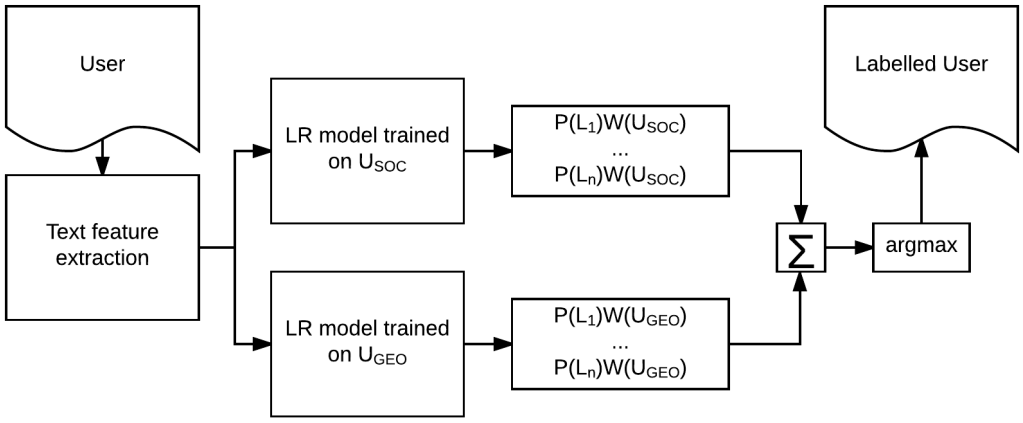

Figure 7.1: Ensemble classification pipeline.

7.4.2 Results and discussion

Table 7.15 shows classification accuracy at predicting U_{SOC} with and without the ensemble. In both the 2-class and 3-class configurations an accuracy boost is observed utilising our ensemble approach. Ultimately, by combining the two classifiers in this ensemble approach we see only a minor increase in predictive performance, with gains of 0.64% and 1.54% for the two- and three-class problems respectively.

Configuration	No ensemble	Ensemble	Gain
2-way	0.7434	0.7498	0.64%
3-way	0.6429	0.6582	1.54%

Table 7.15: Mean U_{SOC} prediction accuracy with and without the ensemble of classifiers.

We also attempted augmenting U_{SOC} by shuffling in profiles from U_{GEO} but found that overall performance decreased. This, in conjunction with the low weight assigned to $W(U_{GEO})$ indicates that the true problems with U_{GEO} lie in the correctness of the labels themselves. In Section 7.5, we attempt to alleviate the affect of these mis-labellings by applying additional filtering steps to U_{GEO}, such that only those labels we are most certain about are used to train models.

7.5 Filtering for profiles with high home location certainty

Numerous findings throughout this work have pointed towards poor judgements of home location being a likely source of error in the initial labelling of the geographically derived training data in U_{GEO}, which results in poor quality labels in derived models. In Chapter 5 we highlighted that the clustering approach used to determine home location was around 80% accurate at the Output Area level. Alongside predicting home location, we also measure home location judgement 'certainty' as the average distance of each point in a users' home cluster, to their home cluster centroid. We also found in Chapter 5 that a correct estimate was much more likely when the the home cluster was denser, or more 'certain'. Currently, our experiments using data derived through our geographic user profiling

dataset generation approach have not taken home location certainty into account, having simply selected profiles annotated through our method at random.

To asses the impact of filtering for profiles with high home location certainty on classifier performance, we took the 4500 most certain profiles for each label to construct a subset of high certainty profiles, as well as a subset of profiles of the same size selected at random. For both subsets, we trained the predictive approach based on TF-IDF n-grams and LRclassifiers described in Section 7.3 and derived average performance metrics using 10-fold cross validation.

7.5.1 Results and Discussion

Table 7.16 presents classifier performance in terms of accuracy when the predictive models were trained on the subset of U_{GEO} where the user's home cluster densities were lowest as well as the random subset, highlighting the gain in classification accuracy for the filtered subset. Table 7.17 additionally presents precision, recall, and F1-score at class level for the models trained on the filtered subset of U_{GEO}.

Introduction of the additional filtering step leads to a clear gain in classification accuracy in the both the 2- and 3-class variants. This implies that many of the labels used to train previous models were indeed likely to be mislabelled, for both NS-SEC and OAC/LAC (Chapter 6). This filtering step is essential in any future application of our geo-location driven user profiling dataset generation approach.

Configuration	**Random Sample**	**Geographic Filtered**	**Gain**
2-way	0.73	0.78	0.05
3-way	0.56	0.60	0.04

Table 7.16: U_{GEO} accuracy improvement by filtering for profiles with denser home clusters.

7.6 Conclusion

In this chapter we expanded and performed the final evaluation of our method for deriving user profiling training data by combing geo-located profiles and demographic data. We applied our method to generate a large, new, dataset (U_{GEO}) for

Prediction	Accuracy	Precision	Recall	F1-score
High	-	0.69	0.67	0.68
Middle	-	0.55	0.50	0.52
Low	-	0.56	0.63	0.60
Average	0.60	0.60	0.60	0.60
High	-	0.77	0.79	0.78
Low	-	0.79	0.76	0.78
Average	0.78	0.78	0.78	0.78

Table 7.17: Classifier performance metrics for U_{GEO} filtered to contain profiles with highly granular location estimates, derived using 10-fold cross validation

a SES variable that has previously been addressed in the literature, the NS-SEC. To meaningfully assess performance of datasets derived through another state-of-the-art method against our own, we additionally gathered the NS-SEC tagged user profiling dataset described in Lampos et al. [134] (U_{SOC}).

We trained a robust user profiling pipeline on both datasets, which achieved good and similar performance metrics in both cases. Given promising performance in isolation, we evaluated the ability of models trained on U_{GEO} to predict the labels in U_{SOC}, and noticed a drop in predictive performance, while still exceeding a random baseline. Given the poor compatibility between the resulting models and opposing datasets, we set about investigating potential explanations for the incompatibility, identified several, and evaluated several approaches for addressing them.

We performed an analysis to investigate the differences in how users in each dataset present themselves in their 'description' field; for the users in U_{SOC}, professional terms were more frequent, and in U_{GEO}, terms relating to geography rose to the top; both observations indicate characteristics of the data generation approach biasing the resulting datasets.

Given the differences in the sorts of user present in both datasets due to differences in labelling approach, we investigated potential methods for combining the two schemes to augment each other. We found that simple combination of the datasets yielded poor performance directly, so instead sought to combine the derived models instead, in an ensemble. Our ensemble approach yielded a modest increase in predictive performance, indicating that the two labelling approaches can indeed

be combined to complement each other.

Having noted that some steps in our method for geo-location driven user profiling dataset generation have the potential to introduce incorrect labels, we investigated what we deemed to be the largest source of error, an incorrectly assigned home location. Our method for assigning home location incorporates a measure of 'certainty', which has up until now not been explicitly constrained upon in our dataset generation method. To investigate the affect of low certainty home locations, we filtered U_{GEO} to contain only those profiles with high certainty labels, while still maintaining dataset balance and per-label examples. Adding this filtration step, yielded a marked boost in predictive performance, and therefore should be included in any further implementations of our for geo-location driven user attribute labelling.

Chapter 8

Discussion and conclusion

Several downsides to current approaches for user attribute labelling in user profiling dataset creation were identified in Chapter 1. To overcome some of these weaknesses, a geo-location driven approach to user attribute labelling was proposed in Section 1.3. Geo-location driven attribute labelling involves identifying the 'home location' of a social media user, and linking them to demographic variables associated with their local area, such as those present in government census data or commercial geo-demographic segmentation datasets. The assumed benefit of this approach is that it can generate large datasets for training downstream user profiling systems, and assign attributes that were not previously feasible to acquire.

In this chapter the findings presented throughout this thesis are discussed with reference to the thesis aims initially presented in Section 1.3. The aims are repeated here, with reference to the discussion included in this chapter:

- Establishing good baseline implementations for user profiling predictive systems based on datasets generated through 'classic' approaches. Discussed in Section 8.1;
- Establishing a strong grounding for our proposed method by evaluating whether simple high level datasets derived from geo-located Tweets can be used to improve performance on a user profiling task. Discussed in Section 8.2;
- Evaluating methods for estimating user home location, determining whether

the state-of-the-art is good enough for use in our proposed method, and improving on the state-of-the-art if necessary. Discussed in Section 8.3;

- Generating novel datasets for user profiling using our proposed method, and evaluating them by developing user profiling predictive systems incorporating established strong baseline approaches. Discussed in Section 8.4; and
- Supplementing and contrasting existing user profiling datasets with ones derived by our methods. Discussed in Section 8.4.1.

For each of the aims above, further works are discussed in the appropriate section listings.

User profiling is by its nature a field that gives rise to ethical concerns. Throughout this thesis we have not commented on the ethical implications of the tools and techniques presented, focussing instead on technical details and empirical results. We discuss the ethical implications of this work in Section 8.6.

8.1 User profiling systems

The geo-location driven user attribute labelling approach proposed in Chapter 1 is not suitable as a user profiling system in its own right, as only a limited proportion of users choose to geo-locate their posts. The main utility of the method is in its ability to generate large datasets annotated with user attributes, which can be used to train user profiling systems. Before applying the method to generate datasets and train user profiling systems, it is important to ensure that the components underpinning the user profiling systems are robust and in line with best practises. A review of the techniques and tools used in the literature to build user profiling systems was performed (Chapter 2), and the typical process undertaken to build a user profiling system was illustrated in Figure 2.1. Broadly, user content and user attributes are used to fit some machine learning model, which is then used to infer the attributes of new users. Linear models, such as logistic regression (LR) or support vector machine (SVM), trained on n-gram feature vectors extracted from user content are used in many successful approaches in both user profiling [13, 58, 64], as well as other fields [161, 162]. As such these components form the basis of the user profiling systems trained throughout this thesis.

To verify the utility of the chosen user profiling system components outside of the context of our geo-location driven user attribute labelling approach, experiments were performed on user profiling datasets derived through 'classic' methods in Chapters 3 and 4. In Chapter 3 SVM classifiers and regressors were trained on n-gram features to predict age, gender and personality. Topic models were experimented with as an additional feature, which provided a minor boost to predictive performance. In Chapter 4, an ensemble composed of a Gaussian process (GP) classifier trained bag-of-word-embedding cluster feature vectors and a logistic regression classifier trained on term frequency-inverse document frequency (TF-IDF) transformed unigram and bigram feature vectors was trained to predict gender and native language variant. The ensemble performed well, and when compared to a strong baseline established from the classification approach in Chapter 3, yielded performance increases for a number of attributes, and performed on-par for others. From the results in these two chapters it was concluded that the proposed baseline user profiling system was indeed suitable for use in experiments on our own novel datasets.

Given the good performance of the systems used in Chapters 3 and 4, similar systems were implemented for the experiments on datasets derived through the proposed geo-location driven user attribute labelling approach. In Chapter 6, user profiling systems incorporating SVM classifiers trained on TF-IDF weighted n-grams were trained to predict two measure of socio-economic status, Output Area Classification (OAC) and Local Authority Classification (LAC). Both systems exceeded a majority baseline, and good results were achieved for LAC. The system trained on OAC performed poorly, although this was found to be a failing of the underlying dataset, not of the user profiling system itself. In Chapter 7, LR classifiers are trained on TF-IDF weighted n-grams for two datasets labelled with a measure of socio-economic status, National Statistics Socio-economic Classification (NS-SEC), derived through different methods. Good performance is achieved by the models trained on both dataset in isolation, but the resulting models are unable to predict as accurately on the dataset derived differently. To overcome this weakness, the resulting models are employed in an ensemble approach to combine their outputs, resulting in improved performance.

The linear model and n-gram components used in the user profiling systems trained in Chapters 6 and 7 performed well, and their simplicity and interpretability allow us to conclude that datasets derived through the proposed geo-location driven user

attribute labelling approach can indeed be used to train informative user profiling models. We chose not to include additional feature sets, such as the topic models or word embedding model clusters (from Chapters 3 and 4), or more modern machine learning approaches such as deep learning in our classification pipelines as our main aim was to show that geo-location derived user profiling datasets could be used to train user profiling systems; the specific components of the user profiling systems were of less interest than achieving consistent, interpretable results when applied to a range of datasets.

Having achieved strong baseline results for user profiling systems trained on datasets derived through the proposed geo-location driven user attribute labelling approach, more recent advances in machine learning could be investigated to train user profiling systems with improved performance. Deep learning methods in particular have advanced in recent years and are now generally seen as the most performant (and therefore default) option for a wide range of machine learning tasks. In future work, deep learning methods for text classification (as discussed in Section 2.5) would be investigated to further improve the predictive performance of developed user profiling systems.

8.2 Establishing utility of geographically derived resources

The proposed geo-location driven user attribute labelling approach relies on the assumption that social media posts (and the users who post them) are representative of the local area in which they are posted. Before jumping in to deriving user profiling datasets directly from geo-located users, in Chapter 4, an experiment was first performed to assess whether a *complementary* user profiling resource could be derived from a collection of geo-located tweets to improve performance of a user profiling system trained on a dataset derived through a 'classic' approach (the PAN 2017 Author Profiling shared task dataset [125, 126, 127]). A large corpus of worldwide Tweets was collected and filtered down to those geo-located within the countries covered in the task's languages, then divided into individual languages (Portuguese, English and Spanish), and further filtered to ensure adequate representation for each country that speaks a language natively. These corpora were used to train Word2Vec word embeddings [34, 128] tailored to each language as it

appears on Twitter.

We assessed the utility of the derived word embedding models by evaluating their effect when included as part of a ensemble classification pipeline in an experiment on the PAN 2017 Author Profiling dataset. Our ensemble approach performed well, and inclusion of the complementary resources improved performance over a strong baseline (SVM and word n-grams) for some user attributes. These results showed that geographically derived resources are indeed useful for user profiling, adding support for the proposed geo-location driven user attribute labelling approach.

Generating large scale datasets in this fashion is of great value in generating language resources representative of how language is used in the real world for use in applications such as language model pre-training [168] although gathering data in this fashion does come with weaknesses that must be accounted for. For example, in Chapter 4, Tweets were collected in the Arabic language, but inspection by native speakers highlighted quality issues with the data, including large quantities of spam posts and Quran quotes. Any future experiments deriving resources from geo-located posts should ensure steps are implemented to ensure such unrepresentative examples of the language are handled. Filtering these posts is likely to require language-specific steps, which reduces the utility of this method for *fully automatic* generation of language specific resources, although recent changes to the Twitter API promise improved spam Tweet detection, which may reduce the burden to perform this step for researchers.

8.3 Accurate home location estimation

The proposed an approach for geo-location driven user attribute labelling involves linking social media users to geo-demographic datasets, such as aggregated census data. This approach requires a reliable method for determining a user's 'home location' at a highly granular, hyperlocal level to accurately link users to their local demographics. In addition, a suitable home location allocation method should incorporate some measure of 'uncertainty', so that users with sparse or highly spread activity can be excluded from derived user profiling datasets.

A suitable method for hyperlocal user home location allocations must perform

accurate coordinate-level predictions or arbitrarily small regions from a list of candidates. State-of-the-art approaches used to assign 'home location' to social media profiles were reviewed in Section 2.4, and four state-of-the-art methods were identified:

- 'first Tweet' involves simply taking the the coordinates of a user's first Tweet as their home location;
- 'location field' converts user-declared location fields into a home location;
- 'grid based' partitions areas into an arbitrary grid and counts the number of Tweets geo-located within each cell; and
- 'geometric median' simply takes the geometric median the coordinates of a user's geo-located tweets.

None of the identified state-of-the-art approaches quite satisfy the desired characteristics of a robust and accurate method for identifying hyperlocal home location identification, so two novel methods for identifying user home location were also proposed. The novel methods acknowledge that social media users do not (in general) post from a single location, and that the most active location is likely to be their primary location. Majority voting measures a user's activity in real world regions as defined by public boundary datasets. Clustering involves applying some clustering algorithm to each user's geo-located Tweets, identifying distinct locations of activity.

To evaluate the existing and proposed methods for home location allocation, a new gold-standard dataset containing users annotated with home location at the hyperlocal (coordinate) level was created. A set of phrases indicating which Tweets were sent from a user's home was curated and used to identify the home location of 1,042 Twitter users. Tweets containing a home-indicative phrase were manually verified for context, and discarded if incorrect or uncertain. True home location was calculated for each user profile as the spatial average of its home-indicative Tweets. Each of the methods mentioned previously were implemented and applied to the gold-standard home location dataset, and predicted locations were compared against the 'true' values. Error distance (in miles) and exact match accuracy (for four types of region) are reported, and the results show the clustering and majority vote approaches outperform the state-of-the-art methods. It is concluded that the clustering method satisfies our requirement of a good method for

assigning home location to Twitter profiles, thanks to good performance at both coordinate and region-level granularity. The clustering method was therefore used in our experiments applying the proposed geo-location driven user attribute labelling approach from Chapter 1.

Introduction of the notion of 'active places' to home location identification yields marked improvements over other approaches that do not utilise them, although our current approach makes the potentially naive assumption that a user's most posted-from location is their home. Clustering for home location identification performs well in the context of user attribute labelling, but is not without errors. It is highly likely that some users post more commonly from an alternate location such as their place of work, study or the homes of friends/family, leading to incorrect location predictions under the current scheme. In future work, the clustering and majority vote methods could be expanded in future work to characterise user's active locations beyond 'home'. Additional types of location could be classified from the candidate places identified by either method in a number of ways. Local land-use data could be interrogated to identify the primary use of a given place, if an area is mostly offices or industrial, it is unlikely to be a user's home. Textual cues and meta-data in Tweets in could be used to train machine learning approaches capable of characterising the identified places given an annotation scheme. To differentiate 'home' and 'work' locations specifically, the proportion of Tweets at each location in and out of canonical work hours could be investigated (building on the work in Cho et al. [155]).

8.4 Novel geographically derived datasets

The proposed method for geo-location driven user attribute labelling was implemented and applied in Chapters 6 and 7 to generate three user profiling datasets covering OAC, LAC, and NS-SEC.

The OAC is a hierarchical classification of socio-economic groups aggregated to the Output Areas (OAs) level in the UK derived from census data. The LAC is a similar scheme, that is applied at the Local Authority District (LAD) level. In Chapter 6, geo-location driven user attribute labelling was performed to create two large user profiling datasets labelled with OAC and LAC; prior to this work, neither attribute had been explored in the user profiling literature. User profil-

ing systems were trained on balanced subsets of both datasets, exceeding random baselines, with poor performance for OAC, but promising results for LAC. Underlying properties of the two datasets were investigated to determine potential explanations for the difference in results.

A likely explanation for the poorer performance of OAC is the geographic properties of the underlying region type (OA); OAs are on average less than a mile long, whereas Local Authoritys (LAs) are much larger at around 35 miles, and as such, small errors in home location prediction are more likely to lead to an incorrect judgement for OAs than LADs. This is illustrated further in Table 5.4; for our chosen home location prediction method (clustering with k-means), 79.85% of prediction were correct at the OA level, versus 97.83% for LADs. A lack of certainty at the home location prediction level feeds through the entire geo-location driven user attribute labelling approach; an incorrect home location prediction leads to a potentially incorrect user attribute label, resulting in noisy data being used to train user profiling systems. Therefore it is essential to minimise the potential for incorrect labels to be used at train time. In Section 7.5 we experimented with a measure of home location judgement certainty, 'home cluster density', and found that by filtering for profiles with highly certain (more densely clustered) home location judgements, resulting user profiling systems were much more accurate in their predictions.

Examining the feature coefficients of the models trained OAC and LAC shed light on the underlying motivations behind each demographic dataset. For LAC, derived models deemed geographic terms most discrimination, with dialect features and local landmarks surfacing as important. For OAC on the other hand, terms stereotypically associated with certain groups, with the rural class referencing country life, and the multicultural group referencing Islamic holidays; only one group, 'Ethnicity Central', was dominated by locale (London) specific features. In future work, steps should be taken to ensure that derived datasets are geographically stratified across the regions of interest, to avoid over-representation of any one specific locale.

Additional analysis indicated that the underlying demographic datasets held characteristics with the potential to make classification difficult. Three of the supergroups in OAC, 2 (Cosmopolitans), 3 (Ethnicity Central), and 4 (Multicultural Metropolitans), refer to users residing in larger cities where mixing of demographic groups is more common; these groups are therefore more similar to each other in

terms of language than other classes, making classification by a machine learning approach difficult. Similar patterns can be seen for other groups as well; 7 (Constrained City Dwellers) and 8 (Hard-Pressed Living) both refer to users in smaller cities and towns, and 1 (Rural Residents) and 6 (Suburbanites) typically border each other. This problem is less pronounced for LAC, although supergroups 3 (London Cosmopolitan), 4 (Suburban Traits), and 7 (Prosperous England) do broadly cover the 'middle class'. In future work, care should be taken to understand such patterns in the underlying data prior to application of the geo-location driven user attribute labelling approach. If a number of classes are indeed functionality equivalent in terms of their members, they could be collapsed into a single class, or a demographic dataset could be discounted if it does not adequately capture individual differences.

8.4.1 Comparability to other methods

Geo-location driven user attribute labelling when applied to the OAC and LAC yielded promising results, but with several weaknesses, most notably the relatively poor performance for OAC and the lean towards geographic cues for models trained on LAC. As both of these datasets cover variables not previously covered in the literature, it was not possible to compare against datasets annotated with the same variables through different means. To enable comparison of geographically derived user profiling datasets against datasets derived for the same attributes via a different method, an existing dataset (referred to as U_{SOC}) from Lampos et al. [134] was acquired covering another measure of socio-economic status (SES), NS-SEC. In parallel, a large novel user profiling dataset (U_{GEO}) annotated with NS-SEC was constructed using geo-location driven user attribute labelling.

Models trained on the two datasets performed similarly well in isolation, but when tested on users labelled using the opposing method performance dropped noticeably. This indicated that the two datasets either contain different 'types' of user, or that the assigned labels are not actually capturing the same information. Analysis of the description fields of the users from both datasets highlighted clear differences in the types of user present in each dataset, with U_{SOC} users skewed towards declaration of their profession and statements that they are not posting behalf of their employer, a clear indication of a person in a 'professional' role.

Both user attribute labelling schemes ostensibly allocate the same labels to users,

but show clear differences in the sorts of user they contain. In Section 7.4 models derived from both datasets were used in an ensemble approach to assess whether or not the two datasets could be used to complement each other, thus overcoming their differences and individual weaknesses. The ensemble approach yielded minor improvements in performance over either of its component classifiers in isolation, highlighting that the two sets of labels derived from disparate annotation schemes can indeed be combined. The ensemble was applied using models trained on U_{GEO} *without* filtration for 'more certain' users (dense home clusters); in future work, we would re-evaluate the ensemble approach on a set of users filtered to only contain those most likely to hold correct home location judgements, as per Section 7.5.

8.5 Applicability to transfer learning

User profiling systems typically rely on large amounts of accurately annotated user profiles as training data, which is expensive and difficult to acquire using classic methods of annotation. Geo-location driven user attribute labelling, proposed in Chapter 1 and evaluated throughout this thesis, is capable of generating large annotated user profiling datasets, but suffers from the potential of incorrect labels introduced through the distantly supervised annotation process. Transfer learning (TL) is a machine learning (ML) approach that attempts to 'transfer' the knowledge learned by a ML model from one set of data to another [169]. The word embeddings trained in Chapter 4 are a basic example of the concept of TL; a single representation is learning for each term in the vocabulary, which are then used to perform classification in a down-stream task.

In future work, geo-location driven user attribute labelling could be used to generate large-scale corpora for pre-training user profiling models for TL, which can then be fine-tuned on smaller-scale hand labelled user profiling training data. TL is typically performed using neural networks, rather than the 'classic' machine learning approaches applied throughout this thesis. The process for pre-training is as follows; a neural network model m_p, is trained on the distantly supervised dataset D. For fine-tuning, a second neural network, m_f is initialised with the same parameters learned by m_p, and trained on the smaller, hand labelled, dataset H. The main advantages of models trained in this way over standard ML are the reduction in training time (may only need to fine-tune an existing model), improved model performance and generalisation (due to picking up cues from a wider set of

data), and a reduction in annotator effort for the same level of performance.

8.6 Ethical considerations

The works performed in thesis were subject to ethical approval from the University of Sheffield Institutional Review Board (IRB). Data collection efforts respected privacy settings by only accessing public tweets and profiles. The Text and Data Mining (TDM) exception to UK copyright law, which allows non-commercial data-mining of public or legally accessible resources, ensures that this data collection did not breach UK copyright law (at the time of data collection).

Steps were taken to ensure no more potential for harm than is found in everyday life. Unlike a lot of research on profiling the characteristics of Twitter users, the work in this thesis could be seen as low risk, as only aggregated data is used, inferred from public disclosure. While potentially sensitive topics such as income, politics and education are considered it must be stressed that they are being studied on a non-individual level and only relate to data available from public sources. Personal details of gathered users were anonymised as far as possible by obfuscating identifying profile information, including names, screen names, and mentions of other users. Disseminated results and statistics are not linked to individual users, and the developed datasets were not available to the public in their raw form.

The geo-location driven user attribute labelling approach implemented and evaluated through this thesis utilises cues present in user profiles and their output that when combined disclose the presence of a personal attribute. Social media is now pervasive across everyday life, and works such as this are valuable in highlighting that the sorts of information we as individuals make available publicly on the internet may reveal more about ourselves than we think.

Knowledge of a user's location can be used for a range of purposes including those with social benefit, such as to aid disaster response [170], track disease outbreaks, reduce bias in demographic analysis by balancing user characteristics [73, 171, 156] and perform geographically linked opinion analysis [172]. However, there is also potential for misuse. When processing the data for the work described here we took care to ensure that geo-location data was processed separately from other

information about the user and, as an extra precaution, we have not made the data publicly available. The findings in Chapter 5 highlight that accurate estimates of some Twitter user's home location can be made using geo-location history and publicly available tools. In addition, we found that some Twitter users state they are 'at home' in geo-located Tweets, implicitly revealing their residential location. This work further highlights the need for users of social media platforms such as Twitter to be aware of the implications of sharing the information they make available in their posts. In particular, when choosing privacy settings or attaching geo-location information to their posts, users should be aware that they may be revealing their location to a variety of actors and that inferences can be drawn from this information as it builds up over time.

8.7 Conclusion

Throughout this thesis we developed the foundations of, and explored our approach for geo-location driven user attribute labelling, showing that user profiling systems can be trained using only geographically derived training data.

Strong baseline components for user profiling systems were explored and validated on existing user profiling datasets (Chapters 3 and 4), allowing us to reliably conclude that reported performance metrics (good or bad) are the result of our novel annotation process, not artefacts of the components underpinning the user profiling system. A complementary language resource was derived using geo-located posts and used to improve performance in a user profiling task (Chapter 4), showing the potential of utilising geo-located posts for user profiling.

Two novel methods for assigning home location to Twitter profiles based on clustering and majority voting across real-world regions were proposed an evaluated (Chapter 5), beating state-of-the-art methods used to generate datasets for user location prediction tasks. Our evaluation of these methods had not been performed previously in the literature at such a fine level, and revealed that many existing approach for 'ground truth' user location are inaccurate.

Three novel datasets were developed using the proposed geo-location driven user attribute labelling approach covering three different measures of SES: OAC, LAC, and NS-SEC (Chapters 6 and 7). User profiling systems trained on LAC, and NS-

SEC performed well in isolation, but systems trained on OAC performed poorly in comparison. Error analysis, comparison against other datasets, and inspection of the underlying demographic datasets revealed several considerations for the application of geo-location driven user attribute labelling approach going forward.

The most clear source of poor or deteriorating model performance was down to errors in the labelling process, specifically incorrect judgements of user home location; efforts must be made to maximise the accuracy of this process in any future application, for example, by filtering only to those profiles with the most 'certain' labels (as per Section 7.5). Properties of the underlying demographic datasets themselves must also be considered: OAC contains several classes that overlap in the type of individual they contain, making classification more difficult; LAC is composed of classes more representative of geographic regions than the users within them, and so models pick geographic features rather than ones indicative of personal characteristics; and NS-SEC is treated as a continuous value, a proportion per class, leading to inherently noisy labels.

Geo-location driven user attribute labelling does yield usable user profiling systems in isolation, but is perhaps best suited as complementary to classic user attribute labelling approaches, for example in ensemble approaches with models trained on less noisy data (as per Section 7.4). In addition, the ability to generate large (albeit weakly labelled) datasets makes geo-location driven user attribute labelling potentially well suited to pre-training deep learning models for transfer learning in user profiling systems.

Bibliography

[1] Danny Poo, Brian Chng, and Jie-Mein Goh. A hybrid approach for user profiling. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, pages 9–pp. IEEE, 2003.

[2] Michael Trusov, Liye Ma, and Zainab Jamal. Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, 35(3):405–426, 2016.

[3] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35–47, 2013.

[4] Claudia Peersman, W Daelemans, and L Van Vaerenbergh. Predicting age and gender in online social networks. pages 37–44, 2011.

[5] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural*, pages 1348–1353. Association for Computational Linguistics, 2013.

[6] Martin Obschonka and Christian Fisch. Entrepreneurial personalities in political leadership. *Small Business Economics*, 50(4):851–869, 2018.

[7] Lindsay Young, Daniel C Kolubinski, and Daniel Frings. Attachment style moderates the relationship between social media use and user mental health and wellbeing. *Heliyon*, 6(6):e04056, 2020.

[8] Jonathan Schler, M Koppel, S Argamon, and J Pennebaker. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs: Papers from the AAAI Spring Symposium*, pages 199–205, 2006.

[9] M. Koppel. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002. ISSN 0268-1145. doi: 10.1093/llc/17.4.401.

[10] Olivier Y de Vel, Malcolm W Corney, Alison M Anderson, and George M Mohay. Language and gender author cohort analysis of e-mail for computer forensics. 2002.

[11] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. pages 149–156, 2011.

[12] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, page 253, 2011. doi: 10.1145/1979742.1979614.

[13] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. pages 37–44, 2010.

[14] Delip Rao and David Yarowsky. Detecting latent user properties in social media. *Proc. of the NIPS MLSN Workshop*, pages 1–7, 2010.

[15] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110 (15):5802–5, 2013. ISSN 1091-6490. doi: 10.1073/pnas.1218772110.

[16] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. Lexical Predictors of Personality Type. 2005.

[17] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2:386–393, 2012. doi: 10.1109/ICMLA.2012.218.

[18] D Nguyen, D Trieschnigg, and T Meder. TweetGenie: Development, Evaluation, and Lessons Learned. *Proceedings of COLING*, 2(1):62–66, 2014.

[19] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. pages 1301–1309, 2011.

[20] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717, 2015.

[21] Office for National Statistics. 2011 census: Aggregate data. 2011.

[22] US Census Bureau. Us census. URL `https://www.census.gov/data.html`.

[23] Umashanthi Pavalanathan and Jacob Eisenstein. Confounds and Consequences in Geotagged Twitter Data. (September):2138–2148, 2015.

[24] Christopher G Gale, Alexander D Singleton, Andrew G Bates, and Paul A Longley. Creating the 2011 area classification for output areas (2011 oac). *Journal of Spatial Information Science*, 2016(12):1–27, 2016.

[25] UK Office for National Statistics. 2011 area classification for local authorities. URL `https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/datasets`.

[26] Rupert A Payne and Gary A Abel. Uk indices of multiple deprivation—a way to make comparisons across constituent countries easier. *Health Stat Q*, 53(22):2015–2016, 2012.

[27] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and Latent Attribute Inference : Inferring Latent Attributes of Twitter Users from Neighbors. pages 387–390, 2012.

[28] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[29] J W Pennebaker and & Booth R J Francis ME. *Linguistic Inquiry and Word Count: A computerized text analysis program.* 2001.

[30] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. pages 624–628, 2005.

[31] James RA Davenport and Robert DeLine. The readability of tweets and their geographic correlation with education. *arXiv preprint arXiv:1401.6058*, 2014.

[32] Alejandro Llorente, Manuel Cebrian, Esteban Moro, et al. Social media fingerprints of unemployment. *arXiv preprint arXiv:1411.3140*, 2014.

[33] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2012. ISSN 15324435. doi: 10.1162/jmlr.2003.3.4-5.993.

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 1–9, 2013. ISSN 10495258. doi: 10.1162/jmlr.2003.3.4-5.951.

[35] Qv Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196, 2014.

[36] Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through Twitter content. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, 2015.

[37] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[38] a Mulac, Jj Bradac, and P Gibbons. Empirical support for the gender-as-culture hypothesis. *Human Communication Research*, 27(1):121–152, jan 2001. ISSN 03603989. doi: 10.1111/j.1468-2958.2001.tb00778.x.

[39] François Mairesse and Marilyn a. Walker. Automatic recognition of personality in conversation. *Proceedings of the Human Language Technology Conference of the NAACL*, (June):85–88, 2006. ISSN 0749596X. doi: 10.3115/1614049.1614071.

[40] François Mairesse and Marilyn Walker. Words mark the nerds: Computational models of personality recognition through language. pages 543–548, 2006.

[41] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, pages 457–500, 2007.

[42] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. pages 180–185, 2011.

[43] Marco Pennacchiotti and Ana-Maria Popescu. A Machine Learning Approach to Twitter User Classification. In *ICWSM*, pages 281–288, 2011. ISBN 9781450308137.

[44] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015. doi: 10.1073/pnas.1418680112.

[45] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[46] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016. ISSN 1432-5012. doi: 10.1007/s00799-015-0156-0.

[47] Dominique Estival, Tanja Gaustad, and SB Pham. TAT: an author profiling tool with application to Arabic emails. *Proceedings of the ...*, pages 21–30, 2007.

[48] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. Author Profiling for English Emails. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, 2007.

[49] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[50] Mayur Rustagi, R. Rajendra Prasath, Sumit Goswami, and Sudeshna Sarkar. Learning age and gender of blogger from stylistic variation. 5909 LNCS:205–212, 2009. ISSN 03029743. doi: 10.1007/978-3-642-11164-8{_}33.

[51] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. *Proceedings of the 2010 conference on Empirical ...*, (October): 207–217, 2010.

[52] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. *Proceedings of the COLING/ACL on Main ...*, (July):627–634, 2006. doi: 10.1177/0266382105060607.

[53] Francisco Iacobelli, AJ Gill, Scott Nowson, and Jon Oberlander. Large scale personality classification of bloggers. *Affective Computing and ...*, 6975 LNCS(PART 2):568–577, 2011.

[54] Tobias Bocklet, Andreas Maier, and Elmar Nöth. Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression. pages 253–260, 2008.

[55] Zubin Jelveh, Bruce Kogut, and Suresh Naidu. Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics. *anthology.aclweb.org*, (2013):1804–1809, 2014.

[56] Zubin Jelveh, Bruce Kogut, and Suresh Naidu. Political language in economics. *Available at SSRN 2535453*, 2014.

[57] Delip Rao, M. Paul, Clay Fink, D. Yarowsky, Timothy Oates, and G. Coppersmith. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. pages 598–601, 2011.

[58] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 09 2013. doi: 10.1371/journal.pone.0073791.

[59] Jalal S. Alowibdi, Ugo a. Buy, and Philip Yu. Language independent gender classification on Twitter. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, (May):739–743, 2013. doi: 10.1145/2492517.2492632.

[60] Edson R D Weren, Anderson U Kauer, Lucas Mizusaki, Viviane P Moreira, J Palazzo M De Oliveira, and Leandro K Wives. Examining Multiple Features for Author Profiling. 5(3):266–279, 2014.

[61] Theo Meder. Why Gender and Age Prediction from Tweets is Hard : Lessons from a Crowdsourcing Experiment. pages 1950–1961, 2014.

[62] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[63] Brendan O'Connor, Noah a Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. (October):1277–1287, 2010.

[64] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.

[65] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Twitter User Geolocation Using a Unified Text and Network Prediction Model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 630–636, 2015.

[66] Delroy L Paulhus and Kevin M Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563, 2002.

[67] Delroy L Paulhus. Toward a taxonomy of dark personalities. *Current Directions in Psychological Science*, 23(6):421–426, 2014.

[68] Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102, 2014. ISSN 01918869. doi: 10.1016/j.paid.2014.01.016.

[69] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. " How Old Do You Think I Am?" A Study of Language and Age in Twitter. *ICWSM*, pages 439–448, 2013.

[70] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 192–199. IEEE, 2011.

[71] Karolina Sylwester and Matthew Purver. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *Plos One*, pages 1–18, 2015. doi: 10.1371/journal.pone.0137422.

[72] Jiwei Li, Alan Ritter, and Eduard Hovy. Weakly Supervised User Profile Extraction from Twitter. pages 165–174, 2014.

[73] Ehsan Mohammady and Aron Culotta. Using County Demographics to Infer Attributes of Twitter Users. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 7–16, 2014.

[74] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015.

[75] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9(6):811–824, 2012.

[76] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[77] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[78] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[79] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions–a content based approach. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 838–843. IEEE, 2011.

[80] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society, 2012.

[81] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 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*, pages 759–768. ACM, 2010.

[82] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. pages 1277–1287, 2010.

[83] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.

[84] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.

[85] David Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[86] Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

[87] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2015.

[88] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[89] Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367. Association for Computational Linguistics, 2015.

[90] Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367. Association for Computational Linguistics, 2015.

[91] B. Alex, C.A. Llewellyn, C. Grover, J. Oberlander, and R. Tobin. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of 10th Language Resources and Evaluation Conference*, 2016.

[92] Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500, 2014.

[93] T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 23(176):88–93, 1975. doi: 10.1179/sre.1975.23.176.88.

[94] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062, 2012.

[95] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview, 2006.

[96] Christopher Manning, Raghavan Prabhakar, and Hinrich Schutze. *Introduction to Information Retrieval*. 2008. ISBN 0521865719. doi: 10.1162/coli.2009.35.2.307.

[97] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[98] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[99] M Fernández-Delgado and Eva Cernadas. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine ...*, 15:3133–3181, 2014.

[100] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[101] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*, 2020.

[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[103] Diana Maynard and Adam Funk. Automatic detection of political opinions in tweets. In *The semantic web: ESWC 2011 workshops*, pages 88–99. 2012.

[104] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009. ISSN 03064573. doi: 10.1016/j.ipm.2009.03.002.

[105] Tom L Beauchamp, James F Childress, et al. *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

[106] Karën Fort and Alain Couillault. Yes, we care! results of the ethics and natural language processing surveys. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1593–1600, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L16-1252`.

[107] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL `https://www.aclweb.org/anthology/P16-2096`.

[108] N. Tirino. *Cambridge analytica: il potere segreto, la gestione del consenso e la fine della propaganda*. Libellula edizioni, 2019. ISBN 9788867355129. URL `https://books.google.co.uk/books?id=OHm5yAEACAAJ`.

[109] Emanuelle Burton, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, and Toby Walsh. Ethical considerations in artificial intelligence courses. *AI magazine*, 38(2):22–34, 2017.

[110] DW Gotterbarn, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Vazansky, and Marty J Wolf. Acm code of ethics and professional conduct. 2018.

[111] Emily M. Bender, Dirk Hovy, and Alexandra Schofield. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*,

pages 6–9, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.2. URL https://www.aclweb.org/anthology/2020.acl-tutorials.2.

[112] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Topic models and n-gram language models for author profiling-notebook for pan at clef 2015. 2015.

[113] Francisco Rangel, Paolo Rosso, Moshe Koppel, and Efstathios Stamatatos. Overview of the Author Profiling Task at PAN 2015. 2015. ISSN 16130073.

[114] Paul T Costa Jr and Robert R McCrae. *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc, 2008.

[115] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[116] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah a Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. *Human Language Technologies*, 2(2):42–47, 2011.

[117] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1039.

[118] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1108. URL https://www.aclweb.org/anthology/D14-1108.

[119] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn:

Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[120] Matthew Michelson and Sofus A Macskassy. What blogs tell us about websites: a demographics study. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 365–374. ACM, 2011.

[121] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, may 2010. ELRA.

[122] Miguel A Alvarez-Carmona, A Pastor López-Monroy, Manuel Montes-y Gómez, Luis Villasenor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: Author profiling task. *Working Notes Papers of the CLEF*, 2015.

[123] Carlos E González-Gallardo, Azucena Montes, Gerardo Sierra, J Antonio Núnez-Juárez, Adolfo Jonathan Salinas-López, and Juan Ek. Tweets classification using corpus dependent tags, character and pos n-grams. 2015.

[124] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. Author profiling using stylometric and structural feature groupings. 2015.

[125] Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September 2014. Springer. ISBN 978-3-319-11381-4. doi: 10.1007/978-3-319-11382-1_22.

[126] Martin Potthast, Francisco Rangel, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In Gareth J. F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International*

Conference of the CLEF Initiative (CLEF 17), Berlin Heidelberg New York, September 2017. Springer.

[127] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *Working Notes Papers of the CLEF 2017 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, sep 2017.

[128] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[129] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Hyperlocal home location identification of twitter profiles. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pages 45–54, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4708-2. doi: 10.1145/3078714.3078719. URL `http://doi.acm.org/10.1145/3078714.3078719`.

[130] Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1239–1248, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1116`.

[131] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[132] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992. URL `https://www.aclweb.org/anthology/J92-4003`.

[133] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[134] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. Inferring the socioeconomic status of social media users based on behaviour and language. 2016.

[135] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[136] Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim. Arabic tweeps gender and dialect prediction. In *CLEF (Working Notes)*, 2017.

[137] Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, and Liviu P Dinu. Including dialects and language varieties in author profiling. *arXiv preprint arXiv:1707.00621*, 2017.

[138] Guillaume Kheng, Léa Laporte, and Michael Granitzer. Insa lyon and uni passau's participation at pan@ clef'17: Author profiling task. In *CLEF (Working Notes)*, 2017.

[139] Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. In *CLEF (Working Notes)*, 2017.

[140] Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. Pan 2017: Author profiling-gender and language variety prediction (notebook for pan at clef 2017, 2nd place). 02 2018.

[141] Alexander Ogaltsov and Alexey Romanov. Language variety and gender classification for author profiling in pan 2017. In *CLEF (Working Notes)*, 2017.

[142] Rodrigo Ribeiro Oliveira. Using character n-grams and style features for gender and language variety classification. In *CLEF (Working Notes)*, 2017.

[143] Nils Schaetti. Unine at clef 2017: Tf-idf and deep-learning for author profiling. In *CLEF (Working Notes)*, 2017.

[144] Liliya Akhtyamova, John Cardiff, and Andrey Ignatov. Twitter author profiling using word embeddings and logistic regression. In *CLEF (Working Notes)*, 2017.

[145] Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak. Author profiling with bidirectional rnns using attention with grus: Notebook for pan at clef 2017. In *CLEF 2017 Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14 September 2017*, volume 1866. RWTH Aachen, 2017.

[146] Sebastian Sierra, Manuel Montes-y Gómez, Thamar Solorio, and Fabio A González. Convolutional neural networks for author profiling. *Working Notes of the CLEF*, 2017.

[147] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from twitter. *PloS one*, 10(7):e0131469, 2015.

[148] Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the W-NUT Workshop*, 2016.

[149] R. W. Sinnott. Virtues of the haversine. *Sky and Telescope*, 68(2):159, 1984.

[150] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[151] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[152] William H Press. Gaussian mixture models and k-means clustering. In *Numerical recipes (3rd edition): The art of scientific computing*. Cambridge university press, 2007.

[153] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[154] J. Robinson. *Evolving English WordBank: A Glossary of Present-Day English Dialect and Slang*. Northern Map Distributors, 2015. ISBN 9781909914896. URL `https://books.google.co.uk/books?id=X9nssgEACAAJ`.

[155] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th*

ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1082–1090. ACM, 2011.

[156] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. User profiling with geo-located posts and demographic data. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 43–48, Austin, Texas, November 2016. Association for Computational Linguistics.

[157] Austin Troy. Geodemographic segmentation. In *Encyclopedia of GIS*, pages 347–355. Springer US, 2008. ISBN 978-0-387-30858-6. doi: 10.1007/978-0-387-35973-1_456. URL `http://dx.doi.org/10.1007/978-0-387-35973-1_456`.

[158] Catherine A Sugar and Gareth M James. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 2011.

[159] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962. doi: 10.1214/aoms/1177704472. URL `https://doi.org/10.1214/aoms/1177704472`.

[160] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt*, pages 95–100. Springer, 2011.

[161] Constantinos Boulis and Mari Ostendorf. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 435–442. Association for Computational Linguistics, 2005.

[162] Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 710–718. Association for Computational Linguistics, 2009.

[163] Aida Ali, Siti Mariyam Shamsuddin, Anca L Ralescu, et al. Classification with class imbalance problem: a review. *Int. J. Advance Soft Compu. Appl*, 7(3):176–204, 2015.

[164] Peter Elias, Margaret Birch, et al. Soc2010: revision of the standard occupational classification. *Economic and Labour Market Review*, 4(7):48–55, 2010.

[165] D. Rose and D. Pevalin. : Re-basing the ns-sec on soc2010: a report to ons. Technical report, 2010.

[166] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[167] Dimitri P Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

[168] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.2. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.2`.

[169] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[170] S. McClendon and A. C. Robinson. Leveraging geospatially-oriented social media communications in disaster response. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 5(1): 22–40, 2013.

[171] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics.

[172] Bahareh Rahmanzadeh Heravi and Ihab Salawdeh. Tweet location detection. In *Computation + Journalism Symposium*, 2015.