Predicting Author Traits Through Topic Modeling of Multilingual Social Media Text

By

Caitlin McCollister

Submitted to the Department of Electrical Engineering and Computer Science and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Master of Science

Bo Luo, Chairperson

Committee members

Arvin Agah

Luke Huan

Date defended: ______________________________

The Thesis Committee for Caitlin McCollister certifies
that this is the approved version of the following thesis :

Predicting Author Traits Through Topic Modeling of Multilingual Social Media Text

Bo Luo, Chairperson

Date approved: ______________________________

Abstract

One source of insight into the motivations of a modern human being is the text they write and post for public consumption online, in forms such as personal status updates, product reviews, or forum discussions. The task of inferring traits about an author based on their writing is often called "author profiling." One challenging aspect of author profiling in today's world is the increasing diversity of natural languages represented on social media websites. Furthermore, the informal nature of such writing often inspires modifications to standard spelling and grammatical structure which are highly language-specific.

These are some of the dilemmas that inspired a series of "shared task" competitions, in which many participants work to solve a single problem in different ways, in order to compare their methods and results. This thesis describes our submission to one author profiling shared task in which 22 teams implemented software to predict the age, gender, and certain personality traits of Twitter users based on the content of their posts to the website. We will also analyze the performance and implementation of our system compared to those of other teams, all of which were described in open-access reports.

The competition organizers provided a labeled training dataset of tweets in English, Spanish, Dutch, and Italian, and evaluated the submitted software on a similar but hidden dataset. Our approach is based on applying a topic modeling algorithm to an auxiliary, unlabeled but larger collection of tweets we collected in each language, and representing tweets from the competition dataset in terms of a vector of 100 topics. We then trained a random forest classifier based on the labeled training dataset to predict the age, gender and personality traits for authors of tweets in the test set. Our software ranked in the top half of participants in English and Italian, and the top third in Dutch.

Acknowledgements

I wish to thank my faculty advisor, Bo Luo, and industry collaborator Shu Huang for their guidance during the rapid development period of the software for the PAN 2015 competition. Without their support and encouragement to further investigate sentiment analysis and topic modelling, I might have settled for yet another text classifier based on n-gram features, and I might not have had nearly as much fun.

To the organizers of the shared tasks at the PAN workshops: Thank you all for the hard work you've put in year after year to make these competitions not only possible, but a pleasure to watch and participate in. I look forward to entering another challenge in the future. *Dank je wel, Vielen Dank, ¡Gracias! Gràcies! Grazie! et Merci.*

To those closest to me and who have known me the longest: thank you for breakfast (or lunch, or bedtime) discussions of Russian literature, global currency markets, and statistical mechanics of biomolecules; for always having one more prime factorization problem ready on the way to school; for reminding me to pull the latest updates before I commit; for making many excellent curries and encouraging me to push onwards in school; and for persuading me to keep my options open for a fallback career in Computer Science just in case the baby seal caretaker thing never works out—I'm not giving up, though.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There has long been a desire to say whether a given collection of words is original or borrowed, in whole or in part. Likewise, spoken or written works often become separated from the identity of their creators, only to have those who read or hear the works to guess whether it was a known author or some other person who wrote them. The rise of the world wide web and, in particular, search engine effectiveness, has made it easier for the general public to locate and digest a variety of knowledge sources on a plethora of topics.

1.1 Motivation

Typically, the goal of author profiling is to uniquely identify the author of a collection of text; the name, or other method of identification, may already be known or it may be of little consequence. Instead of trying to determine exactly who wrote a text, author profiling tries to describe the person who wrote it by introducing additional information about their temporary or permanent state.

One of the benefits of approaching this question in a computer-assisted or fully-automated manner is the ease with which the same analysis steps can be carried out on many authors' work consistently and rapidly. That capability allows the researcher the opportunity to speculate about the distribution of the predicted attributes in the population represented in the data, or to make relative judgments of the form "person A is more *something* than person B."

The purpose of shared tasks like the Author Profiling task at PAN is to provide a common dataset and direction so that the research community can compare different approaches to accomplishing a single goal. While PAN is part of the CLEF initiative (Conference and Labs of the Evaluation Forum, or Cross-Language Evaluation Forum), its distinct focus is on detecting plagiarism and abuse of social media platforms.

From 2013 to 2015, the three tasks set out for the PAN section were Plagiarism Detection, Authorship Attribution and Author Profiling. The Authorship Attribution task addresses the question: “given a set of apparently similar texts, are they all written by the same author?” In contrast, Author Profiling presents sets of documents written by each of a collection of authors, and tries to determine traits about those authors based on their writing.

1.2 Document Overview

Chapter 2 of this thesis gives historical and motivational support for the idea of shared tasks in computing, particularly information retrieval and natural language processing. It also introduces some relevant terms from the fields of personality psychology and psychometrics, the systematic evaluation of psychological attributes, permanent or temporary (Srivastava, 2014). Finally, we illuminate the very beginnings of what has come to be known as Twitter, a microblogging service that is the source of data for the PAN 2015 Author Profiling shared task.

In Chapter 3, we give a more detailed description of how the PAN workshops function, what variety of shared tasks its organizers have hosted, and describe how all of those tasks relate to one another.

Chapter 4 describes the specific configuration of the PAN 2015 Author Profiling task in which we made our software submission based on topic modelling. We will describe the properties of the 2015 dataset and contrast them with those of the previous two years of the Author Profiling task.

Finally, in Chapter 5 we include the full Notebook Paper (Mccollister et al., 2015), submitted to the working notes of the evaluation labs of the CLEF 2015 conference, describing our particular method of predicting author traits through topic modeling of multilingual social media text.

Chapter 2

Background

2.1 Shared tasks in Information Retrieval

As global communication and collaboration has become increasingly common, a number of *shared tasks* have been made available to people interested in computational disciplines such as information retrieval, bioinformatics, and computational linguistics. One of the most famous is the Text Retrieval Conference (TREC), a yearly event that hosts competitions in multiple tracks such as high-recall methods for legal document retrieval, ranking algorithms for web search results, and real-time decision support systems. From its beginning in 1992, the conference has received significant economic investment and dataset contributions from the National Institute of Standards and Technology (NIST) and other United States government entities.

One of its tracks that drew one of the more international participant pools was the Cross-Language Track, which focused on finding documents relevant to a search query no matter what language the documents are written in. While this track is no longer held at TREC, research in related areas has continued at least two other recurring workshops around the world.

In Japan, the National Institute of Informatics (NII) has held the NII Testbeds and Community for Information access Research (NTCIR) workshops since 1999 (Kudo, 2010). NTCIR supplies a variety of test datasets in several East Asian languages as well as English, and provides a framework for systematic evaluation and comparison of approaches to specific information retrieval tasks. Some of its tasks have involved searching temporally diverse knowledge bases for specific events, as well as patent retrieval and translation.

Meanwhile, in Europe, a series of workshops known as the Cross-Language Evaluation Forum (CLEF) began in 2000 (Ferro, 2014). The workshops were held in conjunction with other, larger conferences such as the European Digital Library Conference (EDLC). In these first several years, CLEF offered tasks in cross-language retrieval for image collections, spoken documents, and genre-specific collections with reference datasets from many languages throughout Europe and Asia. It has since expanded into a conference of its own, with invited talks and papers as well as labs which are open to participation from any interested party. Beginning in 2010, the organization changed its name to the Conference and Labs of the Evaluation Forum, conveniently maintaining the same acronym. It also underwent a slight shift in focus away from pure document retrieval to more varied goals, such as content annotation, music retrieval and plagiarism detection.

One of the new labs to join CLEF in 2010 had been running at other conferences for several years already and needed a permanent home: **PAN**, a workshop on **P**lagiarism Analysis, **A**uthorship Identification, and **N**ear-Duplicate Detection. As part of CLEF 2010, PAN offered two evaluation labs in plagiarism detection and Wikipedia vandalism detection.

2.2 Evaluating Shared Task Submissions

The PAN competition organizers provide labeled training datasets each year, which are publicly available for download from the time the competition is announced. There is no cost to the participants in the competition. All results are analyzed together in an overview paper written by the organizers after the competition ends, and participants are encouraged to submit a notebook paper describing their approach so that others may learn from it.

Software submissions are accepted from registered teams or individuals for a period of time after the training data is released and the goals of the task are set out. At the end of the submission period, the organizers evaluate the submissions by running them against a hidden test dataset, which is not made publicly available until several years later, so that parts of it can be used as a consistent benchmark for several years in a row.

All groups are provided with a virtual machine environment which can be used for development, so all teams at least have a working development environment with their choice of operating system installed from the start. However, while teams can use their own resources during the development period, and this may be an advantage when it comes to optimizing parameter settings evaluating the effectiveness of different approaches, the final “runs” of the software must be performed on the provided virtual machines. This requirement serves several purposes.

First, it acts as a common ceiling on the raw computational resources the teams can rely on, at least during the phases of training a model on an arbitrary training dataset and testing on an arbitrary test dataset. The relative performance of different approaches would be difficult to judge if some teams had at their disposal an entire cluster of machines for high performance parallel computing, while others had only modest consumer-grade hardware.

Second, it allows the competition organizers to archive part or all of the virtual machine image with each run of the software submission. This provides a degree of repeatability in a field where it is often difficult to reproduce the results of a previous study if it relies on a particular configuration of the operating system, userspace environment settings, and externally sourced software libraries.

Even if the current version of some specialized packages for machine learning or natural language processing are no longer supported, or even available, five years in the future, the virtual machine images could still be restored and the software could be evaluated on new datasets. This need does in fact arise quite frequently. For those people responsible for collecting and curating a dataset for text analysis, especially benchmarking of state-of-the-art techniques, it is desirable have a baseline performance established by several well-studied techniques to ensure the new dataset will be helpful in distinguishing the relative performance of new methods.

2.3 Personality Theory and Measurement

Categorizing and cataloging the spectrum of human personality has been a significant research topic in psychology for many decades. The traditional approach to assessing an individual’s personality with respect to a common taxonomy involves answering questions in a carefully constructed questionnaire. The subject's responses are used to calculate a numeric score for each of several factors, the number and nature of which vary depending on the psychological model being used.

One of the most established models, which has been in progress in some form for at least 80 years (Allport & Odbert, 1936) is the **Big Five** or **Five-Factor Model.** As a taxonomy for personality, it contends it is possible to describe the lifelong personality traits of non-pathological individuals according to numeric values along five dimensions (Soto & John, 2016). Once again, their names and philosophical have been heavily disputed and rearranged by various research groups over time (John et al., 1988), but one such labeling uses the names **extraversion**, **agreeableness**, **conscientiousness**, **neuroticism**, and **openness**.

However, many personality questionnaires are time-consuming for the subject to complete, and difficult for psychologists to standardize and interpret across different languages and cultures (Guido et al., 2015; Romero et al., 2012). In recent years, another approach is gaining popularity: rather than ask a subject to respond to a fixed set of questions, perhaps we can try instead to develop a representation of the individual's personality based on automatic processing of text he or she has already written in a natural setting. Given a representation in an appropriate feature space, developed using a labeled training dataset, we can apply our choice of machine learning techniques for classification or regression to predict the personality traits, age, or gender of an individual.

2.4 Origins of Twitter

The Twitter microblogging service originated in 2006 as an unnamed software tool for communicating between employees at a company called Odeo, in San Francisco, California (Pederson, 2011). Odeo started as a video broadcasting and podcast delivery service, but its founding members also had combined experience in using and implementing instant messaging services, blogging platforms, and emergency services dispatching systems. Several of its founding employees collaborated on an idea that Dorsey had been experimenting with for several years: a program that would monitor an email address for incoming messages, sent by Dorsey from his BlackBerry 850 handset, forward them to a list of his friends, and finally collect his friends' replies and send them in one email to the address associated with his BlackBerry (Sarno, 2009). Dorsey had actually been fairly content with this level of functionality, except for the fact that as the only of his friends who owned an email-enabled mobile device, he was the sole originator of new message chains in the system.

As SMS (Short Message Service) technology gained wider adoption in the United States, and more people had the capability to send messages of up to 160 characters from mobile devices, Dorsey and a select few coworkers at Odeo began to reconsider the idea as a possible service. The central concept was that users of the service would create an account identified by a short user name and store a list (or lists) of friends who were also using the service (Sarno, 2009). The user could write a message that they likened to one line of a journal entry, relating what they were doing and where they were doing it at the moment they wrote the message. The contribution of the service would be the convenience of only needing to send the message to one place in order to reach an easily-expandable list of people, and storing the immediate responses to the broadcast messages.

In order to minimize the amount of effort and robust software development required to implement the idea, they decided it utilize existing communications protocols of email and SMS, always conforming to the length and character set constraints of SMS messages. While there existed methods of sending longer messages by sacrificing some of the message length for header data that would supposedly enable end-user devices to display multiple messages as one long one, the results were inconsistent and the Odeo group believed it wouldn't be worth the trouble of trying to use this capability. Thus, they decided to limit the user names themselves to 15 characters, and save room for an @ symbol, a colon, and a space before the text content of the message. While this left 142 characters of the 160 characters in the SMS format, they chose to limit messages themselves to 140 characters which they found easier to explain to new users (Sarno, 2009).

Chapter 3

PAN Workshop Tasks and Goals

3.1 Plagiarism Analysis

In the broad field of plagiarism analysis, one of the first steps to solving a problem is a choice between *corpus-based* and *intrinsic* techniques. Corpus-based techniques utilize the patterns and information contained in a wider collection (or *corpus*) than the text under investigation (Gollub et al., 2013). That document and the whole corpus may be similar in language, content, length, time of creation, and cultural origin, but this is not always the case. Furthermore, there is no minimum or maximum size of the corpus. Some applications might use only a dozen essays as examples, while others may use text extracted from billions of webpages. No extreme on any of these scales is better or worse in an absolute sense. Depending on the application, the utility of a corpus for plagiarism analysis can be either its specificity or its generality.

A corpus-based approach to plagiarism detection might be to choose a document representation and compare a suspect document to others that have been placed in the same representation. The primary difficulty is the potentially enormous volume of source documents from which the suspect one may have improperly borrowed. A plagiarizing author is likely to take advantage of whatever content is available, and unlikely to explicitly point out where the content came from. To make the plagiarism detection task feasible, one needs a combination of highly efficient document-to-document comparison methods, and clever ways of narrowing down the list of source documents to compare against.

Performing intrinsic plagiarism analysis requires identifying passages of a document that seem somehow out of place (Gollub et al., 2013). As a first step, many methods segment the document based on transitional words and phrases (such as “contrary to popular belief,” or “studies have shown that . . . ”), or paragraph and section breaks. Once a document is segmented in this fashion, one can begin the process of comparing segments against each other with respect to any number of text analysis features.

Stylistic features include sentence length, usage of active or passive voice, or the relative proportion of different parts of speech. Content-based features could include the portion of unique words used in a segment, portion of words used in other sections but not in the suspicious segment, or the presence or absence of field-specific words or jargon.

In practice, automated intrinsic plagiarism analysis is usually only the first step in finding and investigating a possible case of plagiarism. If and when part of a text has been flagged as suspicious this would be followed by searches against external data sources—or the entire world wide web—for the original source, and ultimately human inspection of the case in context.

At least one type of plagiarism detection task has been hosted at PAN every year since 2009, attracting a consistent following of attendees. Most recently, the organizers have begun to test the concept of having a task for the construction of corpora to be used in plagiarism analysis. The motivation for such a task is that the in-depth nature of techniques used in this field, and the required access to substantial source material. In particular, the organizers hoped contributors would contribute collections of writing samples annotated with “known” author identifiers, and either real-life examples of plagiarism or those that the contributors have constructed themselves. The response so far has been enthusiastic, and in 2015 included a corpus for plagiarism analysis in Farsi. As none of the current organizing committee members speak this language, having that dataset available for future years will add welcome diversity to the shared tasks and hopefully attract more participants from outside Europe and North America.

3.2 Wikipedia Vandalism Detection

The Wikipedia vandalism detection task, offered in 2010 and 2011, explored automated methods of determining whether a change made to an article constituted vandalism or was a well-intentioned modification. Articles on the online encyclopedia can be edited by anyone, even without creating and logging into an account. Users with a vested interest in the topic of an article, or those simply seeking thrills on an encyclopedic scale often attempt to make changes to articles which introduce false or irrelevant information, erase all or most of the content in the article, or litter it with general profanity. As a safety mechanism, when a change is made by an anonymous author or to any page that is known to be a controversial topic, the change is routed to volunteer human editors for approval before being added to the version of the page shown to the public. Some automated measures are in use by Wikipedia to screen the most obvious vandalism changes right away. However, the volunteer editing community of the English-language Wikipedia has expressed a preference over time for keeping this layer of human oversight in place, so further development of automated methods for detection was deemed largely unnecessary.

3.3 Wikipedia quality flaw detection

In 2012, PAN offered a Wikipedia quality flaw detection task (Anderka & Stein, 2012). Rather than providing existing articles and changes to those articles that may be destructive, the PAN organizers assembled a corpus of over 1.5 million articles from the English-language Wikipedia, about 15 percent of which have been “tagged” by actual readers as having at least one quality flaw. The flaws under consideration could concern the factual content of the article, the writing style, or the validity of external references.

The shared task was framed as a one-class classification problem for each of ten specific flaws: given the content of an article, predict whether the article exhibits that flaw. What seemed a promising concept, however, the task was rather more difficult than intended. Inherently, there are few or no articles that are known negative examples—articles which have been examined and marked as not having a flaw—because articles are assumed to begin in an unflawed state and those that are flawed are only sometimes tagged as such. An article tagged with one type of flaw is not necessarily corrected before other changes are made, potentially introducing intertwined problems of multiple types.

Some of the flaws themselves involve subjective judgment in their very definition, such as when an article is an advertisement for a product or service: assuming one could detect a relationship between article text and an advertisable entity, it may be one that genuinely warrants factual documentation. The difference between overt or misleading advertisement and legitimate public knowledge was just one example of the difficulty in automatically identifying a target that is both conceptually fuzzy and logistically messy. The task attracted only three participants, and marked the end of Wikipedia-related tasks at PAN.

3.4 Author Identification

Beginning in 2011, PAN has offered some variation of an *author identification* task. Like many other experts in the field, the competition organizers distinguish between two types of scenarios encountered in this category: authorship attribution and authorship verification. Both of them deal with trying to identify the author of a written work, but they differ in the number of possible conclusions to be framed and evaluated.

Authorship attribution refers to a problem in which the practitioner is given known examples of the writing style of multiple authors as well as a collection of anonymous documents. In this type of problem, also described as *closed-class*, the anonymous documents are not necessarily written by a single unknown author, but the author of each anonymous document is assumed to be among those in the provided example set. This is essentially a text categorization problem in which each category is a known author. Text categorization problems have been solved through the use of clustering or classification methods to a high degree of accuracy. Some challenges that an authorship attribution solution has to overcome include an imbalanced number of writing samples by the known authors, having multiple genres of writing included for some or all of the known authors (such as a mix of autobiographical passages or surrealist fiction), or being presented with anonymous documents written in a different language than those in the labeled set.

Authorship verification, by contrast, is an *open-class* problem because the question to answer is whether a given anonymous document was written by an author in a set of known candidate authors. Most of the PAN author verification tasks have dealt with a special case of this problem, in which there is only a single candidate author, and the objective is to predict whether an anonymous document was written by that person or someone else. Depending on how much outside knowledge is encoded and given to a piece of software about the wide variety of writing in existence, it can be very difficult for the system to make this distinction if it is only given labeled examples that truly were written by the author in question. While this problem is already difficult to break down conceptually, in practice there may be even more complications: the anonymous work could be very short and the example base long and diverse enough that the unknown sample resembles parts of it but not others. In an inverse scenario to that of plagiarism detection, in which an author wants a piece of misappropriated writing to be acknowledged as their own, the anonymous author may be actively trying to avoid being recognized as a known person.

The PAN Author Identification track of 2012 (Juola, 2012) was especially popular, with 25 participants solving several versions of the problem on a dataset of fiction works in English. Participants could choose to solve some or all of the problems, including authorship attribution and verification. It also introduced as something similar to intrinsic plagiarism analysis, in which the task was to segment single documents written by multiple authors, and determine which of the listed authors wrote exactly which parts of the document. Unlike in plagiarism analysis, this task has potential applications even in situations where all authors involved were acting with benign motives, such as automatic annotation of collaborative work such as multiple-author academic papers.

PAN 2013 was the first time the author identification dataset consisted of works in languages other than English, adding Spanish and Greek to the mixture (Juola & Stamatatos, 2013). The training dataset consisted of an imbalanced number of problems across the three languages—10 in English, 5 in Spanish, and 20 in Greek—and each problem was an instance of open-class, single candidate author verification. In a surprise move, however, the organizers selected a test dataset that was balanced over the three languages. Since the 2013 competitions, the organizers have demonstrated a deliberate intent to somehow alter the composition of examples between the training and test sets, either through the class balance or imbalance, or even by adding in extra document genres.

Chapter 4

PAN Author Profiling Task

Each year, the task adds new prediction targets to the scope of the problem and modifies others. Packaged with the tweet data for each language in the training set is a file containing the actual values of each attribute to be predicted with an anonymous identifier string assigned to each author. Due to Twitter, Inc.'s privacy policies regarding distributing user data, the actual account names are stripped from the files.

The organizers reserve the right for the hidden test datasets to differ from the released training datasets in number of authors, documents per author, and class balance or imbalance. The testing dataset will be drawn from the same genre(s) of writing corresponding to those in the training dataset.

4.1 Provided Training Data, Years 2013 and 2014

In 2013, the author profiling task was to predict the age and gender of internet users from the content of their blog posts (Rangel et al., 2013). Only authors who were willing to specify a gender of either "male" or "female" were included in the data collection. All the included authors fell into one of three age groups:

- 13 to 17 years old (4 year span)
- 23 to 27 years old (4 year span)

- 33 to 47 years old (14 year span)

It is worth noting that the age group boundaries are not adjacent (there is a six year gap between each consecutive pair), and that they do not all span the same number of years. This is presumably due to the relatively lower number of consistent Twitter users in older age groups, which makes it harder to find willing volunteers.

In 2014, the goal was again to predict each author's gender and age group. However, the representation of age changed to five groups:

- 18 to 24 years old (6 year span)
- 25 to 34 years old (9 year span)
- 35 to 49 years old (14 year span)
- 50 to 64 years old (14 year span)
- 65+ years old

The dataset consisted of documents in English and Spanish (labeled and separated from each other), and four document types: social media (such as Facebook status updates), blog posts, hotel reviews, and Twitter messages.

The 2014 Twitter subsection consisted of about 380,000 tweets from 478 authors. Some authors had as many as 1000 tweets provided, but some only a few hundred. This training data is still available for download and can be used as a supplement to train or validate the approaches in subsequent years. However, participants in future iterations of the competition will need to take notice of the differences in labeled attributes, number of samples per author, and various levels of preprocessing and cleanup carried out by the competition organizers before releasing the datasets.

4.2 Provided Training Data in 2015

In 2015, the author profiling dataset consisted entirely of Twitter messages, rather than having multiple genres of writing for multiple languages, all of which would have to be collected, evaluated and interpreted with substantial effort.

4.2.1 Number of Authors and Tweets

Of the 2015 datasets, English and Spanish have the most authors at 152 and 100 respectively, while Italian and Dutch have only 38 and 34 authors. There are a maximum of 100 tweets per author and the average tweet count per author is roughly 95, depending on whether one counts exact duplicate tweets only once. The total number of authors in the four languages is 324, and the number of tweets 31,000. Clearly, the vastly reduced number of tweets per author compared to the previous year, and the size imbalance between the four languages' datasets is intended to make this task challenging even for teams who participated in the past.

4.2.2 Languages

At the broadest level, the authors in this dataset are divided into four groups based on their preferred tweeting language. The four languages covered are **English**, **Spanish**, **Dutch** and **Italian**, primarily due to the nationalities and native languages of the competition organizers who recruited volunteer authors and collected the data.

The language affiliated with an author is their most commonly used or personally preferred language on Twitter, and we assume that they are at least proficient enough in using the language to write their own tweets. It is the language in which they answered the questionnaire asking their age, gender, and 10 questions related to five aspects of their personality. However, it is by no means the only language they ever tweet in. Many of the authors in the PAN 2015 Author Profiling training dataset do in fact have tweets that are clearly not in their assigned language, or that are a mixture of languages within the same tweet.

4.2.3 Gender and Age Group

Gender is still one of the prediction targets for authors in all languages, but age labels are provided only for English and Spanish authors. The boundaries of the age groups changed once again, and this time they are:

- 18 to 24 years old (6 year span)
- 25 to 34 years old (9 year span)
- 35 to 49 years old (14 year span)
- 50+ years old

Since there is no training data for age in Italian and Dutch, participants do not make an age prediction for those authors.

4.2.4 Personality

Likely the most significant change in prediction targets for 2015, this dataset includes labels for five personality-related qualities. The following short descriptions are taken from the two questions from the BFI-10 Personality Inventory for each of the five attributes (Rammstedt & John, 2007).

Extraversion

High: *outgoing, sociable*
Low: *reserved*

Agreeableness

High: *generally trusting considerate and kind*
Low: *tends to find fault with others*

Conscientiousness

High: *does a thorough job*
Low: *tends to be lazy*

Neuroticism

High: *gets nervous easily*
Low: *is relaxed, handles stress well*

Openness

High: *has an active imagination*
Low: *has few artistic interests*

Chapter 5

Notebook for PAN at CLEF 2015

5.1 Introduction

Over time, it is common for a single Twitter user to publish tweets related to multiple aspects of his or her life which may be quite independent of each other. For example, a user might write about his or her professional occupation while at work or attending a conference, post pictures of family members while at home or on vacation, and link to news articles about international politics while reading on the train during the daily commute. Even when examining fewer than one hundred tweets per author, as is the case with the PAN'15 training corpus, most authors' Twitter streams are effectively a mixture of distinct subjects or topics. Our approach to the PAN'15 Author Profiling task (Rangel et al., 2015) is motivated by the expectation that authors will produce language that points to a variety of different or even contradictory traits, and the observation that certain common themes do appear repeatedly even across authors and target classes.

We use the topic modeling software package MALLET (McCallum, 2002) to construct models of 100 topics each for the four languages in the scope of the PAN'15 Author Profiling task. The topics in these models are essentially groups of words that may be semantically related and are frequently observed near each other in a collection of training documents. To ensure we had a sufficiently large body of examples to build such models, we collected our own corpora of Twitter messages in English, Spanish, Italian and Dutch. We also use MALLET to infer the most likely distribution over the generated topics that could have produced any given tweet instance, allowing us to represent tweets as concise 100-element document-topic distribution vectors. These representations serve as inputs to a set of classifiers that make predictions for unknown authors' age, gender, extroversion, stability, agreeableness, conscientiousness, and openness.

5.2 Background

5.2.1 Previous PAN Author Profiling Approaches

Some of the more successful entries in previous years, especially PAN'14 (Pardo et al., 2014), are those that acknowledge the substantial diversity of authors within the target classes for predictions. The PAN'14 solution by López-Monroy et al (López-Monroy et al., 2014), one of the top-ranked entries for accuracy in both the English and Spanish Twitter subcorpora, extracts weighted word frequency features from documents and compares the values to those typical of various subprofiles. Those subprofiles are subsets of authors within a target class, such as females aged 18 to 24, who were grouped together by a clustering algorithm based on the most distinguishing words in their writing. By generating and using more fine-grained target classes, the software can train a model that recognizes and accommodates a variety of writing styles and subjects that map to the same original target class.

Another PAN'14 entry by Weren et al (Weren et al., 2014b), which was further refined in a follow-up paper (Weren et al., 2014a), demonstrated the potential effectiveness of information retrieval based features, such as the cosine similarity of a given test document and the labeled training documents. In this implementation, a set of similarity features was found to be more discriminative for age and gender than several common readability measures or the prevalence of dictionary words and punctuation marks. Treating incoming test documents as queries in a document retrieval system and using a combination of aggregate functions on the top-ranked results allows the classification to be based on the most closely related training documents even if many dissimilar documents exist within the correct target class.

5.2.2 Previous Work in Topic Modeling

Several research groups have pursued the use of topic modeling, including Latent Dirichlet Allocation (LDA) (Blei et al., 2003), either to gain insight about the processes involved in social media communication (Bamman et al., 2014), or to make predictions about authors and the text they produce online (Owoputi et al., 2013). Schwartz et al have conducted large-scale studies based on millions of English-language Facebook status updates written by tens of thousands of users, and have published several resulting linguistic resources that they claim to be accurate predictors of authors' age, gender, and personality (Schwartz et al., 2013). They were able to collect the status messages, which are often of a similar length to Twitter messages, from volunteers who specified their age and gender directly, and completed a personality profiling questionnaire yielding numeric values for the same "big five" personality traits that we aim to predict in the Author Profiling task. The group has distilled and made available two types of resources based on this work. The first type includes weighted lexica of one- to three-word phrases that are most discriminative for high or low values of the measured personality traits, as well as for several age bins and males versus females. We have made use of these lexica to compute twelve of the features in our "secondary" feature set, described in section 5.3.1.2. The other resource made available by Schwartz et al is a set of word clusters consisting of the top 20 words representing topics in an LDA-derived topic model of 2000 topics.

To explore the viability of using topic models to generate features for the kind of data in the Author Profiling task, we initially implemented a set of 2000 features corresponding to the topic word clusters published by Schwartz et al. In these features, we summed the number of words in the cluster that appeared in the tweet at least once, then weighted that sum according to the global term frequencies of the matching words and the length of the tweet. Our initial experiments using these features for English tweet classification looked promising, but faced one significant challenge: the studies from which the topics and word clusters were derived came from exclusively English-language text, and the features were not particularly useful for our non-English subcorpora. Conducting additional massive studies on Spanish, Italian and Dutch speaking Twitter users with known gender, age and personality was beyond the scope of our entry in the Author Profiling task. This dilemma inspired the collection of our own unlabeled Twitter corpora for all four languages, with fewer total documents than Schwartz et al used, but more than the number supplied as labeled data in the PAN'15 training corpus. The resulting four corpora are described further in section 5.3.1.1.

5.3 Software Design and Implementation

One of our earliest design decisions was whether to treat all of a given author's tweets as a single body of text, cluster them together by content or in fixed-size chunks, or process them as independent documents all associated with the same author. Our intuition was that the best way to account for high intra-author variation in tweet subject matter and style would be for our software to treat individual Twitter messages as instances in a classification problem, and pool the predictions for all of an author's tweets to make a single prediction per (author, attribute) pair at the end of the testing phase. We implemented and tuned our software for the individual tweet representation, but included a configuration flag to allow concatenating all tweets per author so that we could test the viability of that representation after later software components were completed.

Although treating each tweet as an individual document entails a greater number of predictions to be made in the classification framework, we avoid a potential explosion in dimensionality by limiting the number of features in our models. In the interest of achieving what we felt were reasonable running times within the provided testing environment, especially if the hidden test datasets turned out to be larger than the training datasets, we decided against using the common bag-of-words or n-gram based representations, in which the size of the vocabulary (and thus the feature set) increases rapidly with the number of instances. Instead, we chose to pursue a topic modeling approach in which tweets are encoded as vectors that describe them as an inferred distribution over a fixed-size set of topics generated using the MALLET topic modeling software.

The number of topics in the LDA-based topic model has to be specified at the start of the model training process, so we made our choice of 100 topics after trying both larger and smaller numbers and noting the effect on training time, peak RAM usage, and discriminative power in terms of the computed information gain of the resulting feature sets. While MALLET can supply default values for most of the possible parameters to its particular implementation of LDA, we modified some to suit our application: we set the alpha parameter to 0.5 due to the short document lengths, used 10,000 sampling iterations, and enabled automatic hyperparameter optimization every 50 iterations. These choices were guided by general background literature on topic modeling (Steyvers & Griffiths, 2006), other studies using MALLET for social media text (Schwartz et al., 2013), and eventually by conducting multiple trials using subsets of the training data. Since our topic models are independent of the labeled training datasets provided for the shared task, they only need to be trained once to generate a set of reusable, serialized model files. Even so, we found we could complete this process on the virtual machine provided to us in the TIRA evaluation framework (Gollub et al., 2012) in under two hours per language.

5.3.1 Feature Extraction

5.3.1.1 Primary Feature Set

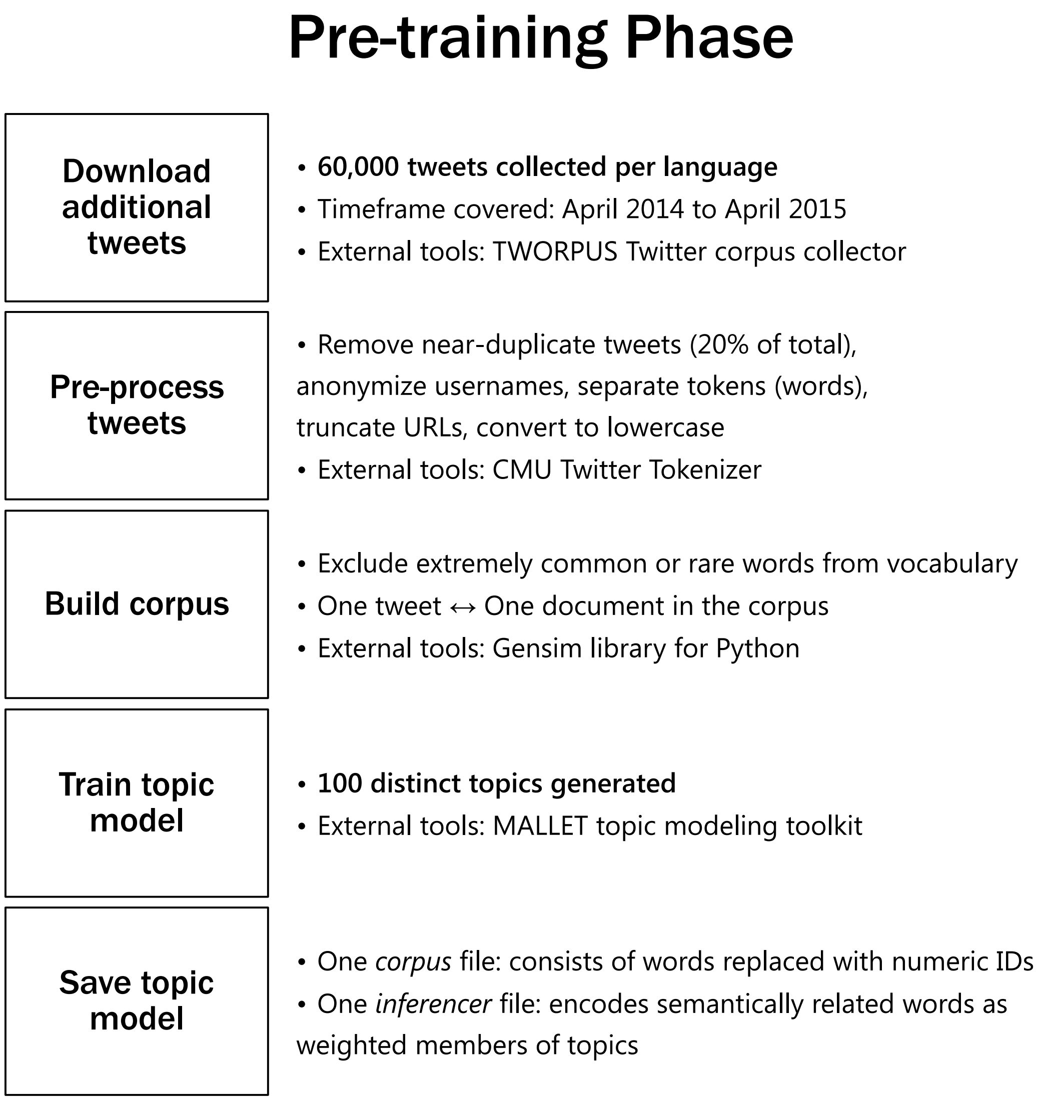
Our topic models are built from datasets of unlabeled Twitter messages which we have collected specifically for this purpose, so that none of the labeled PAN training data is used to define the topics themselves. This was accomplished using the freely-available Twitter corpus-building tool, TWORPUS (Bazo et al., 2013), which can be downloaded and run locally as a web-based application. The application connects to a centralized archive of Twitter message IDs, the user IDs that wrote them, and language tags assigned by a language detection algorithm. Because only the relevant IDs and language tags are stored in the central archive and distributed to TWORPUS users, who then use an included Twitter crawling utility to download the actual message content, the application is compliant with the terms of Twitter's developer agreement forbidding the redistribution of full tweet text and metadata.

We collected four Twitter corpora (one for English, Spanish, Italian and Dutch) spanning the time period from April 2014 to May 2015, with tweets as evenly distributed as possible throughout that period; this was still subject to the availability of the requested number of tweets for each language in the central TWORPUS archive. After retrieving the full text of over 60,000 tweets per language, we used a custom script to remove duplicate or near-duplicate tweets such as simple retweets and bulk-generated advertisements, still leaving over 50,000 tweets per language. No specific action was taken to allow or disallow multiple tweets from any given author; we found that roughly 90 percent of the collected tweets are the only messages collected from their respective authors.

The tweet text from our downloaded TWORPUS corpora needed to be preprocessed in the same fashion that our training and test data would be: we convert all text to lowercase and use the tokenizer included in the CMU Twitter Part-of-Speech Tagger tools (Owoputi et al., 2013). We performed several additional steps on just the model-training input text: we removed lists of language-specific stopwords provided in NLTK (Bird et al., 2009), and use the Python library Gensim (Řehůřek & Sojka, 2010) to filter out extremely common or rare terms from our downloaded tweets. In our initial trials of our topic models as classification features, we found that removing such terms from the model-training input resulted in more coherent and discriminative topics. The construction of the primary feature set so far is depicted in Figure 5.1.

In the training phase of our software, we again use MALLET to infer the distribution over topics for the labeled training documents that were supplied in the PAN'15 corpus. This yields a 100-element vector for each single-tweet instance. Those topic vectors are used as inputs to train a classifier for each of the 26 (language, attribute) pairs being predicted for the Author Profiling task. The functionality of the training phase of our software is shown in Figure 5.1.

In the testing phase, we compute the topic distribution vectors of incoming test documents using the same topic model definitions as we did in the training phase. This final phase for extracting our primary feature set is shown in Figure 5.3.



Pre-training Phase

Download additional tweets

- **60,000 tweets collected per language**
- Timeframe covered: April 2014 to April 2015
- External tools: TWORPUS Twitter corpus collector

Pre-process tweets

- Remove near-duplicate tweets (20% of total), anonymize usernames, separate tokens (words), truncate URLs, convert to lowercase
- External tools: CMU Twitter Tokenizer

Build corpus

- Exclude extremely common or rare words from vocabulary
- One tweet ↔ One document in the corpus
- External tools: Gensim library for Python

Train topic model

- **100 distinct topics generated**
- External tools: MALLET topic modeling toolkit

Save topic model

- One *corpus* file: consists of words replaced with numeric IDs
- One *inferencer* file: encodes semantically related words as weighted members of topics

Figure 5.1: Pre-training phase of PAN 2015 Author Profiling software

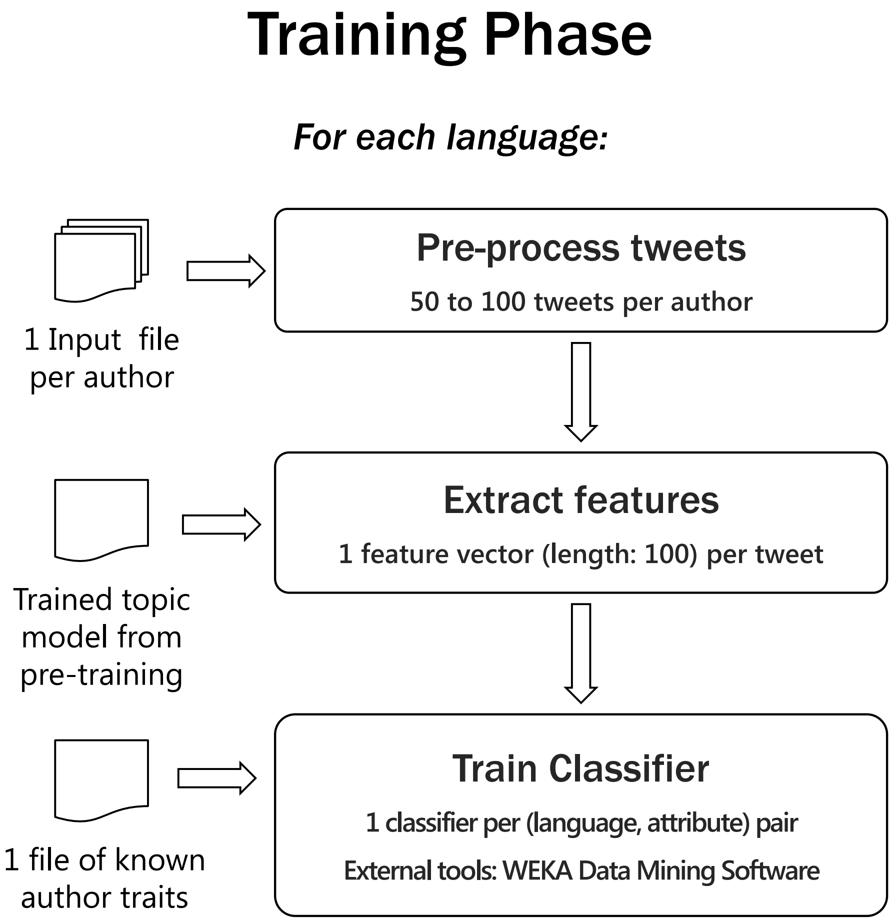

Training Phase
For each language:
1 Input file per author
Pre-process tweets
50 to 100 tweets per author
Trained topic model from pre-training
Extract features
1 feature vector (length: 100) per tweet
1 file of known author traits
Train Classifier
1 classifier per (language, attribute) pair
External tools: WEKA Data Mining Software

Figure 5.2: Training phase of PAN 2015 Author Profiling software

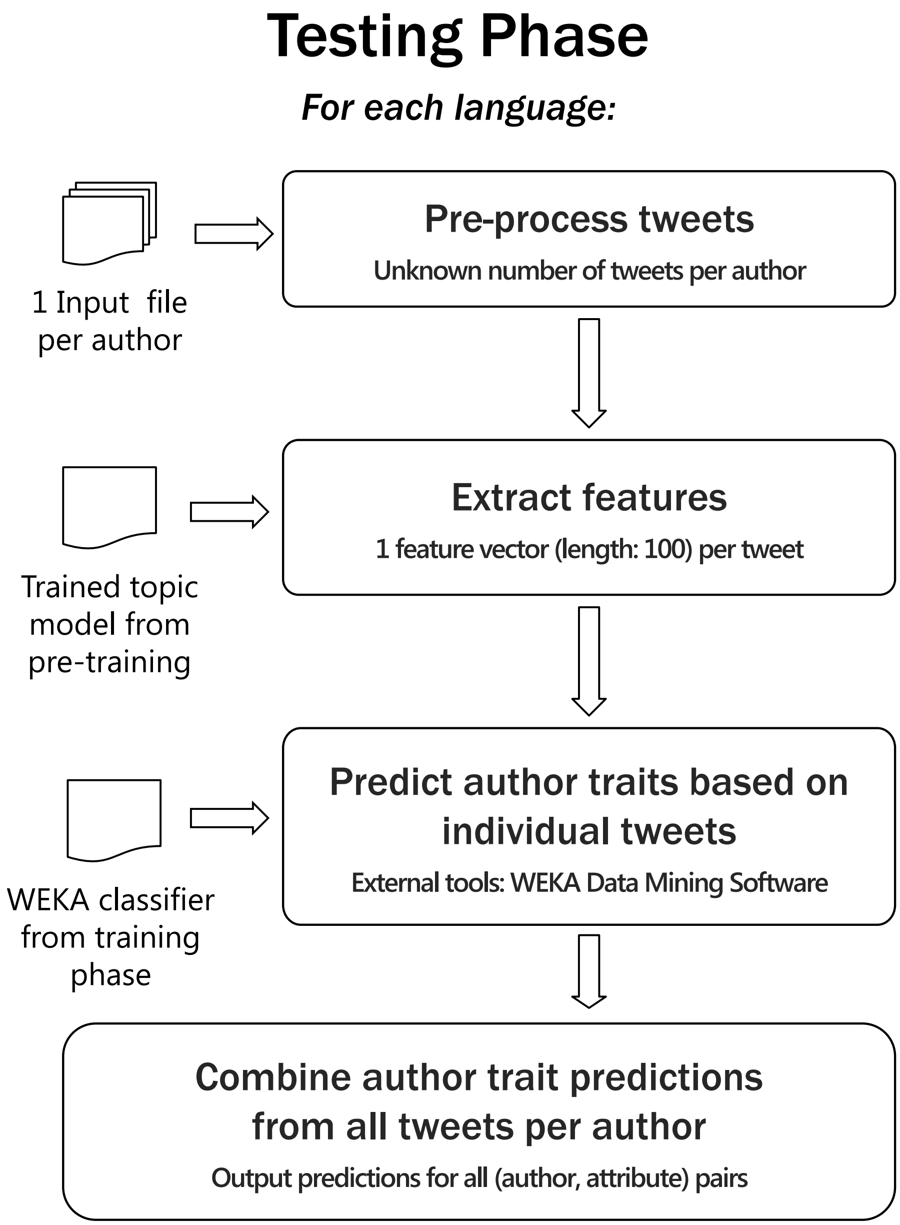

Testing Phase
For each language:
1 Input file per author
Pre-process tweets
Unknown number of tweets per author
Trained topic model from pre-training
Extract features
1 feature vector (length: 100) per tweet
WEKA classifier from training phase
Predict author traits based on individual tweets
External tools: WEKA Data Mining Software
Combine author trait predictions from all tweets per author
Output predictions for all (author, attribute) pairs

Figure 5.3: Testing phase of PAN 2015 Author Profiling software

5.3.1.2 Secondary Feature Set

In order to establish a reference for how well our topic model features performed on the task compared to more conventional methods, we implemented another set of features in the Python programming language which we could evaluate alongside our primary set. We built separate models using the two sets of features, used the same preprocessed data as input, and used the same types of classifiers for nominal and numeric target classes. While some of these features are based on published word lists or clusters derived from exclusively English datasets, the presence of emoticons, hashtags and conveniently universal profanities makes most of them still useful even on the non-English PAN'15 subcorpora. Our secondary feature set is described below:

- **Token count and length**. 3 numeric features: Number of tokens, average number of characters per token, maximum number of characters per token in tokenized tweet text.
- **Special word classes**. 4 numeric features: Proportion of words (tokens) containing at least one non-alphabetic character, proportion of words that are URLs, username mentions, or hashtags.
- **Position-specific special word classes**. 6 binary features: Whether the first or last word is a URL, username mention, or hashtag.
- **Special character classes**. 3 numeric features: Proportion of non-whitespace characters that are punctuation, accented alphabetic characters, or digits 0 through 9.
- **Personality and Gender phrases**. 12 numeric features: From the study of Facebook status updates by Schwartz et al (Schwartz et al., 2013), we combine the 100 most correlated words, phrases and emoticons for high and low values of the five personality traits being predicted, so that 10 features represent the number of such words present in a single tweet and normalized for tweet length. Similar features were created for typically male or female language elements.

- **VADER Sentiment Analysis scores**. 4 numeric features: Computed using the VADER sentiment analysis library (Hutto & Gilbert, 2014). "Positive" and "Negative" sentiment scores range from 0 to 1, and estimate the proportion and intensity of positive and negative words and phrases. "Neutral" indicates the proportion of sentiment-neutral words in the text. "Compound" is a sum of positive and negative scores, normalized to the range [-1, 1].

The VADER engine is fast, accounts for varying degrees of sentiment polarity, and is designed to handle the informal, short messages of social media text. However, because it makes heavy use of English modifiers and negation structures that are context-sensitive, we only use the sentiment analysis features with the English-language subcorpus.

5.3.2 Classification and Prediction

Given the above schemes for feature extraction on the training and test datasets, we use the computed feature vectors as inputs to a classifier created for each (language, attribute) pair by calling the WEKA software package (Hall et al., 2009). This design choice was motivated by the desire for a framework in which we could experiment with a wide variety of methods for classification and regression without making significant modifications to the data processing and file formatting components of our software. Throughout the development period, we were able to observe the effects of other optimizations or design choices, such as those in the feature extraction components, when combined with different types of classification and regression models.

We modeled the gender and age group attributes in the PAN'15 Author Profiling task as discrete classification, or what WEKA calls “nominal class” problems. While we did try using various discretization methods to transform the five personality attributes from real-valued regression to classification problems, we did not find a clear advantage to either approach over all four languages and all five attributes. Thus, for the sake of simplicity in implementation, we treated all personality attributes as “numeric class” problems in WEKA.

For both the nominal and numeric classes, we build one of WEKA's “attribute selection” filters into the classification or regression model at the time it is trained, so that the same subset of “attributes” (which we call “features” in section 5.3.1.1) will be used on the training and test data. Our motivations for applying a feature selection method at this stage of the software are mostly performance-related. It dramatically decreases the time required for training and testing, and keeps the peak RAM utilization safely within the 4 gigabytes allotted to our virtual machine in the TIRA evaluation framework (Gollub et al., 2012), even if the software is to be evaluated on larger datasets in the future.

Our final configuration choices for the classification and regression components of the software are as follows:

- **Nominal attributes** (age group and gender): FilteredClassifier
  - **Filter**: AttributeSelection using CfsSubsetEval with BestFirst forward search
  - **Ensemble method**: RotationForest (Rodriguez et al., 2006) using base classifier REPTree
- **Numeric attributes** (personality traits): FilteredClassifier
  - **Filter**: AttributeSelection using CfsSubsetEval with GreedyStepwise forward search
  - **Ensemble method**: Bagging (Breiman, 1996) using base classifier REPTree

In the testing (prediction) phase of our software, for each (language, attribute) pair, all feature vectors computed for the testing instances are submitted to the trained WEKA model at once, along with the author ID so that the predictions returned by WEKA can be grouped by author. For example, when making predictions for (English, extroverted), if one of the English authors has 100 tweets in the dataset, the predictions from the WEKA classifier will include 100 floating-point predictions of the author’s “extroverted” attribute, ranging from [-0.5, 0.5] as per the Author Profiling task specifications. For numeric attributes (the five personality traits) we take the median value of all the individual predictions. For nominal attributes (age group and gender) we take the discrete class label that occurred most frequently in the individual predictions. This process is conducted once with the WEKA models trained on our primary feature set (document-topic vectors) and again using those trained on the secondary feature set.

The final step in our classification and prediction procedure is to resolve any differences in the predicted values generated by models using the two feature sets. For numeric attributes, we simply take the mean of the two floating-point values. For nominal attributes, we found in cross-validation experiments on the PAN'15 training datasets that the two methods usually agreed. However, in cases where the predictions differed, our primary feature set model was correct more often except on the English-language subcorpora, where the secondary feature set seemed to slightly outperform the topic model features. We suspect this is due to some of the features in our secondary feature set being exclusively used for English data (as in the sentiment analysis features) or based on lexica containing mostly English words. Therefore, when making our final predictions for nominal attributes, we choose to accept the prediction made by the primary feature set model in Spanish, Italian and Dutch; in English, we use the nominal class label predicted by the secondary feature set model.

5.4 Results and Conclusion

Table 5.1 shows the prediction accuracy of our official entry to the PAN'15 Author Profiling task. The columns "Age," "Gender," and "Both" contain the fraction of authors classified correctly, while we list the RMSE for the personality attributes:

Table 5.1: PAN 2015 Author Profiling Competition Performance

Language	**Global**	**Gender**	**Age**	**Both**	**RMSE**	**Agr.**	**Con.**	**Ext.**	**Open**	**Sta.**
English	0.67	0.73	0.72	0.51	0.16	0.15	0.14	0.15	0.15	0.22
Spanish	0.57	0.68	0.50	0.32	0.17	0.17	0.16	0.19	0.14	0.21
Italian	0.70	0.56	–	–	0.15	0.15	0.13	0.13	0.16	0.20
Dutch	0.84	0.81	–	–	0.14	0.15	0.14	0.15	0.09	0.17

We believe we have demonstrated that topic modeling is a promising direction for further research in prediction tasks such as author profiling. Our software achieved accuracy levels at or above average in most subtasks, among roughly 20 participating teams. We see possible avenues of improvement in the construction of our topic models through more informed selection of the LDA parameters, as well as the option of building multiple independent models with different starting conditions and combining the resulting predictions. As for our particular implementation choices, we might be able to improve our accuracy if we devoted more effort to optimizing the classifier training and testing, thus avoiding the need to use feature selection filters beyond what is inherent in the Bagging and RotationForest ensemble methods.

While WEKA was a useful experimental tool for trying different combinations of features and classifier settings, there was some overhead involved in formatting our data in WEKA compatible temporary files and calling the program with its required Java environment from own software written in Python. Now that we have a vision for a successful combination of features based on topic modeling, together with ensemble methods of classification, we plan to further refine these techniques and apply them to other prediction problems in the future.

References

Allport, G. W. & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, 47, 171–220.

Anderka, M. & Stein, B. (2012). Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In *CLEF 2012 Evaluation Labs and Workshop*.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 135–160.

Bazo, A., Burghardt, M., & Wolff, C. (2013). Tworpus - an easy-to-use tool for the creation of tailored twitter corpora. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science* (pp. 23–34). Springer Berlin Heidelberg.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR*, 3, 993–1022.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.

Ferro, N. (2014). *CLEF 15th Birthday: Past, Present, and Future*. Technical Report 2, ACM SIGIR Forum.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. (2013). Recent Trends in Digital Text Forensics and its Evaluation: Plagiarism Detection, Author Identification, and Author Profiling. In *CLEF 2013 PAN*.

Gollub, T., Stein, B., & Burrows, S. (2012). Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In B. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)* (pp. 1125–1126).: ACM.

Guido, G., Peluso, A. M., Capestro, M., & Miglietta, M. (2015). An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values. *Personality and Individual Differences*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1), 10–18.

Hutto, C. & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Weblogs and Social Media*.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2(3), 171–203.

Juola, P. (2012). An Overview of the Traditional Authorship Attribution Subtask: Notebook for PAN at CLEF 2012.

Juola, P. & Stamatatos, E. (2013). Overview of the author identification task at pan 2013. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, CLEF*, (pp. 23–26).

Kudo, T. (2010). Creating an Age Where Anyone Can Find the Information They Truly Need. *NII Today*, (34), 4–7.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Pineda, L. V. (2014). Using intra-profile information for author profiling. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Working Notes for CLEF 2014 Conference, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings* (pp. 1116–1120).: CEUR-WS.org.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Mccollister, C., Huang, S., & Luo, B. (2015). Building Topic Models to Predict Author Attributes from Twitter Messages. *CLEF 2015 Labs and Workshops, Notebook Papers*.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In L. Vanderwende, H. D. III, & K. Kirchhoff (Eds.), *Proceedings of NAACL 2013* (pp. 380–390).: ACL.

Pardo, F. M. R., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daelemans, W. (2014). Overview of the author profiling task at PAN 2014. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Working Notes for CLEF 2014 Conference, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings* (pp. 898–927).: CEUR-WS.org.

Pederson, J. P. (2011). Twitter, Inc. *International Directory of Company Histories*, 118.

Rammstedt, B. & John, O. P. (2007). 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.

Rangel, F., Rosso, P., Koppel, M., & Stamatatos, E. (2013). Overview of the Author Profiling Task at PAN 2013.

Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. In L. Cappellato, N. Ferro, G. Jones, & E. San Juan (Eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*: CEUR-WS.org vol. 1391.

Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 1619–1630.

Romero, E., Villar, P., Gómez-Fraguela, J. A., & López-Romero, L. (2012). Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. *Personality and Individual Differences*.

Sarno, D. (2009). Twitter Creator Jack Dorsey Illuminates the Site's Founding Document, Part I.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9).

Soto, C. J. & John, O. P. (2016). The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model With 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power. *Journal of Personality and Social Psychology*.

Srivastava, S. (2014). Measuring the Big Five Personality Factors.

Steyvers, M. & Griffiths, T. (2006). Probabilistic Topic Models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning* chapter Probabilistic Topic Models. Lawrence Erlbaum Associates.

Weren, E. R. D., Kauer, A. U., Mizusaki, L., Moreira, V. P., Palazzo Moreira de Oliveira, J., & Wives, L. K. (2014a). Examining multiple features for author profiling. *JIDM*, 5(3), 266–279.

Weren, E. R. D., Moreira, V. P., & Palazzo M. de Oliveira, J. (2014b). Exploring information retrieval features for author profiling. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Working Notes for CLEF 2014 Conference, September 15-18, 2014.*, volume 1180 of *CEUR Workshop Proceedings* (pp. 1164–1171).: CEUR-WS.org.