Natural Language Engineering **24** (5): 695–724. © Cambridge University Press 2018
doi:10.1017/S1351324918000244

Multilingual SMS-based author profiling: Data and methods

MEHWISH FATIMA[1], SABA ANWAR[1], AMNA NAVEED[1], WAQAS ARSHAD[2], RAO MUHAMMAD ADEEL NAWAB[1], MUNTAHA IQBAL[3] and ALIA MASOOD[1]

[1]*Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan*
e-mails: {adeelnawab,amnanaveed,mehwish.fatima,sabaanwar}@ciitlahore.edu.pk, aliamasood2020@gmail.com
[2]*Department of Computer Science & IT, Superior University, Lahore, Pakistan*
e-mail: waqas.arshad@superior.edu.com.pk
[3]*Al-Khwarizmi Institute of Computer Science, University of Engineering & Technology, Lahore, Pakistan*
e-mail: muntaha.iqbal@kics.edu.pk

(*Received 5 May 2017; revised 17 May 2018; accepted 17 May 2018; first published online 26 June 2018*)

Abstract

In the recent years, many benchmark author profiling corpora have been developed for various genres including Twitter, social media, blogs, hotel reviews and e-mail, etc. However, no such standard evaluation resource has been developed for Short Messaging Service (SMS), a popular medium of communication, which is very useful for author profiling. The primary aim of this study is to develop a large multilingual (English and Roman Urdu) benchmark SMS-based author profiling corpus. The proposed corpus contains 810 author profiles, wherein each profile consists of an aggregation of SMS messages as a single document of an author, along with seven demographic traits associated with each author profile: gender, age, native language, native city, qualification, occupation and personality type (introvert/extrovert). The secondary aims of this study include the following: (1) annotating the proposed corpus for code-switching annotations at the lexical level (approximately 0.69 million tokens are manually annotated for code-switching) and (2) applying the stylometry-based method (groups of sixty-four features) and the content-based method (twelve features) for gender identification in order to demonstrate how our proposed corpus can be used for the development and evaluation of various author profiling methods. The results show that the content-based character 5-gram feature outperformed all the other features by obtaining the accuracy score of 0.975 and F_1 score of 0.947 for gender identification while using the entire corpus. Furthermore, our proposed corpora (SMS–AP–18 and code-switched SMS–AP–18) are freely and publicly available for research purpose.

1 Introduction

Automatic author profiling is the classification of written text into different classes (e.g. age, gender, etc.) based on the stylistic features chosen by the authors (Rangel

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

et al. 2015). This profiling can be used in different applications, e.g. forensic analysis (Corney *et al.* 2002), marketing intelligence (Glance *et al.* 2005), sentiment analysis and classification (Oberlander and Nowson 2006).

To develop and evaluate new methods for automatic author profiling, some standard evaluation resources are required. Previous studies have focused on developing benchmark corpora for English and other European languages (Peersman, Daelemans and Van Vaerenbergh 2011; Mikros 2012; Wanner 2015; Shrestha *et al.* 2016). The series of PAN author profiling competitions played a vital role in this regard. A set of benchmark corpora of English and other European languages, based on various genres of social media platforms has been developed as an outcome of these competitions (Rangel *et al.* 2013, 2014, 2015, 2016, 2017). However, the author profiling task has not been thoroughly explored for South Asian languages, particularly for Urdu and Roman Urdu.

Urdu, originating from the Persian and Sanskrit, is the lingua franca of Pakistan, but its numerous variations are being used in the South Asian region, particularly in India and Bangladesh. In Pakistan, there are more than 200 million speakers of Urdu[1]. However, Pakistan has been a British Colony since 1857 and after its independence in 1947, English continued to be the official language and the medium of education for upper and middle echelons of society. As a result, English, as a second language, converged with Urdu and gave birth to the popular dialect called Roman Urdu, which uses English alphabets and Urdu semantics. Roman Urdu is the preferred choice for digital communications (Ali and Mahmood Aslam 2012), social media platforms like Twitter (Javed and Afzal 2013; Javed *et al.* 2014; Afzal and Mehmood 2016), Facebook (Fatima *et al.* 2017) and blogs (Mukund and Srihari 2012; Bilal *et al.* 2016). Roman Urdu has developed as a distinctive communication genre because of the ease and comfort it offers to its users who possess average or below average literacies. Furthermore, this particular genre is flexible in nature, as it has absorbed words from other, ethnically diverse languages like Punjabi, Siraiki, etc. The resulting multilingual culture is strongly manifested through the digital communications on social media platforms.

The aim of our work is to broaden current knowledge of multilingual practices in Pakistan. Multilingualism is the 'co-presence of two or more languages in a text, speech or society' (Meylaerts 2010). In broad terms, multilingualism can be defined as the utilization of more than one language simultaneously. More than half of all Europeans claim to speak at least one other language in addition to their native language (Eurobarometer 2006). The past decade has seen a rapid increase in the usage of internet-based technology that resulted in forming a global, multilingual culture; garnering research interest in multilingual corpora (Calvo *et al.* 2017; Fatima *et al.* 2017; Malmasi and Dras 2017). Critics have also analyzed the shifting patterns of language choices, interactional features of multilingualism and distribution of code-switching models (Bhatt and Bolonyai 2011; Ali and Mahmood Aslam 2012; Eleta and Golbeck 2014; Halim and Maros 2014).

[1] `https://en.wikipedia.org/wiki/Urdu` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

This paper examines the short textual interactions on cell phones, which are the most popular medium of interaction, globally and in Pakistan, primarily because of their cost effectiveness. According to Munro and Manning (2012), there is no difference between tweets (posts on Twitter) and mobile SMS (Short Messaging Service). Still, 75% studies of media communications are based on tweets and research on SMS has been overlooked, despite its evident popularity as a medium of communication. A probable reason behind this research limitation can be the data collection process of SMS, which is more challenging than any other social communication genre because of ethical reasons (See Section 3.1 for the challenges faced in generating our proposed corpus).

SMS provides a unique and innovative communicative platform to the modern cellphone users. The 160-character limit on SMS text encourages the use of slang, abbreviations, copula deletion or phonetic spellings during the construction of the syntax, which is already affected by the multilingual background of the users. For instance, in Pakistan, people with English medium educational background are rather proficient in using Roman Urdu and often construct the words according to their personal preferences; educational and social background, etc., which results in varied spellings of the same word. Adding to the complexity of spellings are those multilayered words that are lexically similar and semantically different. For example, bus means a 'vehicle' in English and bus means 'enough' in Urdu.

As discussed earlier that benchmark corpora are needed to develop, evaluate, compare and analyze the author profiling methods for various genres and languages. The previous work on author profiling has been limited to the corpora generated from emails (Corney *et al.* 2002), tweets (Rangel *et al.* 2014), hotel reviews (Rangel *et al.* 2014), blogs (Rangel *et al.* 2013, 2014) and other social media platforms like Facebook (Rangel *et al.* 2014; Fatima *et al.* 2017). The existing corpora are predominantly monolingual which were developed for English and other European languages (Corney *et al.* 2002; Glance *et al.* 2005; Rangel *et al.* 2013, 2014). Thus, these evaluation resources are not helpful for the individualistic SMS genre, specifically for the Pakistani multilingual context. Therefore, to foster the research in this important and widely used medium of digital communication, it is necessary to develop benchmark multilingual corpora based on SMS messages.

With this in mind, our study has designed a benchmark multilingual (English and Roman Urdu), SMS-based author profiling corpus (hereafter called SMS–AP–18 corpus). The proposed multilingual SMS-based corpus is a collection of more than 84,500 SMS messages and consists of 810 author profiles (approximately 104.56 messages per profile on average with a standard deviation of 64.87). One of the biggest challenges of working on SMS text is its length restriction. To resolve this, each profile in our proposed corpus consists of an aggregation of SMS messages as a single document/profile of an author. Each profile considers the following author traits: (1) gender, (2) age, (3) native language, (4) native city, (5) qualification, (6) occupation, and (7) personality type (introvert/extrovert). To demonstrate how our proposed corpus can be used for the development and evaluation of author profiling methods, we applied two methods: (1) stylometry-based method (sixty-four features) and (2) content-based method (twelve features) on our proposed corpus for gender

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

identification task. Furthermore, the SMS–AP–18 corpus is manually annotated for code-switching, which makes it suitable for language shift analysis for future studies. The SMS–AP–18 corpus can also be utilized for the development, evaluation and comparison of existing and new methods for multilingual author profiling, linguistic examination and correlation of author traits (e.g. gender, age), information retrieval and Roman Urdu detection.

This paper is divided into seven sections, including the introduction as Section 1. Section 2 discusses the existing work that has been done on SMS corpora and author profiling in other genres. Section 3 explains the corpus generation process, characteristics and distribution of the proposed corpus. Section 4 describes the code-switching annotations for the proposed corpus. Section 5 outlines the experimental setup that includes author profiling methods, datasets, evaluation methodology and evaluation measures. Section 6 presents the results with the gender identification task. Finally, conclusions are drawn in Section 7.

2 Related work

In the literature, there are many examples of developing benchmark corpora using SMS messages in linguistic and sociolinguistic analysis (Fairon and Paumier 2006; Treurniet *et al.* 2012), machine translation systems (Sridhar *et al.* 2014; Song *et al.* 2014), SMS text normalization (Oliva *et al.* 2013) and spam detection (Giannella, Winder and Wilson 2015). Among the existing SMS-based corpora, NUS SMS corpus[2] is the largest and most widely used one. This corpus was developed primarily to improve the predictive text in mobile devices (How and Kan 2005). Its first version was released in 2004, with 10,000 SMS messages (in English), and the second version came out in 2010 with increased figures. The final corpus release, in 2013, consisted of two sub-corpora in English (41,790 messages) and Chinese (30,020 messages) containing a total of 71,000 SMS messages (Chen and Kan 2013). In the NUS corpus, many profiles were not associated with demographic information because an online survey form collected demographics and many participants had submitted their SMS data without completing the forms.

A developing body of literature has adapted NUS SMS corpus for authorship analysis. Mohan, Baggili and Rogers (2010) experimented on NUS English SMS corpus for author identification by using character *N*-gram approach. The corpus had 4,635 messages and most of the messages were tagged with user ids. For author identification task, at first step, messages were grouped into 132 groups based on author ids. Subsequently, only twenty-eight authors were selected, who had at least fifty messages, which ensued in a reduced corpus of 1,400 messages in English. The character *N*-grams were applied to this reduced dataset and accuracy of 65%–72% was achieved with this system. Silessi, Varol and Karabatak (2016) performed gender identification on a subset of the NUS English corpus, which consisted of 41,002

[2] This corpus has been publicly available since 2004 and is updated on a monthly basis since 2013 `http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/sms/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

SMS messages wherein 28,355 messages were by male authors and 12,647 by female authors. They applied *N*-gram tokenization to the gender classification task and got 64%–72% accuracy approximately.

NUS SMS corpus has also been used for forensic authorship analysis task (Ishihara 2011, 2012, 2014). That corpus contained 38,193 messages by 228 authors from USA, Singapore and India, and 71% of authors were male, 16% were female and 13% were unknown (Ishihara 2011, 2012). The distribution of corpus among native speakers (English), nonnative speakers and unknown were 69%, 30% and 1%, respectively. The researchers applied likelihood ratio for authorship analysis using *N*-gram approach (Ishihara 2011, 2014), vocabulary richness and lexical features (Ishihara 2012), and got the best result of their system in terms of log-likelihood-ratio cost that was 0.46. Ragel, Herath and Senanayake (2013) also utilized NUS SMS corpus for identifying the best experimental conditions for authorship detection/identification. For this purpose, they created a profile of each author treating it as a known author and at the same time creating a similar profile from testing data and treating it as unknown. The researcher applied unigram modelling and similarity score on various sized training and test sets and achieved approximately 85% accuracy. Kretchmar and Zhao (2014) used NUS SMS corpus for author identification by using a combination of tri-gram tokenizer and kernel support vector machine. Interestingly, the researcher achieved an accuracy of 65%–95% considering short and long sized messages.

A number of studies have used combined datasets of SMS and tweets. Kiritchenko, Zhu and Mohammad (2014) experimented on a dataset, composed of Anglophonic tweets, SMS messages and movie reviews, for sentiment analysis. The researchers used data from SemEval-2013[3] competition that consisted of tweet-based training set and two other testing sets in which one was based on tweets and other on SMS messages. The dataset was manually annotated for positive, negative and neutral sentiments in a message as well as for the overall polarity label for the messages. Kiritchenko *et al.* (2014) applied word *N*-grams, noncontiguous *N*-grams, character *N*-grams, all-caps, POS and hash tags, negation feature and sentiment lexicons. Their results show that maximum accuracy of 85.5% was achieved in the message level task. Aboluwarin *et al.* (2016) performed the sentiment analysis on SMS messages and Twitter datasets for digital forensics. They considered three sentiments: positive, negative or neutral emotions and the nondictionary words were used as features along with other sentence level features. These words were normalized before classification and got promising results for classifying the sentiments.

The previous studies have also explored tweet-based datasets for the author profiling task. To date, the PAN[4] competitions provide major contributions of benchmark corpora for identifying different author traits, particularly age and gender, for various languages and genres. In PAN competitions, from 2014 to 2017, one out of four subcorpora consisted of tweets in English and Spanish (Rangel *et al.* 2014), English, Spanish, Italian and Dutch (Rangel *et al.* 2015), and English, Spanish

[3] `https://www.cs.york.ac.uk/semeval-2013/` Last visited: 18-05-2018
[4] `http://pan.webis.de/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

and Dutch (Rangel *et al.* 2016), respectively. All corpora were balanced for gender class. In PAN 2014 competition (Rangel *et al.* 2014), a tweet-based dataset was manually selected and annotated for same genre author profiling. The participants demonstrated author profiling techniques such as stylometry-based (frequencies, punctuations, POS, HTML use, readability measures and many statistics), content-based (bag of words, words *N*-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words, etc.) and the combinations of techniques for their proposed systems. For Twitter, the highest accuracies of gender were 73% for English, and 65% for Spanish, respectively (Rangel *et al.* 2014). In PAN 2015 (Rangel *et al.* 2015), the Twitter corpus was collected in four languages; English, Spanish, Italian and Dutch. The participants explored different stylometry-based features (frequencies, punctuations, POS, verbosity measures and many statistics besides Twitter specific ones such as mentions, hashtags and URLs) and content-based features (bag of words, words *N*-grams, term vectors, TF-IDF *N*-grams, named entities, dictionary words, slang words, ironic words, sentiment words, emotional words) for same genre author profiling (Rangel *et al.* 2015). The highest accuracies for gender were 85.9% for English, 96.6% for Spanish, 86% for Italian and 96.8% for Dutch (Rangel *et al.* 2015).

In 2016, PAN competition task shifted from same genre author profiling to cross-genre author profiling (Rangel *et al.* 2016). The training and testing dataset of PAN 2014 were merged for this year competition. The training was carried out on tweets, and the testing dataset constituted of blogs, social media and hotel reviews in English, Spanish, and Dutch languages (Rangel *et al.* 2016). The participants employed different stylometry-based features (frequencies, punctuations, POS, Twitter-specific elements, readability measures and so forth) and content-based (bag of words, word *N*-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words and so on) for cross-genre author profiling (Rangel *et al.* 2016). The highest accuracies achieved in this task were 74.36% for English, 73.21% for Spanish and 56.80% for Dutch (Rangel *et al.* 2016). In PAN 2017 (Rangel *et al.* 2017), the task was gender and language identification for Twitter data considering four languages: Arabic, English, Spanish and Portuguese. All subcorpora were semiautomatically annotated for gender and language. The participants applied stylistic (frequencies, punctuations, POS, Twitter-specific elements, readability measures), content-based (bag of words, word *N*-grams, term vectors, named entities, dictionary words, slang words, contractions and sentiment words) and deep learning approaches (Recurrent Neural Networks, Convolutional Neural Networks) for this task (Rangel *et al.* 2017). For gender identification, the highest accuracies were 80.31% for Arabic, 82.33% for English, 83.21% for Spanish and 87% for Portuguese (Rangel *et al.* 2017).

Layton, Watters and Dazeley (2010) collected English tweet dataset containing 200 tweets per user by 14,000 different users for author identification. They applied Source Code Authorship Profile (SCAP) approach and structural clues approach. The applied approaches attained the accuracy of 72.9% and 55.5% for SCAP and structural clues respectively. Boutwell (2011) collected the English Twitter dataset of 500 tweets per user from fifty different users. They applied *N*-gram models on the Twitter dataset for authorship classification and reported 40.3% accuracy. Deitrick

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

et al. (2012) collected 36,238 tweets using Twitter Streaming API.[5] Tweets-based author profiles were manually annotated and profiles with undefined gender were removed. The final corpus consisted of 3,031 tweets in English. They applied modified balanced winnow algorithm with a combination of numerous existing features along with unigram and bigram. The results showed 82%–97.98% accuracy with various parameter configurations. Miller, Dickinson and Hu (2012) also performed gender identification on Twitter corpus using *N*-gram approach. They used the Twitter Streaming API[5] to collect 36, 238 unlabelled tweets. Subsequently, these tweets were manually labelled for gender trait. The final corpus consisted of 3,000 English tweets, in which 60% of tweets were written by females and the remaining ones by males. They applied *N*-gram features with six feature selection algorithms and achieved 93%–99% accuracy.

Alowibdi, Buy and Yu (2013) performed gender classification on Twitter data by using colour-based features. The dataset consisted of 53,326 profiles of which 30,898 were male and 22,428 were female and they achieved an accuracy of 65%–71%. Schwartz *et al.* (2013) collected almost 15% of all public tweets in the duration of May 2009 to March 2010 for author identification. They applied character and word *N*-grams along with k-signature and achieved fair improvements in accuracy. Kebede, Tefrie and Sohn (2015) worked on a Twitter dataset of 2,761 tweets for author classification task. They applied bag of words, most frequent words, hashtags and clustered bag of words models on the dataset and attained the accuracy of 53%–100%. Vicente, Batista and Carvalho (2016) experimented on English and Portuguese Twitter corpora for gender classification task. The English dataset contained 65*k* labelled users and the Portuguese 58*k* labelled users. They applied four features excerpted from different sources such as user name and screen name, user description, content of the tweets and profile picture. They reported 93.2% and 96.9% accuracy for English and Portuguese datasets, respectively.

The studies cited in this paper reveal that tweets have been used extensively for author profiling task. Although tweets and SMS messages are similar because its allowed character length (140 for a tweet and 160 for an SMS) is short and these normally contain informal text. However, there are some major differences between SMS and tweet messages, for example, SMS usage is considerably greater than Twitter (see Section 1). Furthermore, SMS is a more popular medium of communication because of its ease of access and user friendly format. Out of the 7.4 billion mobile connections in the world, 3.6 billion connections are associated with basic (nonsmart) phones.[6] Therefore, to communicate via texting, only SMS message is left as an option for users who are not connected through the Internet. In the recent years, there has been an increased use of SMS for authentication, security and confirmation of online accounts (Gmail, Facebook, Banks, etc.).[7] Contrary to the

[5] `https://dev.twitter.com/streaming/overview` Last visited: 18-05-2018

[6] `https://www.apifonica.com/blog/sms-marketing-statistics-2016.html` Last visited: 18-05-2018

[7] `https://thenextweb.com/future-of-communications/2015/02/16/people-still-using-sms-2015/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

SMS mechanics, a broadcast mechanism is applied to tweet messages. Tweets either have a default public setting and anyone with internet access can broadcast them, or there is a privacy restriction for followers only.[8] Moreover, SMS messages are targeted towards smaller audiences and a unicast or multicast can be applied to SMS messages.[9] So, SMS messages can safely be considered as private conversations while tweets are public in nature. People have a tendency to follow the varying conversation pattern for private and public spheres.[10] Consequently, tweets-based datasets cannot be considered as a substitute of SMS-based datasets, despite the fact that collection of tweets-based datasets is quite simple and easy.

To summarize, the available SMS- and Twitter-based corpora are predominantly monolingual and not suitable for multilingual author profiling task. In literature, we found one SMS corpus, which contains multilingual (Roman Urdu and English SMS messages (Ali and Mahmood Aslam 2012) to evaluate the behaviour of code switching. The corpora contains only 1,000 multilingual SMS messages collected from university students (where only gender information is associated with each SMS message). Undoubtedly, this corpus is very small and cannot be used to properly train and test multilingual SMS-based author profiling methods.

Our study contributes a large benchmark multilingual (Roman Urdu and English) SMS-based author profiling corpus (contains more than 84,500 SMS messages and a single profile consists of an aggregation of SMS from one author) along with seven author traits (age, gender, native city, native language, qualification, occupation and personality type). In addition, our proposed corpus is manually annotated for code switching. To the best of our knowledge, such a multilingual (Roman Urdu and English) SMS-based author profiling corpus has not been developed in the yesteryears.

3 Corpus generation process

The main goal of this study is to develop a large benchmark multilingual (Roman Urdu and English) SMS-based author profiling corpus. This section presents the challenges confronted during the accumulation of a large benchmark SMS-based author profiling corpus and discusses the contributors of the corpus, data collection requirements, data collection methodologies, anonymization and standardization process of corpus, the characteristics of proposed corpus and distribution of the author profiles in the proposed corpus.

3.1 Challenges in SMS-based author profiles collection

There are two possible ways to collect genuine SMS text data: (1) gather messages from mobile network operators or (2) ask mobile phone owners to share their data.

[8] https://support.twitter.com/articles/14016 Last visited: 18-05-2018

[9] http://smallbusiness.chron.com/difference-between-twitter-texting-63893.html Last visited: 18-05-2018

[10] http://www.axialent.com/uploads/paper/archivo/Public_and_Private_-Conversations_by_Fred_Kofman.pdf Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

The SMS text is stored in the database of a mobile network operator. However, mobile network operators mostly prefer not to release their database even for the research purpose due to their privacy concerns and customers' care policies. In order to gather the data from phone owners, a restricted access must be ensured because of the obvious privacy reasons. While collecting SMS texts directly from users, the targeted messages are stored on the users' mobile phones, therefore a collection of numerous SMS messages from multiple phone owners requires a lot of effort. Moreover, the private nature of SMS texts is mainly responsible for the scarcity of publicly-accessible SMS-based corpora and makes it a challenging task to gather a large amount of data (Chen and Kan 2013).

3.2 Contributors

For SMS–AP–18 corpus generation, we targeted three types of contributors to ensure the authenticity of the contributed data: (1) friends and family, (2) colleagues and (3) university students. The essential reason for selecting these groups of contributors was that most people are reluctant to share their private data, which is required for this study. Therefore, we organized few presentations and meetings with the target groups to elaborate the nature of research, research goals, data collection methodology and data privacy. In these presentations and meetings, we showed the target contributors our sample profiles and demographics associated with their profiles. The sample profiles were developed by an actual SMS data of the principal investigators of this project.

It is worth noting that all the contributors were volunteers from Pakistan who shared their data without any incentive. Most of the contributors were educated because a contributor using Roman Urdu for SMS messaging must have the knowledge of English as a second language (see Section 1). This data is mainly collected from the various universities of Lahore, which is the capital of Punjab.

3.3 Data collection requirements

For data collection, the contributors were requested to share their own written SMS texts on any topic because the versatility in topics produces realistic and diversified data. The *own written or sent* SMS messages restriction is an intrinsic requirement for modelling the author profiling. We are associating a person's SMS text data with his/her demographic information, therefore, forwarded SMS messages would have produced anomalies in the collected data. The contributors were also requested to provide their demographic traits: (1) gender, (2) age, (3) native city, (4) native language, (5) qualification, (6) occupation and (7) personality type (introvert or extrovert). These author traits will be helpful in developing and evaluating methods for author profiling.

3.4 Data collection methodologies

The common data collection methodologies used in previous studies for collecting SMS messages are as follows: (1) *Manual Transcription*, (2) *Website Collection*, (3)

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Automatic Transcription and (4) *Forwarding* (Chen and Kan 2013). The *Manual Transcription* (or copying messages into a file) method is suitable if the goal is collecting a small amount of data. However, it takes time and effort of the contributors. The *Website Collection* method enables to collect freely available SMS texts from different online websites that provide huge repositories of messages. Yet, it is difficult to trace out the demographic information of the author who created the SMS messages. Therefore, this method is not suitable for collecting the data for author profiling task. The *Automatic Transcription* (or SMS backup through software) is the most convenient and flexible option among the all. This method requires less time and effort in installation of an appropriate software. Moreover, it provides control and traceability to users for filtering out the data (if required). However, this method can be used for smartphone users only who have an advanced to expert knowledge of mobile usage. For simple mobile users, the *Forwarding* method is very effective as it maintains the originality of an SMS message, although it takes time and effort of the forwarding author.

The selection of suitable data collection methodology is important for collecting a realistic and large data. The choice of methodology for corpus generation primarily depends upon the nature of research questions, the nature of data required and how to access the data. The nature of our research is sensitive because we want to predict the author traits from their SMS messages while (as mentioned earlier) SMS messages are exclusively personal data of a contributor. In addition, we need to keep our data collection methodology as simple as possible for the contributors who do not possess a smart-phone or having basic mobile usage skills. According to a study conducted in 2015, only 11% of Pakistani population have smart phones[11]. Therefore, we had to consider numerous contributors, with or without a smart-phone, in order to generate a reasonably large benchmark SMS messages corpus for the author profiling task. Accordingly, we selected *Forwarding* and *Automatic Transcription* methods for constructing the proposed SMS–AP–18 corpus because these methods are well suited to our targeted contributors.

3.5 Guidelines for contributors

Each contributor was requested to share two types of data: (1) SMS texts (the data) and (2) the demographic information (the metadata). We developed some general rules for all contributors, to facilitate their sharing of the data and to ensure the authenticity and quality of collected data: (1) share *sent* or *outgoing* SMS messages only, (2) do not share *forwarded* messages (e.g. poetry, jokes, quotes, marketing ads, etc.), (3) do not share *incoming* messages and (4) do not share any personal or sensitive information, for example name, bank account details, address, e-mail ids, etc.

The procedure of sharing the data (SMS messages) and metadata (demographic information) were different for *Forwarding* and *Automatic Transcription* methods. For sharing data through *Forwarding* method, contributors were asked: (1) to forward

[11] `https://www.techjuice.pk/only-11-percent-of-pakistanis-own-smartphones/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

their SMS messages on a mobile number (which was dedicated to the purpose of data collection only) and (2) to share their demographic information in a single SMS (as the last message). While sharing the data through *Automatic Transcription* method, the contributors were required: (1) to take the back-up of their *sent* SMS messages in a CSV file, (2) add their demographic information in that CSV file or in e-mail content and (3) e-mail the CSV file to us.

We considered all the possible moral and legal implications for the collection of SMS–AP–18 corpus. The most important factor was obtaining the consent of the contributors which was taken via e-mail or SMS message (at the time of data sharing).

3.6 Anonymization and preprocessing of corpus

The author profiles collected from the contributors were anonymized and preprocessed. To ascertain the data anonymity, each contributor's profile is stored with a unique identifier and the demographic information is associated with the given unique identifier only.

After anonymization, the collected data was preprocessed through manual inspection. We performed the following steps for preprocessing of the data: (1) participant's name, mobile numbers and e-mail ids (e.g. 'Regards Ali', 'My Skype id is: AliKhan', etc.) were removed from messages if found, (2) forwarded SMS texts (including jokes, poetry, greetings, etc.) were excluded if found and (3) duplicate SMS texts in a single profile were removed. Duplicate entries could have been sent by the contributor either by mistake or (s)he could have actually typed that SMS more than once for sending to the same or different people. In either case, we kept one text per author.

3.7 Standardization of corpus

After anonymization and preprocessing of the author profiles, the corpus was finalized and converted into the standard XML format (similar to PAN author profiling corpora (Rangel *et al.* 2013, 2014, 2015, 2016, 2017)). Thus, each XML document is an individual profile containing SMS messages (data) of the contributor in standard XML format. Also, there is a text file, *truth file*, containing the demographic information (metadata) of each profile.

3.8 Characteristics of corpus

The proposed SMS–AP–18 corpus contains 810 profiles of total 84,694 SMS messages and 104.56 messages per profile on average with a standard deviation of 64.87. There are total 732,219 words (tokens) and 50,974 word types (unique tokens).

Table 1 shows the fifteen most frequent words in the SMS–AP–18 corpus after removal of English stop words. By observing the Table 1, it can be discovered: (1) the most frequent words are in short forms, (2) almost one third proportion of the frequent words can be considered in both languages and their meaning depends on

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Table 1. *The fifteen most frequent words in SMS–AP–18 corpus after removal of English stop words*

Sr.	Word	Frequency	Language	Meaning/ Form
1	k	10,507	Roman Urdu/ English	En-Ru: Abbreviation of Ok
2	koi	2,209	Roman Urdu	Ru: Anybody or Nothing
3	main	2,129	Roman Urdu/ English	En: Primary, Ru: I or inside
4	Han	2,058	Roman Urdu	Ru: Yes
5	yar	1,981	Roman Urdu	Ru: Pal, Friend
6	kal	1,882	Roman Urdu	Ru: Yesterday or Tomorrow depending on context
7	bhai	1,325	Roman Urdu	Ru: Brother
8	uni	1,296	Roman Urdu/ English	En-Ru: Abbreviation of University
9	time	1,259	English	En: Period of Time
10	mein	1,241	Roman Urdu	Ru: I or inside
11	kam	1,215	Roman Urdu	Ru: Task, work
12	msg	1,214	Roman Urdu/ English	En-Ru: Abbreviation of Message
13	bat	1,133	Roman Urdu/ English	En: animal and playing tool, Ru: Talk, gossip or playing tool similar to English depending on context
14	nahi	1,128	Roman Urdu	Ru: No
15	sath	1,045	Roman Urdu	Ru: With, Along-with, Accompanied

the contextual position in a sentence structure and (3) some Roman Urdu words have different lexical structure but the same meaning, hence reflecting the spelling variation in Roman Urdu language.

If we consider the sentence or message level structure in SMS–AP–18 corpus, an SMS message can be categorized on language basis as follows: (1) complete SMS in English, e.g. '*Rafa is the champ*', (2) complete SMS in Roman Urdu, e.g. '*Bhai na bat tha na baallll*' (Brother there was no ball and bat) and (3) multilingual SMS (contains text in both English and Roman Urdu), e.g. '*Dont know that, mn apne cousin se puchun ga*' (Don't know that, I'll ask my cousin). By this observation, it can be inferred that the educated people in Pakistan normally use a mixed variation of native (Urdu) and second language (English) for their digital communication.

In SMS–AP–18 corpus, each profile consists of an XML document in which all the messages of an author are stored, therefore a profile is constructed of multiple messages. There are total 610 male profiles having total 64,968 messages and 106.50 messages per male profile on average, while there are 200 female profiles holding total 19,726 messages and 98.63 messages per female profile on average. In female profiles, the minimum number of messages in a profile is six and the maximum number of messages in a profile is 336. In male profiles, the minimum number of messages in a profile is one and the maximum number of messages in a profile is 410.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

3.9 Distribution of corpus

Table 2 shows the distribution of SMS–AP–18 corpus profiles based on seven different traits. The number of possible values for each trait is different. We have divided all the possible values of *Age* into three bins: (i) 15–19, (ii) 20–24 and (iii) 25—xx. The *Gender* has distinctly two discrete values: (i) male, and (ii) female. As there was a large range of names of the native cities, we selected the cities with higher values and combined all the cities with lower values in *other* category. Therefore, the *Native City* trait has five categories: (i) Lahore, (ii) Faisalabad, (iii) Sialkot, (iv) Multan and (v) other. Similarly, for native language, we opted for languages having higher values and aggregated lower valued languages in *other* category. So, the *Native Language* has four possible values: (i) Urdu, (ii) Punjabi, (iii) Saraiki, and (iv) other. The *Personality Type* has only two distinct values: (i) introvert (unsocial), and (ii) extrovert (social). The *Qualification* is categorized into five categories on the basis of data found in SMS–AP–18 corpus: (i) intermediate (12 years), (ii) undergraduate, (iii) graduate, (iv) post-graduate, and (v) Ph.D. or Ph.D. scholar. Finally, we divided the *Occupation* into three categories: (i) teacher, (ii) student and (iii) other.

Considering the age trait, the majority of the profiles fall into two age groups: teens (1,519) and early twenties (2,024). The higher representation of young people is probably due to two factors: (1) SMS messaging is very common among young people in Pakistan, and (2) one of three types of the contributors are university students. We can see the dominance of male profiles (610 males (75%) and 200 females (25%)) for gender trait. These statistics may reflect upon the discriminatory gender practices in Pakistani patriarchal society and the lack of education for females that is linked with such practices.

In our data, 50% profiles belong to Lahore and other prominent cities are Sialkot, Faisalabad and Multan. As per expectations, Urdu and Punjabi dominated the native language marker, with Urdu being the national language and Punjabi being the dialect of Punjab (the largest province of Pakistan). Regarding the personality type, 55% profiles are of the introvert and remaining are of extroverts. In education level, the majority of the profiles are fallen in *intermediate* (12 years of education) and *undergraduate*. Similarly, for the occupation, the majority of the profiles are *students*. Our proposed SMS–AP–18 corpus is freely and publicly available for the research purpose.[12]

4 Code-switching annotations for SMS–AP–18 corpus

Multilingualism is known as code-switching in linguistics. It occurs when a speaker flip-flops between two or more languages in the context of a single conversation. Usually, the multilingual speakers utilize the factors of multiple languages when interacting with each other. Thus, the use of more than one linguistic variety in

[12] The SMS–AP–18 corpus is publicly available for the research purpose under the Natural Language Processing (NLP) Group, COMSATS University Islamabad, Lahore Campus, licence. The corpus can be obtained through e-mail (contact details must be provided). Complete details of the license are given in a file along with the corpus.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Table 2. *Distribution of SMS–AP–18 corpus profiles into different traits*

Demographic Trait	Categories	# of Profiles
Age	15–19	292
	20–24	424
	25−xx	94
Gender	Female	200
	Male	610
Native city	Lahore	405
	Faisalabad	23
	Sialkot	54
	Multan	12
	Other	316
Native language	Urdu	393
	Punjabi	378
	Saraiki	20
	Other	19
Personality type	Introvert	445
	Extrovert	365
Education	Intermediate	331
	Undergraduate	364
	Graduate	23
	Postgraduate	86
	Ph.D./Ph.D. scholar	6
Occupation	Student	726
	Teacher	52
	Other	32

a manner, which is consistent with the syntax and phonology of each variety is defined as code-switching. The linguistic researchers are intriguingly studying code-switching for sociocognitive (Bhatt and Bolonyai 2011; Ali and Mahmood Aslam 2012), sociolinguistics analysis (Halim and Maros 2014), analyzing the intersection of languages in social media (Eleta and Golbeck 2014), etc. Our proposed SMS–AP–18 corpus is annotated for code-switching. In the following sections, the annotation guidelines, annotation process and interannotator agreement are described.

4.1 Annotation guidelines

We identified the following eight tags, which can be assigned to tokens in our proposed SMS–AP–18 corpus: (1) Roman Urdu, (2) English, (3) Named Entity, (4) Numeric, (5) Punctuation, (6) Symbol, (7) Expression and (8) Emoticon. We also

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Table 3. *Code-switching annotation rules*

Sr.	Rule	Tag
1	Roman words	Roman Urdu
2	Blended words	Roman Urdu
3	Roman phonetics (Context-based)	Roman Urdu
4	English words used as Urdu words (Context-based)	Roman Urdu
5	English words	English
6	English abbreviations	English
7	English phonetics (Context-based)	English
8	Names and places	Named Entity
9	Date, time or numbers	Numeric
10	Punctuation and its repetition	Punctuation
11	Special characters and its repetition	Symbol
12	Sound expression	Expression
13	Emoticon and patterns	Emoticon

derived three rules for tagging a token into one of the above-mentioned tags. These rules are presented in Table 3.

(1) *Roman Urdu:* The rule 1–4 says that all the words written in Urdu or any other regional language are considered as Roman Urdu. The examples of Urdu-based words are '*Kia*', '*Raat*', '*Din*', '*Subah*' and so on. The blended words are a fusion of Urdu and English in an informal way, e.g. '*Tello*' is a blend of English word Tell and Urdu word Btao and both are equivalent in meaning, '*Eatain*' is a blended word of '*Eat*' and Roman Urdu word '*Khaain*', etc. The word that makes the phonetics of the Roman Urdu word in the context of the sentence is also included in Roman Urdu. The letter and number homophones are also considered in this tag if a word is phonetically Roman Urdu. For the understanding of Roman phonetics, consider the sentence: '*32 chali gai*' that can be translated to English as '*There is a power outage*', here '*32*' denotes to the word '*Batti*' which means power. Moreover, the English words which are written as Urdu words in the context are also considered in this category. As it is mentioned earlier that many English words are part of the Urdu vocabulary as well, sometimes in same meanings and sometimes in different meanings. For example, '*Bus kro*', here Bus is an English word for a vehicle but in the context of this sentence, it is meant enough in Urdu, so it is included in Roman Urdu on the basis of context and sentence structure. Another example of this is '*main jaon?*' where 'main' is an English word, however, it is representing a Roman Urdu word in this sentence.

(2) *English:* The rule 5–7 says that all the words written in English (that are not tagged by above mentioned rules) are considered as English. The example of English-based words are '*Reached*', '*Home*', '*While*' and so on. Abbreviation refers to a short form of a word or phrase, e.g. '*jk*' can be used as descriptor in place of '*just kidding*', '*btw*' can be used as abbreviation in place of '*by the way*', '*DIY*' for '*do it yourself*', etc. The word that making phonetics for English words in the context of the sentence is also included in English. For English phonetics, spellings

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

are considered as the shorter forms of their representation in the standard language. The nonstandard spellings contain at least one nonstandard alphabet to represent the spoken pronunciation of the word. For example, '*ov*' is used in place of '*of*', '*klaim*' as '*claim*' and so on. Letter and number homophones are also included in this tag if the word is phonetically English, e.g. '*c u 2morow*' is abbreviated and homophonic of sentence '*See you tomorrow*'.

(3) *Named Entity*: The names of people and places belong to this category, e.g. '*Ali*' is a masculine name, '*Alina*' is a feminine name, '*Babar Chowk*' is a place name, '*Vehari*' is a city name, '*NB*' is abbreviation of new block in COMSATS, '*C6*' is the room number in block C, '*CIIT or Ciit*' is the abbreviation of COMSATS, and '*UOL*' is abbreviated form of University name, etc. (4) *Numeric:* Date, time and number strings are included in this category. For example, '*5pm*', '*10-May*', '*6/10*', '*3214504500*', etc. (5) *Punctuation:* This denotes the punctuation symbols such as '*?*', '*.*', '*,*','*!*' and so on. The repetition of punctuation marks is also included in this category, e.g. '*enjoying?????*' here '*?????*' is a punctuation mark, etc. (6) *Symbol:* In this, special characters and their repetitions are included such as '$$$$', '#', '@' and so on. (7) *Expression:* Sound expressions or onomatopoeic are distinguishing features of SMS language, which refer to natural or sometimes nonhuman sounds representing expressions, e.g. '*hahaha*' and '*hehehe*' are used to represent laughter and joy. Such sound expressions are tagged as expression. (8) *Emoticon:* It refers to the facial expressions of human beings, e.g. expression for sadness is ': (', smile is ':)', wonder is '*o.O*' or '*:O*', etc. The '*Emoji codes*' of smart phones are also included in this category.

4.2 Annotation process and interannotator agreement

Our proposed SMS–AP–18 corpus was manually annotated by five annotators (A, B, C, D and E) for code-switching. The reason for selecting five annotators is that the corpus is annotated at word level and there are approximately 0.69 million tokens (including punctuation marks, special characters and emoticons) in the corpus. All of the annotators are authors of this study, native speakers of Urdu, having a high level of English proficiency and graduate level of education. The author profiles from 1 to 405 were annotated by annotators A and B, while the author profiles from 406 to 810 were annotated by annotators C and D. The annotations were carried out in two steps. In the first step, a subset of profiles from the corpus was annotated (annotator A and B annotated profiles from 1 to 10, and annotators C and D annotated profiles from 406 to 415) by following the annotation guidelines. After carrying out the initial annotations, the conflicts were discussed by four annotators and annotation guidelines were revised. In the second step, remaining corpus was annotated by four annotators by following the updated annotation guidelines. Then, the conflicting words were annotated by annotator E.

After annotating the whole corpus, Interannotator Agreement (IAA) was computed which was 0.978 out of 1. As it can be noted that the IAA score is very good reflecting that the annotation protocol is very well designed and annotators followed the guidelines substantially during the annotation process. After the

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Table 4. *Examples of code-switching from SMS–AP–18 corpus*

Sr.	SMS message with code-switching tags	Translation of SMS message
1	Sport ⟨*English*⟩ day ⟨*English*⟩ 4 ⟨*Numeric*⟩ bar ⟨*RomanUrdu*⟩ cancel ⟨*English*⟩	Sports day was cancelled four times
2	Suno ⟨*RomanUrdu*⟩ Kaha ⟨*RomanUrdu*⟩ ho ⟨*RomanUrdu*⟩ , ⟨*Punctuation*⟩ reply ⟨*English*⟩ kro ⟨*RomanUrdu*⟩	Listen, where are you, reply me
3	Uski ⟨*RomanUrdu*⟩ 2 ⟨*Numeric*⟩ Year ⟨*English*⟩ Duration ⟨*English*⟩ Ha ⟨*RomanUrdu*⟩ Aur ⟨*RomanUrdu*⟩ 2178 ⟨*Numeric*⟩ Fee ⟨*English*⟩ hay ⟨*RomanUrdu*⟩	Its duration is 2 years and its fee is 2178
4	I ⟨*English*⟩ Am ⟨*English*⟩ At ⟨*English*⟩ Home ⟨*English*⟩ . ⟨*Punctuation*⟩ Ap ⟨*RomanUrdu*⟩ Jag ⟨*RomanUrdu*⟩ Rahy ⟨*RomanUrdu*⟩ Hn ⟨*RomanUrdu*⟩	I am at home. Are you awake?
5	Ahan ⟨*RomanUrdu*⟩ Gud ⟨*English*⟩ Ji ⟨*RomanUrdu*⟩ :* ⟨*Emoticon*⟩ Nd ⟨*English*⟩ Nxt ⟨*English*⟩ Kb ⟨*RomanUrdu*⟩ He ⟨*RomanUrdu*⟩ ? ⟨*Punctuation*⟩	Okay, good :* and when is the next?

code-switching annotations, the corpus was standardized in the XML format. The code-switched manually annotated corpus is freely and publicly available for the research purpose.[13]

Table 4 shows some examples of code-switching annotations from SMS–AP–18 corpus. The second column of the table contains multilingual SMS messages along-with code-switching annotations for each token. For each example, the corresponding English translation is provided in the next column.

5 Experimental setup

This section exhibits how our proposed SMS–AP–18 corpus can be used for the development and evaluation of author profiling techniques. The following sections describe dataset, techniques used for author profiling, evaluation methodology and evaluation measures in detail.

[13] The code-switched SMS–AP–18 corpus is publicly available for the research purpose under the Natural Language Processing (NLP) Group, COMSATS University Islamabad, Lahore Campus, licence. The corpus can be obtained through e-mail (contact details must be provided). Complete details of the licence are given in a file along with the corpus.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

5.1 Dataset

The set of experiments presented in this study are aimed to predict an author's gender from his or her SMS-based profile. In our proposed multilingual SMS–AP–18 corpus, there are two classes for gender trait: (1) male (610 profiles) and (2) female (200 profiles).

As our proposed multilingual SMS–AP–18 corpus is imbalanced in terms of gender trait, therefore a subcorpus from the original dataset is also considered for gender identification task. We selected all the female profiles and first 200 male profiles from the original dataset for the balanced corpus (hereafter called SMS–AP–GB–18 corpus[14]).[15] In SMS–AP–GB–18 corpus, gender trait is balanced: (1) male (200 profiles) and (2) female (200 profiles).

5.2 Author profiling methods

The methods for automatic author profiling can be broadly divided into three categories: (1) stylometry-based method – which aims to capture an author's writing style using different statistical features, (2) content-based method – which aims to identify an author's traits using content of the profile and (3) topic-based method – which utilizes the information about different topics presented in a profile to identify an author's trait(s). We intend to investigate the performance of stylometry-based and content-based methods for gender prediction task on SMS–AP–18 and SMS–AP–GB–18 corpora for this study.

5.2.1 Stylometry-based method

We applied the various groups of sixty-four different features (eleven are lexical or word-based features, forty-seven are character-based features and six are vocabulary richness measures) for the stylometry-based approach. In previous studies, stylometry-based features were used to predict an author's traits, particularly age and gender (Schler *et al.* 2006; Argamon *et al.* 2009; Goswami, Sarkar and Rustagi 2009; Nguyen, Smith and Rosé 2011; Santosh *et al.* 2013; De-Arteaga *et al.* 2013; Przybyła and Teisseyre 2015; Pervaz *et al.* 2015); Fatima *et al.* (2017). Flekova, Ungar and Preotiuc-Pietro (2016) used stylistic- or stylometry-based features for age and income prediction on Twitter data. Soler and Wanner (2016) used stylistic features considering character-based, word-based, sentence-based, dictionary-based and syntactic features for gender identification task.

The first group is composed of the following lexical- or word-based features: (1) average sentence length in characters, (2) count of unique words, (3) count

[14] The SMS–AP–GB–18 corpus is publicly available for research purpose under the Natural Language Processing (NLP) Group, COMSATS University Islamabad, Lahore Campus, licence. The corpus can be obtained through e-mail (contact details must be provided). Complete details of the licence are given in a file along with the corpus.

[15] Note that we also applied the SpreadSubsample filter of Weka for class balancing. The results obtained by using SpreadSubsmaple filter were similar as the results of our manually built SMS–AP–GB–18 corpus. `https://weka.wikispaces.com/Classifying+large+datasets` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

of words, (4) count of sentences, (5) count of paragraphs, (6) average words per paragraph, (7) percentage of question sentences, (8) percentage of words with length 3, (9) percentage of words with length 4, (10) average word length and (11) average sentence length in words.

The second group is consisted of the following character-based features: (1) percentage of semicolons, (2) percentage of punctuation characters, (3) percentage of commas, (4) count of white spaces, (5) number of vertical lines, (6) count of upper case characters, (7) count of multiple question marks, (8) count of multiple exclamation marks, (9) count of underscore, (10) count of tildes, (11) count of at sign, (12) count of semi colons, (13) count of less than sign, (14) count of greater than sign, (15) count of ampersands, (16) count of asterisk, (17) count of question marks, (18) count of dollar sign, (19) count of equal sign, (20) count of digits, (21) count of slashes, (22) count of left parentheses, (23) count of right parentheses, (24) count of left curly braces, (25) count of right curly braces, (26) count of left square brackets, (27) count of right square brackets, (28) count of percentage sign, (29) count of plus sign, (30) count of tabs, (31) count of single quotes, (32) count of full stops, (33) count of exclamation marks, (34) count of ellipsis, (35) count of dashes, (36) count of comma, (37) count of colons, (38) count of brackets, (39) count of apostrophe, (40) count of characters without spaces, (41) count of characters, (42) percentage of upper case letter to N (where N = total no of characters in an author profile), (43) percentage of digits to N, (44) percentage of tabs to N, (45) percentage of special characters to N, (46) percentage of letters to N and (47) percentage of white-spaces to N.

Every text document consists of a group of unique words known as *document vocabulary*. Vocabulary richness functions try to measure the diversity of vocabulary in a given text, i.e. how rich is the vocabulary. The third group is a combination of the following vocabulary richness measures (Stamatatos, Fakotakis and Kokkinakis 2000; Cheng, Chandramouli and Subbalakshmi 2011; Kubát and Milička 2013): (1) yuleK measure, (2) hapax legomena, (3) simpsonD measure, (4) sichelS measure, (5) honoreR measure and (6) brunetW measure.

The fourth group, which is named as all features, we combined all the above mentioned sixty-four features consisting of word-based, character-based and vocabulary-richness-based measures. The fifth and last group is named as the best features, which are obtained by applying Correlation-based Feature Selection Subset Evaluator (Cfs Subset Evaluator)[16] on group of all features. The results of Cfs Subset Evaluator are the following features on SMS–AP–18 corpus: (1) percentage of question sentences, (2) count of tabs, (3) count of left parentheses, (4) percentage of tabs to N and (5) count of Dashes; and on SMS–AP–GB–18 corpus: (1) count of paragraphs, (2) percentage of question sentences, (3) count of tabs, (4) count of left parentheses, (5) percentage of digits to N, (6) percentage of tabs to N and (7) count of brackets.

[16] We used Weka's (Witten, Frank and Hall 2011) implementation of above-mentioned feature selection evaluator. `http://www.cs.waikato.ac.nz/ml/weka/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

5.2.2 Content based method

The content-based method for author profiling is based on *N*-gram features extracted from the text along with feature selection method to reduce the feature space. Mikros and Perifanos (2013) used the *N*-gram approach for author profiling on Greek tweets corpus. Burger *et al.* (2011) applied the word and character *N*-gram approach for gender classification on a multilingual corpus. Nguyen, Smith and Rosé (2011) used the unigram for predicting age. Miller *et al.* (2012) also applied the *N*-gram approach for gender identification task. Ishihara (2012, 2014) used the *N*-grams on NUS SMS corpus for authorship analysis. Layton, Watters and Dazeley (2012, 2013) applied the *N*-gram along with some stylistic features for authorship analysis. Van de Loo, De Pauw and Daelemans (2016), Fatima *et al.* (2017) also applied the word and character *N*-grams for age and gender identification task.

We applied both the word-based and character-based *N*-gram models for this study[17] along with three popular feature selection methods: (1) Information Gain (IG) (Yang and Pedersen 1997), (2) Gain Ratio (GR) (Ikonomakis, Kotsiantis and Tampakas 2005) and (3) Chi Square (CS) (Yang and Pedersen 1997). The length of *N* was varied from 1 to 3 for word-based *N*-gram models, whereas it was varied from 2 to 10 for character-based *N*-gram models.

5.3 Evaluation methodology

The task of identifying an authors' gender is treated as a supervised document classification task. For gender identification, we have binary classification task and goal is to distinguish between two classes: (1) male and (2) female.

We used the ten-fold cross validation for experiments to better estimate the performance of stylometry-based and content-based approaches. Four different machine learning algorithms were explored for the classification task including J48, Random Forest, Support Vector Machine and Naive Bayes.[18] The numeric scores generated by 64 stylometry features are used as input to these classifiers (see Section 5.2.1). For content-based approach, the word and character *N*-gram features extracted from the text are used as input to the machine learning algorithms (see Section 5.2.2). Three feature selection methods: Chi Square, Information Gain and Gain Ratio are used to select the most discriminating features from the set of features generated using content-based methods.

5.4 Evaluation measure

In this study, the evaluation was carried out by using the most widely used evaluation measures, i.e. accuracy and F_1 measures. The accuracy can be expressed as the ratio

[17] We used Weka's (Witten *et al.* 2011) implementation of above mentioned models considering the following parameters: IDFTranform = true, TFTransform = true, MinTermFreq = 3, MultiStopwordsHandler, WordsToKeep = 2,500 and the remaining with default values. `http://www.cs.waikato.ac.nz/ml/weka/` Last visited: 18-05-2018

[18] We used Weka's (Witten *et al.* 2011) implementation of above-mentioned classifiers. `http://www.cs.waikato.ac.nz/ml/weka/` Last visited: 18-05-2018

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

of correctly predicted instances to total instances.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

where TP is defined as number of true positives – correctly classified positive instances, TN is defined as the number of true negatives – correctly classified negative instances, FP is defined as the number of false positives – incorrectly classified negative instances and FN is defined as number of false negatives – incorrectly classified positive instances (Sokolova, Japkowicz and Szpakowicz 2006).

Usually, it is considered that if the accuracy of a model is high then the model is the best fit (Sokolova *et al.* 2006). However, it is not true for all scenarios. The accuracy performs well for symmetrical or balanced datasets, where the values of false positive (FP) and false negatives (FN) are almost same. Therefore, accuracy is not considered as a good measure if data is highly unsymmetrical or imbalanced. As our original dataset is imbalanced for gender class, so considering only the accuracy is not a good measure to evaluate the performance of author profiling techniques (Del Gaudio, Batista and Branco 2014). Therefore, F_1 measure is also considered to evaluate the performance of author profiling methods because it provides the unbiased reflections for skewed data or an uneven class distribution.

The F_1 measure can be defined as the harmonic mean of *Precision* and *Recall*. Therefore, it takes both the false positives (FP) and false negatives (FN) into account. The mathematical formulation of F_1 measure is as follows:

$$F_1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{2}$$

where *Precision* is the proportion of the predicted positive instances that were correct (see Equation (3)), and *Recall* is the proportion of positive instances that were correctly classified (see Equation (4)).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

6 Results and analysis

For experiments, we applied stylometry- and content-based methods with multiple features on SMS–AP–18 and SMS–AP–GB–18 corpora (see Section 5). The results of these experiments are presented in this section. The following terminologies are used for the discussion of the results and table presented below. The 'Feature' means the group of stylometry-based features or the *N*-gram models of word and character for the content-based method which are used for the classification task. The 'CLF' means the machine learning algorithm, which produced the best results (NB means Naive Bayes, RF means Random Forest and SVM means Support Vector Machine). The 'CLF/FSM' denotes the combination of classifier and Feature Selection Method. The feature selection methods are used for selecting the most discriminating features

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Table 5. *Best results of stylometry- and content-based methods*

	SMS–AP–18			SMS–AP–GB–18		
	Stylometery-based method					
Feature	CLF	ACC	F_1	CLF	ACC	F_1
	MCC	0.753		MCC	0.500	
Group of best features	J48,NB	0.864	0.703	J48, RF	0.825	0.811
Group of all features	RF	0.877	0.722	RF	0.850	0.857
Group of char features	**J48**	**0.889**	**0.727**	**RF**	**0.875**	**0.872**
Group of word features	J48	0.840	0.581	RF	0.750	0.737
Group of vocab richness	RF	0.778	0.357	RF	0.725	0.744
	Content-based method					
Features	CLF/ FSM	ACC	F_1	CLF/ FSM	ACC	F_1
Word 1-gram	**NB/ CS, IG, GR**	**0.926**	**0.850**	**RF/ GR**	**0.950**	**0.947**
Word 2-gram	SVM/ CS, IG, GR	0.926	0.833	RF/ GR, NB/ CS, IG	0.925	0.923
Char 5-gram	**SVM/ CS, IG, GR**	**0.975**	**0.947**	**RF,NB/ CS, NB/IG, GR**	**0.950**	**0.947**
Char 6-gram	NB/ CS, IG, GR	0.963	0.930	RF/ CS, IG, GR	0.950	0.950
Char 7-gram	NB/ CS, IG, GR	0.963	0.930	RF/ CS, IG, GR	0.950	0.950
Char 8-gram	SVM/ CS, IG, GR	0.951	0.895	RF/ IG	0.950	0.919

from the feature space (IG refers to Information Gain, GR refers to Gain Ratio and CS refers to Chi Square) because *N*-gram space is usually very huge and expensive in terms of computation. 'ACC' denotes the accuracy score and the 'F_1' denotes the F_1 score obtained by applying a technique on both corpora.

In the classification task, the performance of a machine learning algorithm can be compared with the baseline approach called the Most Common Category (MCC). Using the MCC approach, the accuracy of a machine learning algorithm is computed by assigning the MCC to all the examples in the dataset. The performance with the MCC approach for gender identification task is 0.753 and 0.500 on SMS–AP–18 and SMS–AP–GB–18 corpora, respectively.

Table 5 shows the best results obtained using stylometry- and content-based methods (see Section 5) on both corpora.[19] Inclusively, the highest results are achieved on content-based features ($Accuracy = 0.975$), ($F_1 = 0.947$) with character 5-gram for SMS–AP–18 corpus, and ($Accuracy = 0.950$), ($F_1 = 0.947, 0.950, 0.919$)

[19] We reported the best result here. The detailed results can be viewed from the following link: `https://www.dropbox.com/s/n0k1p9rx9zcdenq/NLE-Result-Tables.xlsx?dl=0`

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

by applying word 1-gram, character 5, 6, 7 and 8-gram for SMS–AP–GB–18 corpus. On SMS–AP–18 corpus, Support Vector Machine classifier produced the best results (using Chi Square, Info Gain and Gain Ratio feature selection methods), whereas on SMS–AP–GB–18 corpus, Random Forest and Naive Bayes classifiers exhibited the highest results (using Chi Square, Info Gain and Gain Ratio feature selection methods).

These results demonstrate that the class balancing has an impact on the performance of author profiling techniques on our proposed corpus. Another interesting point is that MCC of SMS–AP–18 corpus (the distribution of male and female profiles is not balanced) is 0.753 and character 5-gram achieved the accuracy 0.975, showing an improvement of 22.2% to MCC. While the MCC of SMS–AP–GB–18 corpus (the distribution of male and female profiles is balanced) is 0.500 and character 5–8-gram outperformed with accuracy 0.950, attaining a rise of 45% from MCC. Similarly, when we consider the F_1 measure, the highest F_1 scores 0.947 and 0.950 are obtained for SMS–AP–18 and SMS–AP–GB–18 corpora, respectively. The best results are achieved by applying character-based *N*-grams on both corpora. This clearly indicates that in SMS–AP–18 corpus male and female authors prefer different sequences or combinations of characters while typing their SMS messages, which makes it easier for the classifiers to clearly discriminate male profiles from female ones. Moreover, on SMS–AP–GB–18 corpus, the best results are also obtained using word 1-gram, which shows that vocabulary used by female and male authors in their SMS messages in SMS–AP–GB–18 corpus is quite different, which makes the gender identification task easier. Also, this affirms the finding that the class balancing does have an impact on gender identification methods on our proposed corpus. Furthermore, the personality traits and social conducts of males and females are very different in Pakistan. Therefore, the possible reason of high accuracy for gender identification task is the selection of discrete topics in SMS messages and distinct typing style between males and females.

While considering the word *N*-gram models, the highest accuracy (0.926) and F_1 (0.850, 0.833) are obtained with word 1 and 2-gram using Naive Bayes as classifier along with Chi Square, Info Gain and Gain Ratio, Support Vector Machine with Chi Square, Info Gain and Gain Ratio, respectively, for gender identification task on SMS–AP–18 corpus. Overall, the word-based features with Naive Bayes and SVM outperformed the baseline MCC approach. For SMS–AP–GB–18 corpus, the highest accuracy (0.950) and F_1 (0.947) are obtained with word 1-gram using Random Forest as classifier along with Gain Ratio. The word 2-gram also performed very well by achieving ($Accuracy = 0.925$), ($F_1 = 0.923$) once again with the combination of Random Forest with Gain Ratio, and Naive Bayes as classifier along with Chi Square, Info Gain and Gain Ratio.

As the length of SMS messages is very restricted, i.e. 160 characters (discussed in Section 1), so usually people prefer to type in short notations and/or slangs. Therefore, it can be anticipated that author profiling methods would perform better on shorter lengths of character and word *N*-gram models. As expected, in case of word *N*-grams the performance decreases as the length of *N* increases on both corpora. Another possible reason for the low performance would be that as the

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

length of N increases number of N-gram features raises very rapidly and may add noisy features rather than the discriminating ones. Also, it gets difficult for the feature selection methods to identify the set of most discriminating features from a very large feature space. Consequently, it gets convoluted to distinguish male profiles from female ones, and hence the performance decreases. Regarding the length of the character N-grams, the best results are obtained for the medium lengths of N, i.e. 5, 6 and 7. The performance drops for shorter ($N < 5$) and longer ($N > 8$) lengths of N-grams. This variation in performance for both word and character N-grams indicates that length of N has an impact on the performance of author profiling techniques on our proposed corpus.

Regarding machine learning algorithms, the highest accuracy and F_1 results are obtained using Support Vector Machine and Naive Bayes on SMS–AP–18 corpus with content-based methods. On the SMS–AP–GB–18 corpus, the best results are achieved by applying Random Forest and Naive Bayes with content-based methods. So, it can be inferred that overall Naive Bayes seems to be producing the most promising results on both corpora. Regarding the feature selection methods, the best results are obtained on all three methods, i.e. Chi Square, Information Gain and Gain Ratio.

Considering the stylometry method, the best results are obtained using the group of character-based features with classifier J48 ($Accuracy = 0.889$), ($F_1 = 0.727$) on SMS–AP–18 corpus. While on SMS–AP–GB–18 corpus, the highest results are also obtained by the group of character-based features with classifier Random Forest ($Accuracy = 0.875$), ($F_1 = 0.872$). This indicates that character-based features are the most suitable for gender identification task on both SMS–AP–18 and SMS–AP–GB–18 corpora. This also shows that the sequences of characters used by males in SMS messaging are different from those used by females, which makes it easier for the classifiers to discriminate the male profiles from the female ones. The highest results using J48 and Random Forest classifiers highlight the fact that Decision Tree algorithms are suitable with stylometry-based method for gender identification task on our proposed corpora. It can also be noted that performance of group of character-based features approach is significantly higher than baseline approach (MCC) on both corpora. As it can be noted from these results, in all stylometry-based approaches, the accuracy score on SMS–AP–18 corpus is quite high as compared to F_1 score, whereas on SMS–AP–GB–18 corpus there is not much difference in the accuracy and F_1 scores. This again demonstrates that the class balancing has an impact on the performance of gender identification techniques. Considering the MCC of SMS–AP–18 corpus that is 0.753 and character-based features achieved the accuracy of 0.889 making an improvement of 13.6% to MCC. On the other hand, MCC of SMS–AP–GB–18 corpus is 0.500 and character-based features outperformed with the accuracy score of 0.875, attaining a rise of 37.5% from MCC. These findings further strengthen the fact that performance of author profiling method is effected by the distribution of classes in a dataset because there is a significant difference in the results obtained with and without class balancing. Similarly, when we consider the F_1 measure, the highest F_1 scores 0.727 and 0.872 are obtained for SMS–AP–18 and SMS–AP–GB–18 corpora, respectively, which

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

clearly shows that results with the SMS–AP–GB–18 are better than the SMS–AP–18. Generally, when all the features are grouped together and/or the best features are computed, usually the performance of such groups are far better than the individual groups. However, the group of all features (($Accuracy = 0.877$), ($F_1 = 0.722$) on SMS–AP–18, ($Accuracy = 0.850$), ($F_1 = 0.857$) on SMS–AP–GB–18) and group of best features (($Accuracy = 0.864$) on SMS–AP–18, ($F_1 = 0.703$) ($Accuracy = 0.825$), ($F_1 = 0.811$) on SMS–AP–GB–18) have not performed well, highlighting the fact that the set of all features and the set of best features are not useful to boost performance on both corpora. The group of word-based features and group of vocabulary richness also performed low. This is likely to happen because the word-based and vocabulary richness features tend to perform better on large documents, where ample and dense lexicons are present. The SMS messages mostly contain short slangs and abbreviations (as discussed earlier in Section 1), which makes it difficult to get a dense vocabulary representation from the SMS dataset. Consequently, it becomes difficult for vocabulary richness measures to clearly discriminate the male profiles from female ones and thus the performance is dropped. The second possible reason of the low performance would be that SMS messages in SMS–AP–18 corpus contain text in two languages (English and Roman Urdu), which also makes it complex for the vocabulary richness measures to accurately quantify the density of vocabulary in author profiles.

Regarding the stylometry-based method, Random Forest produces the highest accuracy and F_1 scores in the majority of the cases for both corpora among all the machine learning algorithms. This reflects that Random Forest algorithm is the most suitable classifier with stylometry-based features for discriminating author profiles containing multilingual text (Roman Urdu and English) as compared to the other machine learning algorithms used in this study. The possible reason for robust performance of Random Forest is its implicit feature selection while building trees.

It can also be observed from these results, that the highest accuracy and F_1 scores obtained using the best content-based approach are significantly higher than those obtained by using the best stylometry-based approach. This highlights the fact that content of author profiles (content-based method) is more useful in gender identification than the simple statistical features (stylometry-based method) on our proposed corpus. To conclude, the content-based features (character 5-gram on SMS–AP–18 corpus, character 5, 6, 7, 8-gram and word 1-gram on SMS–AP–GB–18 corpus) are the most efficacious features for the gender identification task. Considering all the best results, Random Forest produced the best results in majority of techniques. Regarding the feature selection methods, all three methods (Chi Square, Information Gain and Gain Ratio) performed equally well.

7 Conclusion and future work

In this paper, we presented a benchmark multilingual (Roman Urdu and English) SMS-based author profiles corpus. Our proposed corpus contains 810 profiles with seven demographic traits (age, gender, native language, native area, personality type, qualification and occupation), where each profile comprises an aggregation

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

of SMS messages from a single author. In addition, our proposed corpus is manually annotated for code-switching. To demonstrate how our proposed corpus can be used for the development and evaluation of author profiling methods, we applied the content-based features (word and character *N*-grams) and different groups of sixty-four stylometry-based features for the gender identification task. As the original dataset was imbalanced for gender class, therefore we performed our experimentation on original (imbalanced) dataset named as SMS–AP–18 corpus as well as on a balanced subset of original corpus named as SMS–AP–GB–18 corpus and compared the performance of author profiling methods on both datasets. Our empirical evaluation showed that the best results ($Accuracy = 0.975$), ($F_1 = 0.947$) and ($Accuracy = 0.950$), ($F_1 = 0.947, 0.950, 0.919$) are obtained by using the content-based character 5-gram on SMS–AP–18 (imbalanced) and word 1-gram, character 5–8-gram on SMS–AP–GB–18 (balanced), respectively, for the gender identification task. The potential avenues of future research work are as follows: collecting more author profiles to increase the size of dataset, analyzing the code-switching effect on gender identification task, applying more novel techniques on the proposed corpus, exploring the behaviour of language-dependent features (e.g. Part of Speech) and identification of other author traits other than gender.

Acknowledgment

We would wish to thank Ms. Javaria Farooqui (PhD scholar at the University of Tasmania, Australia) for assistance in proofreading that greatly improved the manuscript. We would also like to show our gratitude to the volunteers of this paper for sharing their personal data with us, and we would also thank our anonymous reviewers for their insights and suggestions, which remarkably aided the manuscript.

References

Aboluwarin, O., Andriotis, P., Takasu, A., and Tryfonas, T. 2016. Optimizing short message text sentiment analysis for mobile device forensics. In *Proceedings of the 12th IFIP WG 11.9 International Conference on Advances in Digital Forensics XII*, New Delhi, India, Springer, pp. 69–87.

Afzal, H., and Mehmood, K. 2016. Spam filtering of bi-lingual tweets using machine learning. In *Proceedings of the 18th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea, IEEE, pp. 710–14.

Ali, I., and Aslam, T. M. 2012. Frequency of learned words of English as a marker of gender identity in SMS language in Pakistan. *Journal of Elementary Education* **22**(2): 45–55.

Alowibdi, J. S., Buy, U. A., and Yu, P. 2013. Language independent gender classification on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM-2013)*, Niagara, Ontario, Canada, ACM, pp. 739–43.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM* **52**(2): 119–23.

Bhatt, R. M., and Bolonyai, A. 2011. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and Cognition* **14**(4): 522–46.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Bilal, M., Israr, H., Shahid, M., and Khan, A. 2016. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, decision tree and KNN classification techniques. *Journal of King Saud University – Computer and Information Sciences*, **28**(3): 330–44.

Boutwell, S. R. 2011. Authorship attribution of short messages using multimodal features. Master's thesis. Naval Postgraduate School Monterey, California.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, United Kingdom, Association for Computational Linguistics, pp. 1301–9.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, **23**(5): 649–85.

Chen, T., and Kan, M. Y. 2013. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation* **47**(2): 299–335.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. 2011. Author gender identification from text. *Digital Investigation* **8**(1): 78–88.

Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. Gender-preferential text mining of E-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference, ACSAC-2002*, Las Vegas, NV, IEEE Computer Society, pp. 282–9.

De-Arteaga, M., Jimenez, S., Mancera, S., and Baquero, J. 2013. Author profiling using corpus statistics, lexicons and stylistic features—notebook for PAN at CLEF-2013. In *Proceedings of CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, Valencia, Spain.

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., and Hu, W. 2012. Gender identification on Twitter using the modified balanced winnow. *Communications and Network* **4**(3): 189–95.

Del Gaudio, R., Batista, G., and Branco, A. 2014. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering* **20**(3): 327–59.

Eleta, I., and Golbeck, J. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior* **41**: 424–32.

Eurobarometer, S. 2006. Europeans and their languages. Technical Report, European Commission.

Fairon, C., and Paumier, S. 2006. A translated corpus of 30,000 French SMS. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, European Language Resources Association (ELRA), pp. 351–4.

Fatima, M., Hasan, K., Anwar, S., and Nawab, R. M. A. 2017. Multilingual author profiling on Facebook. *Information Processing & Management* **53**(4): 886–904.

Flekova, L., Ungar, L., and Preotiuc-Pietro, D. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Giannella, C. R., Winder, R., and Wilson, B. 2015. (Un/Semi-)supervised SMS text message SPAM detection. *Natural Language Engineering* **21**(4): 553–67.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T. 2005. Deriving marketing intelligence from online discussion. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD-2005)*, Chicago, Illinois, USA, pp. 419–28.

Goswami, S., Sarkar, S., and Rustagi, M. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the 3rd International AAAI Conference of Weblogs and Social Media (ICWSM'09)*, San Jose, California, AAAI Press, pp. 214–7.

Halim, N. S., and Maros, M. 2014. The functions of code-switching in Facebook interactions. *Procedia-Social and Behavioral Sciences* **118**: 126–33.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

How, Y., and Kan, M. Y. 2005. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services (HCI-2005)*, Salzburg, Austria.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. 2005. Text classification using machine learning techniques. *WSEAS Transactions on Computers* **4**(8): 966–74.

Ishihara, S. 2011. A forensic authorship classification in SMS messages: a likelihood ratio based approach using *n*-gram. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, Canberra, Australia, pp. 47–56.

Ishihara, S. 2012. A forensic text comparison in SMS messages: a likelihood ratio approach with lexical features. In *Proceedings of the 7th International Workshop on Digital Forensics & Incident Analysis (WDFIA-2012)*, Crete, Greece, pp. 55–65.

Ishihara, S. 2014. A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using n-grams. *International Journal of Speech, Language & the Law* **21**(1).

Javed, I., and Afzal, H. 2013. Opinion analysis of bi-lingual event data from social networks. In *Proceedings of Emotion and Sentiment in Social and Expressive Media (ESSEM 2013) A Workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013)*, Turin, Italy, Citeseer, pp. 164–72.

Javed, I., Afzal, H., Majeed, A., and Khan, B., 2014. Towards creation of linguistic resources for bilingual sentiment analysis of Twitter data. In *Proceedings of the 19th International Conference on Application of Natural Language to Information Systems*, Montpellier, France, Springer, pp. 232–6.

Kebede, A. M., Tefrie, K. G., and Sohn, K. A. 2015. Anonymous author similarity identification. In *Proceedings of the 5th International Conference on IT Convergence and Security (ICITCS-2015)*, Kuala Lumpur, Malaysia, pp. 1–5.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* **50**: 723–62.

Kretchmar, M., and Zhao, Y. 2014. Text message authorship classification using Kernel Support Vector Machines. In *Proceedings of the International Conference on Computational Science and Computational Intelligence (CSCI-2014)*, vol. 2, Las Vegas, Nevada, USA, IEEE, pp. 215–8.

Kubát, M., and Milička, J. 2013. Vocabulary richness measure in genres. *Journal of Quantitative Linguistics* **20**(4): 339–49.

Layton, R., Watters, P., and Dazeley, R. 2010. Authorship attribution for twitter in 140 characters or less. In *Proceedings of the 2nd Cybercrime and Trustworthy Computing Workshop (CTC-2010)*, IEEE, pp. 1–8.

Layton, R., Watters, P., and Dazeley, R. 2012. Recentred local profiles for authorship attribution. *Natural Language Engineering* **18**(03): 293–312.

Layton, R., Watters, P., and Dazeley, R. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering* **19**(1): 95–120.

Malmasi, S., and Dras, M. 2017. Multilingual native language identification. *Natural Language Engineering* **23**(2): 163–215.

Meylaerts, R. 2010. *Multilingualism and translation*. Handbook of translation studies, Amsterdam, John Benjamins Publishing, vol. 1, pp. 227–30.

Mikros, G. K. 2012. Authorship attribution and gender identification in Greek blogs. In *Proceedings of the Methods and Applications of Quantitative Linguistics*, University of Belgrade, Serbia, pp. 21–32.

Mikros, G. K., and Perifanos, K. 2013. Authorship attribution in Greek tweets using author's multilevel n-gram profiles. In *Proceedings of AAAI 2013 Spring Symposium: Analyzing Microtext (SAM-2013)*, Stanford, USA, AAAI Press.

Miller, Z., Dickinson, B., and Hu, W. 2012. Gender prediction on Twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science* **2**(24): 143–8.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Mohan, A., Baggili, I. M., and Rogers, M. K. 2010. Authorship attribution of SMS messages using an n-grams approach. Technical Report, CERIAS2010-11, College of Technology.

Mukund, S., and Srihari, R. K. 2012. Analyzing Urdu social media for sentiments using transfer learning with controlled translations. In *Proceedings of the Second Workshop on Language in Social Media (LSM-2012)*, Montreal, Canada, Association for Computational Linguistics, pp. 1–8.

Munro, R., and Manning, C. D. 2012. Short message communications: Users, topics, and in-language processing. In *Proceedings of the 2nd ACM Symposium on Computing for Development (ACM DEV-2012)*, Atlanta, Georgia, pp. 4:1–10.

Nguyen, D., Smith, N. A., and Rosé, C. P. 2011. Author age prediction from text using Linear Regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH-2011)*, Portland, Oregon, Association for Computational Linguistics, pp. 115–23.

Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions (COLING-ACL-2006)*, Sydney, Australia, pp. 627–34.

Oliva, J., Serrano, J. I., Del Castillo, M. D., and Igesias, A. 2013. A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering* **19**(01): 121–41.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents (SMUC-2011)*, ACM, pp. 37–44.

Pervaz, I., Ameer, I., Sittar, A., and Nawab, R. M. A. 2015. Identification of author personality traits using stylistic features—notebook for PAN at CLEF 2015. In *Proceedings of Evaluation Labs and Workshop – Working Notes Papers (CLEF-2015)*, Toulouse, France. CEUR-WS.org.

Przybyła, P., and Teisseyre, P. 2015. What do your look-alikes say about you? Exploiting strong and weak similarities for author profiling—Notebook for PAN at CLEF 2015. In *Evaluation Labs and Workshop – Working Notes Papers (CLEF-2015)*, Toulouse, France. CEUR-WS.org.

Ragel, R., Herath, P., and Senanayake, U. 2013. Authorship detection of SMS messages using unigrams. In *Proceedings of the 2013 IEEE 8th International Conference on Industrial and Information Systems (ICIIS-2013)*, Sri Lanka, IEEE, pp. 387–92.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. 2013. Overview of the author profiling task at PAN 2013. In *Proceedings of Evaluation Labs and Workshop – Working Notes Papers (CLEF-2013)*, Valencia, Spain.

Rangel, F., Rosso, P., Potthast, M., and Stein, B. 2017. Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs, CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

Rangel, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. 2015. Overview of the 3rd author profiling task at PAN 2015. In *Proceedings of Evaluation Labs and Workshop – Working Notes Papers CLEF-2015*, Toulouse, France. CEUR-WS.org.

Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., and Daeleman, W. 2014. Overview of the 2nd author profiling task at PAN 2014. In *Evaluation Labs and Workshop – Working Notes Papers (CLEF-2014)*, Sheffield, UK. CEUR-WS.org.

Rangel, F., Rosso, P., Verhoeven, B., Daeleman, W., Potthast, M., and Stein, B. 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Proceedings of Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Évora, Portugal, pp. 750–84.

Santosh, K., Bansal, R., Shekhar, M., and Varma, V. 2013. Author profiling: Predicting age and gender from blogs—notebook for PAN at CLEF 2013. In *Evaluation Labs and Workshop – Working Notes Papers (CLEF-2013)*, Valencia, Spain.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, vol. 6, Palo Alto, California, AAAI Press, pp. 199–205.

Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1880–91.

Shrestha, P., Rey-Villamizar, N., Sadeque, F., Pedersen, T., Bethard, S., and Solorio, T. 2016. Age and gender prediction on health forum data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*. European Language Resources Association (ELRA).

Silessi, S., Varol, C., and Karabatak, M. 2016. Identifying gender from SMS text messages. In *Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, California, USA, IEEE, pp. 488–91.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Proceedings of the Australian Conference on Artificial Intelligence*, vol. 4304, pp. 1015–21.

Sokolova, M., and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* **45**(4): 427–37.

Soler, J., and Wanner, L. 2016. A semi-supervised approach for gender identification. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, Portorož, Slovenia, European Language Resources Association (ELRA).

Song, Z., Strassel, S., Lee, H., Walker, K., Wright, J., Garland, J., Fore, D., Gainor, B., Cabe, P., Thomas, T., Callahan, B., and Sawyer, A. 2014. Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland, European Language Resources Association (ELRA).

Sridhar, V. K. R., Chen, J., Bangalore, S., and Shacham, R. 2014. A framework for translating SMS messages. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*, pp. 974–83.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* **26**(4): 471–95.

Treurniet, M., De Clercq, O., Van Den Heuvel, H., and Oostdijk, N. 2012. Collecting a corpus of Dutch SMS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation Conference (LREC-2012)*, Istanbul, Turkey, European Language Resources Association (ELRA), pp. 2268–73.

Van de Loo, J., De Pauw, G., and Daelemans, W. 2016. Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics* **5**(1): 46–60.

Vicente, M., Batista, F., and Carvalho, J. P. 2016. Improving Twitter gender classification using multiple classifiers*. In *Proceedings of the 8th European Symposium on Computational Intelligence and Mathematics (ESCIM-2016)* , Sofia, Bulgaria, pp. 121–7.

Wanner, L. 2015. Multiple language gender identification for blog posts. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pp. 2248–51.

Witten, I. H., Frank, E., and Hall, M. 2011. *Data mining: Practical machine learning tools and techniques* (3rd ed.), The Morgan Kaufmann Series in Data Management Systems, Elsevier Science.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML-1997)*, Nashville, TN, USA, Morgan Kaufmann Publishers Inc., pp. 412–20.

https://doi.org/10.1017/S1351324918000244 Published online by Cambridge University Press