Information Processing and Management 53 (2017) 886–904

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Multilingual author profiling on Facebook

Mehwish Fatima[a,*], Komal Hasan[b], Saba Anwar[a], Rao Muhammad Adeel Nawab[a]

[a] Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan
[b] Department of Computer Science, COMSATS Institute of Information Technology, Vehari, Pakistan

ARTICLE INFO

Article history:
Received 4 July 2016
Revised 17 March 2017
Accepted 27 March 2017
Available online 12 April 2017

Keywords:
Authorship
Author profiling
Multilingual corpus
Facebook
Roman Urdu
Stylometry
N-gram

ABSTRACT

Author profiling is the identification of demographic features of an author by examining his written text. Recently, it has attracted the attention of research community due to it's potential applications in forensic, security, marketing, fake profiles identification on online social networking sites, capturing sender of harassing messages etc. We need benchmark corpora to develop and evaluate techniques for author profiling. Majority of the existing corpora are for English and other European languages but not for under–resourced South Asian languages, like Roman Urdu (written using English alphabets). Roman Urdu is used in daily communication by a large number of native speakers of Urdu around the world particularly in Facebook posts/comments, Twitter tweets, blogs, chat blogs and SMS messaging. The construction of sentences of Urdu while using alphabets of English transforms the language properties of the text. We aim to investigate the behavior of existing author profiling techniques for multilingual text consisting of English and Roman Urdu, concretely for gender and age identification. We here focus on author profiling on Facebook by (i) developing a multilingual (Roman Urdu and English) corpus, (ii) manually building of a bilingual dictionary for translating Roman Urdu words into English, (iii) modeling existing state-of-the-art author profiling techniques by using content based features (word and character N–grams) and 64 different stylistic based features (11 lexical word based features, 47 lexical character based features and 6 vocabulary richness measures) for age and gender identification on multilingual and translated corpora, (iv) evaluating and comparing the behavior of above mentioned techniques on multilingual and translated corpora. Our extensive empirical evaluation shows that (i) existing author profiling techniques can be used for multilingual text (Roman Urdu + English) as well as monolingual text (corpus obtained after translating multilingual corpus using bilingual dictionary), (ii) content based methods outperform stylistic based methods for both gender and age identification task and (iii) translation of multilingual corpus to monolingual text does not improve results.

1. Introduction

Author profiling is a sub task of authorship analysis where objective is to identify the traits of an author (age, gender, native language etc.) by analyzing his written text (Rangel, Rosso, Potthast, Stein, & Daelemans, 2015). In the last decade,

* Corresponding author.
E-mail addresses: mehwish.fatima@ciitlahore.edu.pk, mehwishfatima.raja@gmail.com (M. Fatima), komalhasan@ciitvehari.edu.pk (K. Hasan), sabaanwar@ciitlahore.edu.pk (S. Anwar), adeelnawab@ciitlahore.edu.pk (R.M.A. Nawab).

http://dx.doi.org/10.1016/j.ipm.2017.03.005
0306-4573/

introduction of social media like Facebook, Twitter, blogs etc., has resulted in the evolution of very large collaborative environments. A very serious and important issue in these collaborative environments are fake profiles (or one person can have multiple profiles for fraudulent and other wrong deeds). For example, Facebook has 1.65 billion monthly active users in the 1st quarter of 2016.[1] According to an article published in 2015, number of fake Facebook profiles could be anywhere near to 170 million.[2] These figures are alarmingly high and indicates that there is a need to develop automatic tools and techniques for the detection of fake profiles from different types of texts like Facebook posts/comments, Twitter tweets, blogs posts etc. Moreover, author profiling has potential applications in marketing (Rangel et al., 2015), public sentiments about ongoing government policies (Anstead & O'Loughlin, 2015), election campaign (Caplan, 2013), security and forensic purposes (Juola, 2015) etc.

To develop and evaluate automatic author profiling techniques, we need benchmark corpora in different languages and genres because the nature of text varies from genre to genre. For instance, tweets are short and informal, whereas emails are normally of moderate length and formally written. Therefore, to properly train the author profiling methods, we need standard evaluation resources for different types of genres. In literature, corpora have been developed in various genres, for example, fiction and non–fiction texts (Koppel, Argamon, & Shimoni, 2002), chatlogs (Lin, 2007), customer reviews (Rangel et al., 2014), emails (Estival, Gaustad, Pham, Radford, & Hutchinson, 2007), blogs and social media (Mikros, 2012; Peersman, Daelemans, & Van Vaerenbergh, 2011; Pham, Tran, & Pham, 2009; Rangel, 2013; Schler, Koppel, Argamon, & Pennebaker, 2006), tweets (Burger, Henderson, Kim, & Zarrella, 2011; Nguyen, Gravel, Trieschnigg, & Meder, 2013). Majority of these corpora are in English language, however, some work has been done in European languages as well like Dutch, Italian, Spanish (Rangel, Rosso, Moshe Koppel, Stamatatos, & Inches, 2013; Rangel et al., 2015; Rangel et al., 2014; Wanner, 2015). Litvinova (2014) used Russian, Zhang, Caines, Alikaniotis, and Buttery (2016) considered Chinese and Villegas, Garciarena Ucelay, Errecalde, and Cagnina (2014) used Spanish text for author profiling.

As mentioned above that majority of existing author profiling corpora are available in English and other European languages and these are monolingual. To the best of our knowledge, there is no author profiling corpus available comprising of profiles with multilingual text – Roman Urdu and English. Roman Urdu (along with English) has become a popular and common mode of communication for social media, blogs, tweets, SMS messaging, product reviews etc., in Pakistan and other areas of world where people use Urdu in their daily communication. Roman Urdu is a name used for Urdu language written in Roman script (using English alphabets). For example, Urdu sentence:

"میں نے آج انگلینڈ اور پاکستان کا کرکٹ میچ دیکھا"

will be written in Roman Urdu as "mein ne aaj England aur Pakistan ka cricket match dekha" and in English as "I watched England and Pakistan cricket match today".

1.1. Multilingual settings and Roman Urdu in social media

This world has become a global village due to emergence of Internet and smart devices. People are more intended to connect with others in virtual world (the online world) instead of real world and this thing has affected the society norms as well. Blogs, Tweets, YouTube, Facebook posts and many other social media applications and websites are used as a platform on which people share their thoughts, their ideas and in some manner their routine. Also, these platforms are connecting people of different races, nationalities and indeed distinct origins (of native language). People use a common language (like English) for communication purpose but somehow they have an inclination to their native languages. Even social media platforms are allowing multilingual settings now a days, e-g, Facebook,[3] Blogs,[4] and Twitter.[5] Therefore multilingual text is growing day by day over the Internet.

Researchers are also hitting multilingual settings for research purpose. Zielinski et al. (2012) investigated the multilingual twitter feeds for emergency events for English and under–resourced Mediterranean languages in endangered zones, particularly Turkish, Greek, and Romanian. Abbasi and Chen (2005) demonstrated experiments on Arabic and English based multilingual dataset for authorship analysis. Yarowsky, Ngai, and Wicentowski (2001) carried out experiments on multilingual dataset based on Chinese, French, Czech and Spanish, and aligned it with English. Severyn, Moschitti, Uryupina, Plank, and Filippova (2016) investigated the behaviors of opinions with multilingual settings of Italian and English language on YouTube.

It has been discussed earlier that Roman Urdu is Urdu written using English alphabets, is common language medium over social media and mobile phones where text is written with QWERTY keyboards. A nationally representative sample of men and women from across the Pakistan were asked, "Usually which language do you use for sending SMS from your mobile phone?". Thirty seven percent (37%) said they send SMS in Roman Urdu, 15% use Urdu typed in Urdu alphabets to send text messages whereas 17% said they type SMS in English. Twenty nine percent (29%) do not send any SMS whereas

[1] www.Statista.com : Last visited: 05-01-2017.
[2] http://www.huffingtonpost.com/james-parsons/facebooks-war-continues-against-fake-profiles-and-bots_b_6914282.html Last visited: 05-01-2017.
[3] https://www.facebook.com/help/community/question/?id=10151544094843003 Last visited: 05-01-2017.
[4] https://en.support.wordpress.com/set-up-a-multilingual-blog/ Last visited: 05-01-2017.
[5] https://www.searchenginejournal.com/how-to-manage-twitter-multi-language-accounts/37891/ Last visited: 05-01-2017.

2% gave no response.[6] As it can be noted from these statistics that the most common medium of SMS communication is Roman Urdu, which highlights it's extensive usage in daily life of Pakistani people.

Roman Urdu script is also gaining attention in research trends. Mukund and Srihari (2012) performed the sentiment analysis on Urdu blog data by using structural correspondence learning for Roman Urdu. Javed and Afzal (2013) performed sentiment analysis on bilingual data (English and Roman Urdu) by collecting tweets related to 2013 elections in Pakistan. Javed, Afzal, Majeed, and Khan (2014) extended this work by adding bilingual (Roman Urdu and English) lexicons. Bilal, Israr, Shahid, and Khan (2015) investigated the behavior of multilingual opinions (Roman Urdu and English) extracted from a blog. Daud, Khan, Daud et al. (2014) also worked on multilingual opinion mining (Roman Urdu and English). Afzal and Mehmood (2016) performed spam classification on tweets data based on English and Roman Urdu languages.

1.2. Research objectives

This study aims to develop a benchmark multilingual (Roman Urdu and English) author profiling corpus of Facebook users and to apply existing stylistic and content based approaches on the proposed corpus for automatic detection of an author's trait. The proposed corpus (hereafter called RUEN–AP–17 corpus (Roman Urdu and English–Author Profiling–2017 corpus)) contains Facebook authors profiles along with their demographic information including age, gender, native language, native region, qualification, occupation and personality (introvert or extrovert). To make the corpus more realistic, each author profile contains: (1) both public and private posts/comments of a user and (2) posts/comments typed by a user (shared and forwarded posts/comments are ignored) (see Section 3 for details). We also manually build a bilingual dictionary containing 7749 entries which translates Roman Urdu words to English.

After constructing the RUEN–AP–17 corpus, we generated a bilingual dictionary manually based on RUEN–AP–17 for converting the multilingual dataset into monolingual (English) dataset. We applied stylistic and content based approaches for age and gender identification on both multilingual and translated datasets.

RUEN–AP–17 corpus will be helpful in: (1) developing and evaluating existing state-of-the-art and new methods for author profiling on multilingual text (Roman Urdu and English), (2) fostering research in an under–resourced language like Roman Urdu, (3) identifying potential threats of terrorism since Pakistan is under terrorist attacks from the last decade, (4) detecting potential cases of harassment and fake profiles of Pakistani (and other speakers who use Roman Urdu on social media) Facebook users and (5) developing and evaluating the methods for spelling variations in Roman Urdu text.

Rest of the paper is organized as follows. Section 2 provides an overview of various corpora that have been developed for author profiling. Section 3 presents the corpus generation process. Section 4 is the discussion about existing stylistic and content based approaches that we applied on our proposed corpus. Section 5 describes the experimental setup (dataset, evaluation methodology and evaluation measures). Section 6 presents the results of experiments on proposed corpus, and their analysis. Finally, Section 7 concludes the paper with possible future research work directions.

2. Related work

In recent years, efforts have been made by the research community to develop benchmark corpora for author profiling task. The most prominent effort in this regard is the series of PAN competitions on author profiling (Rangel et al., 2013; Rangel et al., 2015; Rangel et al., 2014). The outcome of these competitions is a set of benchmark corpora for identifying different author traits, particularly age and gender in different languages and geners. The PAN–AP–13 corpus (Rangel et al., 2013) contained author profiles from English and Spanish blog posts for age and gender prediction tasks. For gender prediction, 50% of the profiles were male and remaining were female. For age prediction, profiles were divided into three age groups: 10s (13–17), 20s (23–27) and 30s (33–47). Training dataset contains 236,600 and 75,900 posts for English and Spanish respectively, whereas testing dataset contained 25,440 posts in English and 8160 posts in Spanish. The PAN–AP–14 corpora (Rangel et al., 2014) were developed for four different genres (Hotel Reviews, Tweets, Social Media and Blogs) in English and Spanish. For gender prediction, in all corpora, 50% of the profiles were male and remaining were female. For age prediction, in all corpora, the profiles were divided into five age groups: 18–24, 25–34, 35–49, 50–64 and 65+. Training data of Social Media corpus contained 7746 and 1272 posts in English and Spanish respectively, while testing data contained 3376 and 566 posts in English and Spanish respectively. In Blogs corpus, training data contained 147 and 88 blogs in English and Spanish respectively, whereas testing data contained 78 and 56 blogs in English and Spanish respectively. Training data of Twitter corpus contained 306 and 178 profiles in English and Spanish respectively, and testing data contained 154 and 90 profiles in English and Spanish respectively. Hotel Reviews Corpus was only available in English and comprised of 4160 and 1642 profiles for training and testing respectively. The PAN–AP–15 corpora (Rangel et al., 2015) extended the author profiling task to personality traits detection alongwith age and gender prediction. Four copora were build using Twitter tweets in four different languages including English, Spanish, Italian and Dutch. Each sub-corpus in all four languages was gender balanced and for age identification task, profiles were divided into four age groups: 18–24, 25–34, 35–49 and 50+. Training dataset included 152 English, 110 Spanish, 38 Italian and 34 Dutch profiles. Whereas testing dataset contained 142 English, 88 Spanish, 36 Italian and 32 Dutch profiles. The PAN–AP–16 corpora[7] were built in English, Spanish and Dutch for age and

[6] http://gallup.com.pk/preferred-medium-for-communicating-with-others-via-smspart-5-in-the-5-part-series/ Last visited: 05-01-2017.

[7] http://pan.webis.de/clef16/pan16-web/author-profiling.html Last visited: 05-01-2017.

gender identification. The training dataset contained Twitter tweets, whereas the testing dataset comprised of profiles other than tweets i.e. focus of competition was on cross-genre author profiling task. To conclude, all these PAN corpora contain monolingual author profiles and developed for English and other European languages.

In literature, other efforts have also been made to develop corpora for author profiling. For example, Koppel et al. (2002) experimented on 566 English documents extracted from BNC (British National Corpus) for gender identification using part of speech characteristics and function words. Argamon, Koppel, Fine, and Shimoni (2003) investigated the problem of gender and genre identification on English corpora of 604 documents taken from above mentioned BNC corpus tagged with fiction vs. non–fiction genre and gender. The corpus was comprised of equal number of female and male authors in each genre (123 each in fiction and 179 each in non–fiction). The corpora comprising of blogs data have been extensively targeted for author profiling experimentation. For example, Schler et al. (2006) developed a corpus of 71,493 English blogs to explore how stylistic and content based features can help to identify age and gender of authors. A corpus of Vietnamese weblogs was constructed by Pham et al. (2009). Blogs of 44 female and 29 male bloggers were collected to predict gender, age, location and occupations of bloggers using different document based, character based, lexical, function words and part of speech tag features. Rosenthal and McKeown (2011) built a corpus of 24,500 English blogs for age prediction using three different categories of features including lexical stylistic, lexical content and online behavior. Mikros (2012) investigated author profiling in Greek language using blogs. GBC (Greek Blog Corpus) was built by taking 50 blog entries of 20 bloggers. Wanner (2015) contributed blog corpora in Spanish, Dutch, French, German and Catalan languages for gender and language identification using stylistic features. Shrestha et al. (2016) developed corpus of health forum for age and gender identification. Shrestha et al. (2016) collected data from online health forum comprising on 84,518 profiles. These profiles are categorized in 5 age groups that are 12–17, 18–29, 30–49, 50–64, 65+ in which female gender class was dominant.

Recently, social networks like Twitter and Facebook has got attention of research community for different text mining tasks. Zhang et al. (2016) collected 40,000 posts from Chinese social media (Siana Weibo) users for age prediction of authors considering 4 different age groups. Burger et al. (2011) explored author profiling task for tweets collected in 13 different languages for gender identification. The corpus contained 4,102,434 tweets from 184,000 authors with division of 55% female and 45% male authors. Nguyen et al. (2013) used a corpus of Twitter tweets in Dutch language for age prediction. Alowibdi, Buy, and Yu (2013) developed a Twitter tweets corpus of Dutch language with 53,326 profiles, of which 30,898 were male and 22,428 were female. Rangel (2013) developed a Spanish corpus of 1200 Facebook comments to investigate how human's emotions correlate with their gender. Schwartz et al. (2013) built a Facebook corpus of 75 millions words from 75,000 users (with consent) to predict gender, age and personality traits as a function of words they use in their Facebook status. Park et al. (2015) experimented on personality assessment of author using a corpus of 66,000 Facebook users of same applications. Another Vietnamese corpus consisting of 6831 forum posts collected from 104 authors was developed by Duong, Pham, and Tan (2016) for identification of same traits as used by Pham et al. (2009) by employing stylistic and content features. Ciot, Sonderegger, and Ruths (2013) built a corpus of 8618 Twitter users in four languages French, Indonesian, Turkish, and Japanese for gender prediction. (Volkova & Yarowsky, 2014) developed a Twitter corpora in English and Spanish for gender prediction. Sap et al. (2014) developed an age and gender prediction lexicon from a corpus of 75,394 Facebook users of MyPersonality[8], a third party Facebook application. Verhoeven, Daelemans, and Plank (2016) developed a twitter based corpus containing six different languages that are Dutch, German, Italian, French, Portuguese and Spanish for gender and personality identification. Rangel and Rosso (2013) considered Facebook comments for gender identification.

To conclude, a number of different author profiling corpora have been developed for different genres and languages. The corpora based on social media texts are mostly generated for English and other European languages using publicly available data. Also, profiles in these corpora contain text in one single language. This paper contributes a multilingual (Roman Urdu and English) Facebook author profiles corpus, which contains both public and private comments/posts of Facebook users typed by them. To the extent of our knowledge, such a corpus has not been developed in the past.

3. Corpus generation process

This section describes the process of data collection and the challenges we faced during the data gathering. This section also enlightens our data collection methodology, representatives of our collected data, characteristics and distribution of corpus, and the potential applications of RUEN–AP–17.

3.1. Challenges in Facebook author profiles collection

Facebook monthly serves around 1.65 billion active users.[9] The data (posts, comments, profile information, etc.) of Facebook users are stored under the privacy policy of Facebook[10], which makes it difficult to gather large amount of data even for research purpose (Zimmer, 2010).

[8] http://mypersonality.org Last visited: 05-01-2017.
[9] www.Statista.com : Last visited: 05-01-2017.
[10] Facebook privacy policy is available at this link: https://www.facebook.com/about/privacy/ Last visited: 05-01-2017.

To generate a large benchmark corpus of Facebook author profiles, one possibility is to use an API (or some other tool) to gather users comments/posts and demographic information (Cvijikj & Michahelles, 2011; Farahbakhsh, Han, Cuevas, & Crespi, 2013). However this approach has two main disadvantages. Firstly, we will only get public posts/comments and private ones will be missed, which will not be a true representation of user's personality. Secondly, we will not be sure about the authenticity of user's demographic information, which is very important for the author profile corpus construction. To overcome the above mentioned problems, another approach is to ask Facebook user's to share their data and demographic information to construct a benchmark author profiling corpus. However, this approach is labor intensive and makes it challenging to gather large number of author profiles.

3.2. Data collection methodology

For building RUEN–AP–17 corpus, we asked Facebook users to share their data (both public and private posts/comments to make the dataset more realistic) and their demographic information with us. We held meetings with the targeted groups for data collection – friends and family, university students and colleagues (see Section 3.3 for the details), and explained the purpose of collecting their Facebook profiles and committed assurance of their privacy. These meetings and sessions had a positive impact on the audience and they understood the nature of the research and targeted data. The participants were asked to email the comments/posts along with their demographic information.

For data collection, participants were not restricted to provide data on certain topics. This helped us to assure that the collected data is diversified and it is representative of true and realistic content (Barasa, 2010). The participants were asked to share at least 500 posts/comments collectively in textual format and the minimum length of a single comment/post should be at least five words. They were also asked to submit only unique comments/posts. This helped us to ensure that the collected data is genuine, realistic and diversified. Each participant had to provide the following demographic information along with his comments/posts: (1) age, (2) gender, (3) native language, (4) native city, (5) qualification, (6) personality type (introvert or extrovert) and (7) occupation. This demographic information will help us to develop and evaluate automatic techniques for identifying profile of an anonymous text. Moreover, the collected dataset will also be helpful in socio–linguistic research.

The selection of posts and comments is very important part of the study. The participants were asked to share only those posts/comments that were written by themselves. They were requested to share original stuff without any editing. The restriction of submitting their own posts/comments was applied for two important reasons. First is the ethical reason, the submission of posts/comments of other person is disallowed as the consent of that person is not guaranteed. This may violate the trust and rights of the original user. Secondly, as we are associating a person's posts/comments with his demographic information, therefore, his profile should not contain comments/posts that are not written by him. In addition, this restriction also discards those posts/comments which are not written by the contributor (poetry, jokes, quotes etc.) and are not helpful in capturing the writing style of the author.

3.3. Representatives and ethical considerations

To collect a large corpus while ensuring the authenticity and least dependencies on the participants, three types of participants were targeted for our study: (1) friends and family, (2) colleagues, and (3) university students. Moreover all the participants were volunteers from Pakistan.

The recruitment of above mentioned groups have advantages over the quantitative methods (which often involve anonymous participation) (Fairon & Paumier, 2006). The advantages of collecting data from three different kind of groups are greater likelihood of ensuring authenticity, greater familiarity with participants backgrounds and ability to acquire personal information (How & Kan, 2005). Another advantage of recruiting friends, family and colleagues is to achieve depth (allowing greater understanding of individual's behavior) as well as breadth (Fairon & Paumier, 2006). The university students were selected as target participants because of the popularity and massive use of social media, i.e., Facebook and Twitter among young adults.

The corpus compilation involves consideration of ethical and legal incriminations of collecting and processing data. It involves obtaining participants consent, ensuring their personal information is safely stored and collected data is anonymous. The main ethical consideration that we have emphasized in our study is to protect the interests and rights of our participants. The consents were obtained with two stage method, first is initial informal consent while explaining the data collection methodology to the participant (which was verbal) and secondly through an email (at the time of data sharing by the participant). Since posts/comments and demographic information of participants is their personal data therefore, this information should be anonymous. To assure the privacy of participants, all real identifications of participants has been removed and replaced with the unique identifiers.

There are total 479 Facebook author profiles in the corpus. Each profile is assigned a unique id and stored in a single text file. The demographic information of all the 479 profiles is stored in a separate file (each profile is linked to it's demographic information). The corpus will be freely available under the license for research purposes.[11]

[11] The RUEN-AP-17 Corpus is publicly available for research purposes under the NLPT Group, CIIT Lahore, license. The corpus can be obtained through email (contact details must be provided). Complete details of license are given in a file along with the corpus.

Table 1
Top ten most frequent words in the corpus other than stop words.

Sr. #	Word	Frequency	Language	Meaning
1	bhai	2757	RU	Brother
2	main	2450	EN / RU	EN: Primary, RU: I
3	allah	2366	RU	God
4	yar	2130	RU	Buddy, Pal, Friend
5	happy	2074	EN	Glad
6	nice	1881	EN	Pleasant, Decent
7	han	1814	RU	Yes, Ok
8	day	1618	EN / RU	EN: Day, RU: Give
9	time	1544	EN	Time
10	love	1528	EN	Affection

Table 2
Distribution of corpus.

Trait	Class	#of Profiles	% of Class
Gender	Male	328	68%
	Female	151	32%
Age	xx–19	170	35%
	20–24	218	46%
	25–xx	91	19%
Language	Urdu	157	33%
	Punjabi	269	56%
	Pashto	26	5%
	Others	27	6%
Native area	Punjab	433	90%
	KPK	33	7%
	Sindh	13	3%
Qualification	College	99	21%
	Under graduation	297	62%
	Post graduation	83	17%
Personality	Introvert	157	33%
	Extrovert	322	67%
Occupation	Student	416	87%
	Others	63	13%

3.4. Corpus characteristics and potential applications

The proposed RUEN–AP–17 corpus contains 479 unique profiles. The lexical analysis of the corpus shows that there are 1,032,899 words (tokens) and 453,986 word types (unique tokens) with average 2156 tokens per profile.

Table 1 shows the 10 most frequent keywords in the corpus (excluding the stop words). It can be observed that 4 frequent keywords are in Roman Urdu, 4 are in English and 2 can be used both in English and Roman Urdu depending upon the context. This clearly shows that the Facebook profiles in RUEN–AP–17 are multilingual i.e. comprises of Roman Urdu and English texts. Another interesting fact is that there are words in the corpus which can be either used in English or Roman Urdu with different meanings, for example, the word “day” means “day” in English and “give” in Roman Urdu.

Table 2 shows the distribution of author profiles based on different author traits. For gender, the number of male profiles is higher than female profiles. The possible reason for this is, the female were more hesitant to share their data as compared to male. Considering the age group, the majority of the participants fall into the early twenties (20–24) group. The reason for this high representation of young people is probably due to two factors: first, social networking is more common among young people and second, one of three types of the participants in our corpus are the university students, which may be the reason for a relatively higher percentage of young contributors. For native language, the dominant languages in the corpus are Punjabi and Urdu. As this study is conducted in the province of Punjab where mostly people speak Punjabi language and Urdu is national language of Pakistan, therefore, most of the author’s native language is Punjabi or Urdu. Therefore, the majority of the geographical areas is Punjab. Regarding qualification, majority of the participants are undergraduate. For personality type, the majority contributors are extrovert.

We believe that our proposed RUEN–AP–17 corpus can be used for different applications. Firstly, spelling variants detection of Roman Urdu words, a single Roman word can be written with different spellings and all the variants can be mapped to a standard Roman word. The spelling variant detection system developed using our proposed multi-lingual dataset can be used for standardizing Roman Urdu words for different tasks. For example, Information Retrieval, Machine Translation, Author Profiling, Sentiment Analysis etc. Secondly, generating different bi-directional bilingual dictionaries, for example, translating Roman Urdu words to English and vice versa, and translating Roman Urdu words to Urdu and vice versa. These bilingual dictionaries can be particularly very helpful in: (i) developing Machine Translation systems for Roman Urdu, English and

Urdu languages, and (ii) normalizing multi-/bi-lingual texts into one single language. Thirdly, identification of surmised or fake Facebook profiles who use Roman Urdu and English texts in their posts/comments. Fourthly, identification and blocking of violent/hatred material over Facebook, Twitter, Blogs etc. written in Roman Urdu and English.

4. Methods for author profiling

Methods for predicting an author's traits from his written text can be broadly categorized into three types of methods: (1) stylistic based methods – which aim to capture an author's writing style using different statistical features, (2) content based methods – that intend to identify an author's traits based on the content of the text, and (3) topic based methods – which are used to identify author's traits based on topics discussed in the text. For this study we aim to explore two existing approaches i.e. stylistic based methods and content based methods. The next sections discuss these methods in more detail.

4.1. Stylometry based methods

Every author's writing or typing style has some definite features which (s)he is using consciously or unconsciously (Reddy, Vardhan, & Reddy, 2016; Stamatatos, Fakotakis, & Kokkinakis, 2000). These writing style patterns can be used to identify different traits of an author. In previous studies, stylistic based features have been used to predict an author's traits, particularly age and gender (Argamon, Koppel, Pennebaker, & Schler, 2009; De-Arteaga, Jimenez, Mancera, & Baquero, 2013; Goswami, Sarkar, & Rustagi, 2009; Nguyen, Smith, & Rosé, 2011; Pervaz, Ameer, Sittar, & Nawab, 2015; Przybyła & Teisseyre, 2015; Santosh, Bansal, Shekhar, & Varma, 2013; Schler et al., 2006). Flekova, Ungar, and Preotiuc-Pietro (2016) used stylistic features for age and income prediction on twitter data. Soler and Wanner (2016) used stylistic features that includes character based, word based, sentence based, dictionary based and syntactic features for gender identification. Rangel and Rosso (2016) performed experiments for author profiling on age and gender considering the stylistic features and impacts of emotions.

The stylistic features can be either language dependent (for e.g. POS Tags based stylistic features) and language independent (for e.g. character based stylistic features). For this study, we applied those stylistic features which are language independent because RUEN–AP–17 corpus is multilingual. We are not applying the language dependent features because two languages are mixed in each profile. Another reason is that normally people don't follow grammar rules in their posts/comments. Three types of stylistic features were applied for author profiling including lexical word based features, lexical character based features and vocabulary richness measures.

4.1.1. Lexical word based features

Lexical word based features represent text as a sequence of tokens forming sentences, paragraphs and documents. A token can be numeric number, alphabetic word or a punctuation mark. These tokens are used to get statistics like average sentence length, average word length etc. (Cheng, Chandramouli, & Subbalakshmi, 2011; Stamatatos, 2009). These features have the ability to get insights of a text in any language without special requirements. We have applied 11 lexical word based features: (1) average sentence length in characters, (2) average sentence length in words, (3) average word length, (4) average words per paragraph, (5) number of paragraphs, (6) number of sentences, (7) percentage of question sentences, (8) ratio of words with length 3 (9) ratio of words with length 4, (10) total number of words and (11) total number of unique words.

4.1.2. Lexical character based features

Character based features consider text as a sequence of characters. A number of character based measurements are defined including punctuation count, digit count, character count, colon count, comma count, question mark count, etc. (Cheng et al., 2011; Stamatatos, 2009). Such features are usually present in every language. We have applied 47 character based features: (1) character count, (2) character count without spaces, (3) count of apostrophe, (4) count of brackets, (5) count of colons, (6) count of comma, (7) count of dashes, (8) count of ellipsis, (9) count of exclamation marks, (10) count of full stops, (11) count of question marks, (12) count of semi colons, (13) count of slashes, (14) count of digits, (15) count of ampersands, (16) count of asterics, (17) count of at sign, (18) count of dollar sign, (19) count of equal sign, (20) count of greater than sign, (21) count of less than sign, (22) count of percentage sign, (23) count of plus sign, (24) count of left curly braces, (25) count of left square brackets, (26) count of left parentheses, (27) count of right curly braces, (28) count of right square brackets, (29) count of right parentheses, (30) count of tabs, (31) count of single quotes, (32) count of tilds, (33) count of underscore, (34) number of multiple exclamation marks, (35) number of multiple question marks, (36) number of upper case characters, (37) number of vertical lines, (38) number of white spaces, (39) percentage of commas, (40) percentage of punctuation characters, (41) percentage of semi colons, (42) ratio of white spaces to N (where N = total no of characters in an author profile), (43) ratio of digits to N, (44) ratio of letters to N, (45) ratio of special characters to N, (46) ratio of tabs to N, and (47) ratio of upper case letter to N.

4.1.3. Vocabulary richness features

Every 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. Vocabulary richness is typically measured

as the ratio V/N where V is the size of the vocabulary of the sample text, and N is the number of tokens of the sample text (Cheng et al., 2011; Stamatatos, 2009). Hapax legomena (words appearing once in a text) and dislegomena (words appearing two times in a text) have also been used as vocabulary richness measures (Stamatatos, 2009). Most of the vocabulary richness functions are text length dependent. For this study, we applied 6 vocabulary richness measures which are given below.

$$BrunetWMeasure : W = N^{\nu - .165}$$

$$HapaxLegomena : V_1 = (number\ of\ words\ appear\ exactly\ once)$$

$$HonoreRMeasure : \mathrm{R} = \frac{100 log \mathrm{N}}{1 - (\frac{V_1}{V})}$$

$$SichelSMeasure : S = \frac{V_2}{\mathrm{V}}$$

$$SimpsonDMeasure : \mathrm{D} = \frac{\sum n(\mathrm{n}-1)}{\mathrm{N}(\mathrm{N}-1)}$$

$$YuleKMeasure : \mathrm{K} = \frac{10^4 \ \sum_{i=1}^{\infty} t^2 V_i - N}{N^2}$$

4.2. Content based methods

As text is composed of words, and words consist of characters, so the order of word/character sequences can provide useful information about the content and style of a particular author. In previous studies content based methods have been widely used for author profiling. Schler et al. (2006) and Santosh et al. (2013) used content based technique for author profiling using blogs data for predicting gender and age. For author profiling, Argamon et al. (2009) used content based approach for predicting different demographic features.

4.2.1. N–gram models

The content based methods for author profiling can be based on N–gram features extracted from the text. "An N–gram is an adjacent string of tokens (characters or words). A set of N–grams can be produced by considering text as a string of tokens and moving a sliding window of one token at a time from the start to the end of the string" (Gencosman, Ozmutlu, & Ozmutlu, 2014; Kešelj, Peng, Cercone, & Thomas, 2003). Mikros and Perifanos (2013) used N–gram approach for author profiling on Greek tweets corpus. Burger et al. (2011) used word and character N–gram approach for gender classification on a multilingual corpus. Nguyen et al. (2011) used unigram for predicting the age. Poulston, Stevenson, and Bontcheva (2015) used N–gram approach for the author profiling. Türkoğlu, Diri, and Amasyalı (2007) used word N–grams on Turkish text. Van de Loo, De Pauw, and Daelemans (2016) used word and character N–grams for age and gender identification. For this study, we applied both word based and character based N–gram models. For word based N–gram models the value of N was varied from 1–3, whereas for character based N–gram models it was varied from 2–10.

4.2.2. Feature selection methods

A corpus normally contains a large amount of text and the N–gram models produce a huge feature space (even in hundred thousands and millions). Many classification algorithms are not able to work with such huge feature spaces. To reduce the feature space, feature selection methods are used to select the most discriminating features and remove the redundant or less informative ones (Yang, Liu, Zhu, Liu, & Zhang, 2012; Yang & Pedersen, 1997).

For this study, we applied three existing popular feature selection methods: (1) Information Gain (IG), (2) Gain Ratio (GR) and (3) Chi Square (Chi). Information gain measures the number of bits of information gained about a category by knowing the presence or absence of a term in a document (Yang & Pedersen, 1997). Information Gain examines each feature individually and decides either it should be included in the reduced set of features or not. Gain Ratio measure is a ratio of Information Gain to the intrinsic information (Karegowda, Manjunath, & Jayaram, 2010; Yang et al., 2012). It implies decision tree algorithm for feature selection and prefers to select features with large number of possible values (Karegowda et al., 2010). Chi Square is a feature selection method that uses document frequency threshold (document frequency is the number of documents in which a term occurs). Chi Square measures the lack of independence between document and the category (Yang et al., 2012; Yang & Pedersen, 1997). The formulas for these three feature selection methods are given below.

$$IG(t) = -\sum_{i=0}^{m} P(c_i) log P(c_i) + P(t) \sum_{i=0}^{m} P(c_i \mid t) log P(c_i \mid t) + P(\bar{t}) \sum_{i=0}^{m} P(c_i \mid \bar{t}) log P(c_i \mid \bar{t})$$

(Yang & Pedersen, 1997)

$$GR(t_k, c_i) = \frac{\sum_{c\in\{c_i,\overline{c_i}\}}\sum_{t\in\{t_k,\overline{t_k}\}} P(t,c)log\frac{P(t,c)}{P(t)P(c)}}{-\sum_{c\in\{c_i,\overline{c_i}\}} P(c)logP(c)}$$

(Ikonomakis, Kotsiantis, & Tampakas, 2005)

$$\mathcal{X}^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

(Yang & Pedersen, 1997)

5. Experimental setup

This section describes the experimental setup used for applying stylometery based methods and content based methods on two corpora: (1) multilingual corpus (or RUEN–AP–17 corpus) and (2) translated corpus - obtained after translating Roman Urdu words in the RUEN–AP–17 corpus into English.

5.1. Multilingual corpus

This section describes the experimental configurations and evaluation measures for the multilingual dataset.

5.1.1. Techniques

For stylistic based approach, we applied 64 different features (11 are lexical word based features, 47 are lexical character based features and 6 are vocabulary richness measures (see Section 4.1)). For content based approach, we applied word and character N–grams. For word N–grams the length of N was varied from 1–3, whereas for character N–grams it was varied from 2–10 (see Section 4.2). Note that for content based approaches text was pre–processed by removing all extra spaces and non ASCII characters while including all punctuation marks and special characters.

5.1.2. Dataset

For these experiments, we have focused on two author traits: (1) gender and (2) age. In the RUEN-AP-17 corpus, for gender identification, there are 328 male profiles and 151 female profiles, whereas for age group identification, 170 profiles fall in the age group of *xx*–19, 218 profiles fall in the age group of 20–24 and 91 profiles fall in the age group of 25–*xx*.

5.1.3. Evaluation methodology

The task of identifying an author's gender and age from his/her text is treated as a supervised document classification task. For gender identification, we have binary classification task i.e. goal is to distinguish between two classes: (1) male and (2) female. For age identification, we have multi–classification task i.e. goal is to categorize age among three classes: (1) *xx*–19, (2) 20–24 and (3) 25–*xx*.

We have used 10–fold cross validation to experiment for better estimation of the performance with stylistic based and content based approaches. Four different machine learning algorithms were used for the classification task including J48, Random Forest, SVM and Naive Bayes.[12] The numeric scores generated by 64 stylistic features are used as input to these classifiers (see Section 4.1). For content based approach, the word and character N–grams features extracted from the text are used as input to the machine learning algorithms (see Section 4.2). Three feature selection methods, Chi Square, Information Gain and Gain Ratio (see Section 4.2.2), are used to select the most discriminating features from the set of features generated using content based methods.

5.1.4. Evaluation measures

For experiments presented in this study, evaluation is carried out using two measures: (1) accuracy and (2) Area Under the ROC Curve (AUC). Accuracy is defined as the ratio between total number of correct predictions n_c over total number of predictions n_p.

$$Accuracy = \frac{n_c}{n_p}$$

Area Under the ROC Curve (AUC) is defined as the probability that randomly chosen positive instance is ranked above than the randomly selected negative one (Witten et al., 2011). AUC score is computed using True Positive Rate (TPR) and False Positive Rate (FPR). TPR is defined as the proportion of positive instances that are correctly classified and FPR is defined as the proportion of negative instances that are incorrectly classified as positives.

[12] We used Weka's (Witten, Frank, & Hall, 2011) implementation of above mentioned classifiers. http://www.cs.waikato.ac.nz/ml/weka/ Last visited: 05-01-2017.

Table 3
Top ten most frequent words in the translated corpus.

Sr. #	Word	Frequency
1	No	20,041
2	You	16,491
3	Okay	9919
4	I	9204
5	Come	8009
6	Be	7443
7	Do	7241
8	Surprise	6670
9	Yes	6320
10	Her	6300

5.2. Translated corpus

To evaluate the effect of machine translation on the multilingual author profiling task, we have manually built a bilingual dictionary, which can be used to translate Roman Urdu words into English. To build the bilingual dictionary, main source was the RUEN–AP–17 corpus.[13] A list of unique words was generated using the raw dataset. Total number of unique words extracted were 84,000. Different spelling variants of the same Roman Urdu word were identified manually and grouped together. After that, 7749 most frequent Roman Urdu words were manually translated into English.

Table 3 shows the 10 most frequent keywords in the translated corpus (excluding the stop words). As can be noted that after translation all the ten most freuqent words are in English language, highlighting the fact that bilingual dictionary is useful in translating Roman Urdu words into English.

All the experiments presented in this study, same experimental settings were applied for both multilingual and translated corpora.

6. Results and analysis

For all tables presented in this section, the following terminologies are used. For stylistic based results, “Feature” means the individual stylistic feature that was used for the classification task. The “Classifier” means a machine learning algorithm which produced the best result for an individual feature (NB means Naive Bayes, RF means Random Forest and SVM means Support Vector Machine). For content based results, “Feature” denotes the word/character N–gram feature that was used for the classification task. The “FSM” refers to a Feature Selection Method that was used for selecting the most discriminating features from the feature space (IG refers to Information Gain, GR refers to Gain Ratio and CS refers to Chi Square). The “Accuracy” defines the accuracy of a classifier for a specific feature.

In classification task, the performance of a machine learning algorithm can be compared with the baseline approach called the Most Common Category (MCC). Using MCC, the accuracy of a machine learning algorithm is computed by assigning the most common category to all the examples in the dataset. For both multilingual (RUEN–AP–17 corpus) and translated corpora, performance using the MCC approach is 0.680 and 0.460 for gender and age respectively.

6.1. Results with multilingual corpus

6.1.1. Results using stylometry based methods

Table 4 shows the best results of lexical word based stylistic features for both gender and age.[14] Overall, the best results are obtained when a combination of all the lexical word based features is used for both age ($Accuracy = 0.604$) and gender ($Accuracy = 0.750$). This shows that combination of different word based features is helpful in improving the performance. In case of gender identification task, best result is also obtained using a single stylistic feature (avgSentenceLengthinWords). This shows that in Facebook comments/posts (RUEN–AP–17 corpus), one of the gender prefers longer comments/posts than the other. Interestingly the performance of all the individual features, for both gender and age, is higher than the baseline approach (MCC).

Regarding classifiers, for age prediction, best results are obtained using the Random Forest classifier and for gender prediction, best results are obtained using Naive Bayes and Random Forest classifiers. This indicates that Random Forest classifier is useful when we have a group of features as input for the classification task, whereas Naive Bayes performs well for single feature.

Best results using lexical character based stylistic features are presented in Tables 5 and 6 for both gender and age. Consider that the highest accuracy is achieved when combination of all the lexical character based features is used (*age* :

[13] We also used SMS messages multilingual text (Roman Urdu and English) which was collected in a separate study.
[14] We have only reported the best results among all the classifiers using Weka Experimenter.

Table 4
Results using lexical word based stylistic features on multilingual corpus.

Feature	Gender		Age	
	Classifier MCC	Accuracy 0.680	Classifier MCC	Accuracy 0.460
avgSentenceLengthInCharacters	NB	0.729	NB	0.511
avgSentenceLengthInWords	**NB**	**0.750**	NB	0.532
avgWordLength	NB	0.688	RF	0.479
avgWordsPerParagraph	RF	0.708	NB	0.500
numOfParagraphs	J48	0.688	J48	0.500
numOfSentences	J48	0.688	J48	0.542
percentageOfQSentences	RF	0.708	RF	0.500
ratioOfWordsWithLength3	NB	0.688	RF	0.500
ratioOfWordsWithLength4	NB	0.688	NB	0.542
totalNumberOfWords	J48	0.688	J48	0.479
totalUniqueWords	J48	0.687	RF	0.520
Word-Combined	**RF**	**0.750**	**RF**	**0.604**

Table 5
Results using lexical special character and punctuation marks character based stylistic features on multilingual corpus.

Feature	Gender		Age	
	Classifier MCC	Accuracy 0.680	Classifier MCC	Accuracy 0.460
countApostrophe	RF	0.708	RF	0.542
countBrackets	RF	0.708	J48	0.521
countColon	**RF**	**0.750**	RF	0.521
countComma	J48	0.688	J48	0.500
countDash	RF	0.688	RF	0.583
countEllipsis	RF	0.729	RF	0.532
countExclamation	**RF**	**0.750**	RF	0.458
countFullStop	NB	0.729	RF	0.532
countQMark	NB	0.729	J48	0.521
countSemicolon	RF	0.708	J48	0.500
countSlash	RF	0.708	RF	0.500
numOfAmpersands	RF	0.708	RF	0.479
numOfAsterics	NB	0.708	RF	0.500
numOfAtSign	RF	0.708	NB	0.521
numOfUnderScore	NB	0.688	NB	0.500
numOfEqualSign	RF	0.729	RF	0.500
numOfGreaterThanSign	RF	0.688	RF	0.583
numOfLessThanSign	RF	0.688	RF	0.583
numOfPercentSigns	RF	0.729	RF	0.479
numOfPlusSigns	RF	0.688	RF	0.458
numOfDollarSigns	RF	0.688	RF	0.479
numOfLeftCurlyBraces	RF	0.688	RF	0.468
numOfLeftSquareBrackets	RF	0.688	RF	0.479
numOfLeftParentheses	RF	0.708	RF	0.500
numOfMulExclamationMarks	RF	0.688	RF	0.521
numOfMulQuestionMarks	RF	0.729	RF	0.500
numOfRightCurlyBraces	RF	0.688	RF	0.479
numOfRightSquareBrackets	RF	0.688	RF	0.479
numOfRightParentheses	**RF**	**0.750**	RF	0.542
numOfSingleQuotes	RF	0.708	J48	0.542
numOfTilds	RF	0.688	RF	0.500
numOfVerticalLines	RF	0.688	RF	0.500
%ageOfCommas	NB	0.688	NB	0.500
%ageOfPunctuationChar	J48	0.688	J48	0.479
%ageOfSemiColons	J48	0.688	J48	0.458

Accuracy = 0.688) and (*gender* : *Accuracy* = 0.792). These results are similar to the lexical word based features as the best results are also obtained with combination of features (see Table 4).

Regarding the performance of individual features, for age prediction, the highest results are obtained using a single stylistic feature "ratioOfLettersToN" (*Accuracy* = 0.596) with Random Forest classifier. In gender identification task, the highest accuracy score for individual stylistic feature is obtained on multiple features "countColon", "countExclamation", "numOfRightParentheses" and "numOfUpperCaseChar" (*Accuracy* = 0.750) with Random Forest classifier. This demonstrates the discriminative behavior of simple features like colon, exclamation marks, right parentheses and uppercase characters on

Table 6
Results using lexical non-punctuation character based stylistic features on multilingual corpus.

Feature	Gender		Age	
	Classifier MCC	Accuracy 0.680	Classifier MCC	Accuracy 0.460
characterCount	RF	0.729	J48	0.521
charCountW/OSpaces	NB	0.688	RF	0.553
digitCount	NB	0.688	RF	0.500
numOfTabs	RF	0.688	RF	0.489
numOfUpperCaseChar	**RF**	**0.750**	RF	0.542
numOfWhiteSpaces	RF	0.708	J48	0.500
ratioOfDigitsToN	RF	0.729	RF	0.583
ratioOfLettersToN	NB	0.688	**RF**	**0.596**
ratioOfSpecialCharToN	RF	0.708	RF	0.479
ratioOfTabsToN	RF	0.708	RF	0.521
ratioOfUCaseLettersToN	NB	0.708	NB	0.458
ratioOfWhiteSpacesToN	NB	0.708	J48	0.532
Character-Combined	**RF**	**0.792**	**RF**	**0.688**

Table 7
Results using vocabulary richness stylistic features on multilingual corpus.

Feature	Gender		Age	
	Classifier MCC	Accuracy 0.680	Classifier MCC	Accuracy 0.460
brunetWMeasure	NB	0.688	NB	0.479
hapaxLegomena	NB	0.688	**RF**	**0.583**
honoreRMeasure	RF	0.688	RF	0.542
sichelSMeasure	J48	0.688	RF	0.521
simpsonDMeasure	RF	0.729	RF	0.500
yuleKMeasure	RF	0.688	RF	0.479
Vocabulary-Combined	**RF**	**0.750**	RF	0.511
All Features (Lexical, character and rvocabulary richness)				
All-Combined	**RF**	**0.813**	**RF**	**0.646**

our proposed multilingual RUEN–AP–17 corpus. Again, all the individual character based stylistic features outperform the baseline approach (MCC) for both gender and age.

Regarding classifiers, Random Forest classifier demonstrated eminent results for age and gender prediction. This indicates that Random Forest classifier is beneficial when we have a group of features as input for the classification task.

The results reported in Tables 5 and 6 exhibit that group of lexical character based features (age: $Accuracy = 0.688$; gender: $Accuracy = 0.792$) are more effective compared to the group of lexical word based features (age: $Accuracy = 0.640$; gender: $Accuracy = 0.750$) presented in Table 4.

The best results of vocabulary richness stylistic features are presented in Table 7 for both age and gender. The result of combination of all the 64 stylistic features "All–Combined" is also reported in this table.

For vocabulary richness measures, inclusively best results are attained with a single vocabulary richness feature "hapaxLegomena" for age ($Accuracy = 0.583$). While the combination of all the vocabulary richness features came with the best results for gender ($Accuracy = 0.750$). These results infer that the combination of different features is efficacious in improving the performance of gender identification task. Table 7 reflects that all the individual features of vocabulary richness exhibited improved results than the baseline approach (MCC) for both gender and age. This behavior is consistent in terms of performance while lexical word based features (see Table 4) and lexical character based features (see Tables 5 and 6) are used.

The combination of all features "All–Combined" gives the highest accuracy for gender ($Accuracy = 0.813$) as compared to lexical word based features (see Table 4), character based features (see Tables 5 and 6) and vocabulary richness measures (see Table 7). Regarding the classifiers, once more Random Forest produces the best results. For age identification task, the combination of all features "All–Combined" demonstrated low performance ($Accuracy = 0.646$) compared to the combination of character based stylistic features ($Accuracy = 0.688$).

To conclude, the results reported in Tables 4 and 7 show that group of lexical word based features and group of vocabulary richness features are equally effective in gender identification task. However for age identification task, group of lexical word based features are more effective than group of vocabulary richness features. The results of different types of lexical features (see Tables 4–7) indicate that group of lexical character based features are more effective compared to group of lexical word based and group of vocabulary richness features in both age and gender identification tasks. It is also found that combination of lexical word based, lexical character based and vocabulary richness (All-Combined) feature is most effective

Table 8
Results using content based features on multilingual corpus.

Feature	Gender			Age		
	FSM	Classifier MCC	Accuracy 0.680	FSM	Classifier MCC	Accuracy 0.460
			Word N–gram			
1gram	**IG**	**RF**	**0.875**	IG	RF	0.688
2gram	CS, GR, IG	RF, SVM	0.813	CS, GR, IG	NB	0.708
3gram	CS, GR, IG	SVM	0.854	GR	RF	0.625
			Character N–gram			
2gram	GR	RF	0.833	IG	RF	0.620
3gram	**GR**	**RF**	**0.875**	CS, GR, IG	RF	0.646
4gram	CS, GR, IG	RF, J48	0.833	GR	RF	0.708
5gram	CS, GR, IG	RF	0.854	CS, GR, IG	NB, RF	0.688
6gram	CS, IG	RF	0.854	**IG**	**RF**	**0.729**
7gram	CS, GR, IG	RF, J48	0.854	CS	RF	0.708
8gram	**GR**	**RF**	**0.875**	IG, GR	RF	0.688
9gram	CS, GR, IG	RF, SVM	0.854	IG	RF	0.708
10gram	GR	RF	0.833	GR	RF	0.708

Table 9
Result using group of stylistic features on translated corpus.

Group	Gender		Age	
	Classifier MCC	Accuracy 0.68	Classifier MCC	Accuracy 0.46
All Combined	**NB**	**0.750**	**RF**	**0.625**
Vocabulary richness	RF	0.688	**RF**	**0.625**
Word based	**RF**	**0.750**	RF	0.583
Character based	**RF**	**0.750**	RF	0.604

in predicting gender while only group of character based features produces the highest accuracy for age. Random Forest has proven to be the most efficient classifier for all three groups of stylistic features.

6.1.2. Results using content based methods

Best results for word and character N–grams features both for gender and age are presented in Table 8.[15] For age classification task, the highest accuracy (0.729) is achieved with character 6–gram. We obtained the highest accuracy (0.875) with word unigram, character trigram and 8–gram for gender identification task. This implies that character N-grams are the most discriminating features for age and gender prediction, on our proposed corpus. From Table 8, it is also demonstrated that all the N–grams outperform the baseline approach (MCC) for both age and gender.

Regarding the performance of feature selection methods and classifiers, Information Gain feature selection method and Naive Bayes classifier can be inferred as a good combination for age identification task. For gender prediction, the combination of Information Gain feature selection method and Random Forest classifier has proven to be more efficient than others. This indicates that Random Forest and Naive Bayes classifiers along with Information Gain feature selection method are effective when we have word N–gram features as input for the classification task.

For character N–gram, the combination of Information Gain feature selection method and Random Forest classifier again gave the highest results for age prediction. For gender prediction, we obtained the utmost results using the Gain Ratio feature selection method and Random Forest classifier. This proposes that if we have character N–grams as input for classification task, the combination of Random Forest classifier and Information Gain/Gain Ratio feature selection method is beneficial.

To conclude, from the results presented in Table 8, it is also deduced that character N-gram models are more effective than word N-gram models in age identification task. However word and character N–gram models are equally effective in gender identification task.

6.2. Results with translated corpus

6.2.1. Results using stylometry based methods

Table 9 shows the best results on group of stylistic features for both gender and age.[16] In general, the highest results for translated dataset are obtained with group of all features, group of word based features and group of character

[15] We have only reported the best results among all the classifiers using Weka Experimenter.

[16] We have only reported the best results among all the classifiers using Weka Experimenter. Also, we used the same experimental settings for the translated corpus as we did for the multilingual corpus (see Section 5).

Table 10
Result using content based features on translated corpus.

Feature	Gender			Age		
	FSM	Classifier	Accuracy	FSM	Classifier	Accuracy
		MCC	0.680		MCC	0.460
			Word N–gram			
1gram	CS, GR, IG	SVM	0.729	CS, GR, IG	SVM, J48	0.583
2gram	CS, GR, IG	NB	0.750	**CS, GR, IG**	**NB**	**0.750**
3gram	**CS, GR**	**SVM**	**0.813**	CS, IG	RF	0.625
			Character N–gram			
2gram	CS, GR, IG	J48	0.708	CS, GR, IG	NB, J48, RF	0.521
3gram	GR	RF	0.745	CS, GR, IG	SVM	0.563
4gram	CS	RF	0.708	CS, GR, IG	J48	0.563
5gram	GR	RF	0.750	CS, GR, IG	NB,RF	0.625
6gram	CS, GR, IG	NB, SVM	0.729	CS, GR, IG	NB, SVM, RF	0.542
7gram	CS, GR, IG	RF, SVM	0.723	CS, GR, IG	SVM	0.604
8gram	CS, IG	J48	0.729	GR	SVM	0.542
9gram	CS, GR, IG	NB,SVM	0.729	CS, IG, GR	NB, J48	0.563
10gram	CS, IG	J48	0.771	CS, IG, GR	SVM	0.479

Table 11
Comparison of results obtained using group based stylistic features for multilingual and translated corpora.

Group	Gender						Age					
	Multilingual			Translated			Multilingual			Translated		
	CLF	ACC	AUC	CLF	ACC	AUC	CLF	ACC	AUC	CLF	ACC	AUC
	MCC	**0.680**		**MCC**	**0.680**		**MCC**	**0.460**		**MCC**	**0.460**	
All	**RF**	**0.813**	**0.782**	NB	0.750	0.656	RF	0.646	0.826	RF	0.625	0.687
Vocab	RF	0.750	0.589	RF	0.688	0.517	RF	0.511	0.665	RF	0.625	0.681
Word	RF	0.750	0.593	RF	0.750	0.640	RF	0.604	0.681	RF	0.583	0.714
Char	RF	0.792	0.828	RF	0.750	0.693	**RF**	**0.688**	**0.795**	RF	0.604	0.759

based features for gender (*Accuracy* = 0.750) and group of word based features and group of vocabulary richness for age (*Accuracy* = 0.625). There is an intriguing point that these highest results are lower than the group based stylistic results of multilingual dataset except the results of group of vocabulary richness features (see Tables 4–7). The possible reason for lower results on translated dataset compared to multilingual is that reasonable amount of stylistic information is lost when Roman Urdu words are translated into English because Roman words with different spelling variants are translated to same English word (see Section 5.2). Interestingly the performance of all the groups of features is higher than the baseline approach (MCC) for both age and gender. Once more the Random Forest outperforms among all the classifiers.

6.2.2. Results using content based methods

The best results for word and character N–grams features both for gender and age are presented in Table 10. Comprehensively the highest results for translated dataset are obtained with Gain Ratio and Chi Square feature selection methods and SVM as classifier using word trigram for gender (*Accuracy* = 0.813). While the highest result for age is obtained on word bigram using Chi Square, Information Gain and Gain Ratio with Naive Bayes (*Accuracy* = 0.750). Intriguingly the highest results for gender is lower than the results of multilingual dataset (see Table 8) for content based approach. While for age, there is slight improvement in accuracy on translated data (see Table 8) as compare to the results of multilingual dataset. This demonstrates that machine translation is not very assistive for improving the performance of author profiling methods on multilingual text (Roman Urdu and English).

The performance of all the content based features is higher than the baseline approach (MCC) for both age and gender. Regarding the feature selection methods and classifiers, almost all feature selection methods performed well when used with SVM, Naive Bayes, J48 and Random Forest for age identification task. For gender prediction, the combination of Chi Square and Gain Ratio feature selection methods and SVM, Random Forest and Naive Bayes classifiers have proven to be more efficient.

6.3. Comparison of results

Table 11 demonstrates the group based results of stylistic features on multilingual and translated dataset for both gender and age. The “Group” refers to group of stylistic features. The “CLF” refers to machine learning algorithm used for the purpose of classification. “ACC” denotes the accuracy of classifier and “AUC” is the score of Area Under the ROC Curve for the classifier.

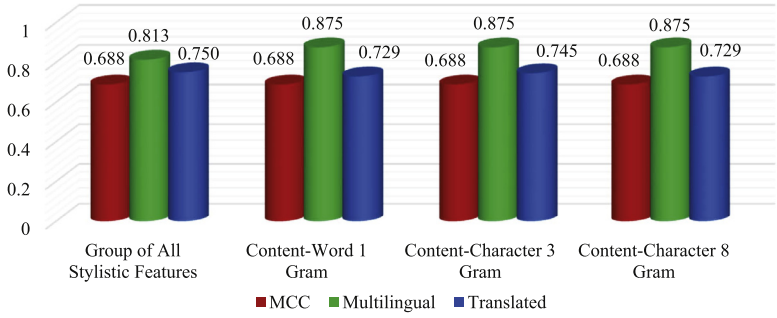

Fig. 1. Comparison of best results for gender among all stylistic features and content based methods.

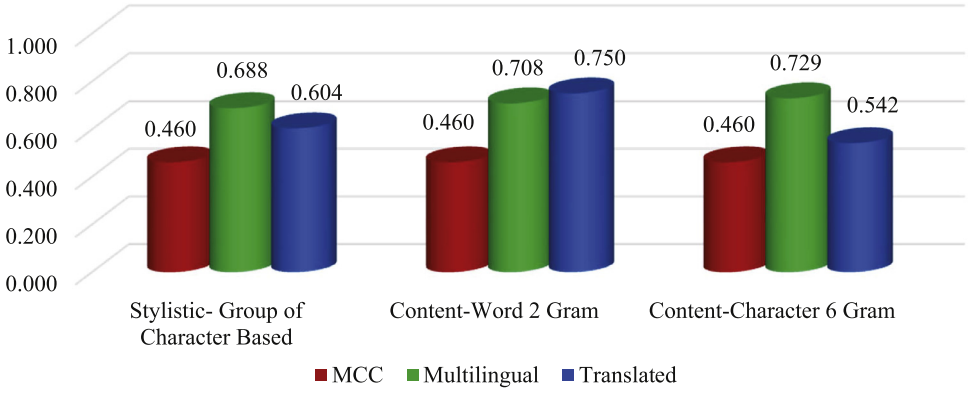

Fig. 2. Comparison of best results for age among all stylistic features and content based methods.

Inclusively the best results are achieved on multilingual dataset with "All Combined" feature for gender (*Accuracy* = 0.813). This accuracy is obtained by using Random Forest classifier with (*AUC* = 0.782). For age, the best results are obtained on multilingual dataset with group of character based features using Random Forest classifier (*Accuracy* = 0.708) and (*AUC* = 0.795). The obtained AUC score for Random Forest demonstrate that classifier performs prosperously in discriminating between different classes of age and gender.

Table 12 depicts the results of content based methods on multilingual and translated datasets for both gender and age. The "FM/CL" refers to the combination of feature selection method and machine learning algorithm used for classification.

The highest results are obtained for age on translated dataset with word 2–gram and we obtained (*Accuracy* = 0.750) and (*AUC* = 0.795) which are slightly higher than the results of multilingual dataset with character 6–gram (*Accuracy* = 0.729) and (*AUC* = 0.789). The values of AUC depict that classifier performance is thorough in distinguishing between different age groups is good.

For gender identification task, comprehensively the best results are obtained on multilingual dataset using word 1–gram, character 3–gram and character 8–gram with (*Accuracy* = 0.875) and (*AUC* = 0.847), (*AUC* = 0.799) and (*AUC* = 0.862) respectively.

To summarize, Fig. 1 presents the utmost results obtained for gender, among all the approaches applied on multilingual and translated corpora. The highest accuracy 0.875 is obtained on multilingual dataset with word 1–gram, character 3–gram and character 8–gram. These results are comparable to the best results obtained in different PAN Author Profiling Competitions.[17] For example, in PAN 2014 Author Profiling Competition the highest accuracy 0.729 and 0.539 were obtained with Spanish and English Social Media corpora respectively for gender (Rangel et al., 2014). In the mentioned competition, on Twitter corpus (for gender), the best accuracy of 0.884 and 0.766 were obtained for Spanish and English languages corpora respectively (Rangel et al., 2014). In PAN 2015 Author Profiling Competition, the highest accuracy 0.9688 on Italian, 0.9659 on Spanish, 0.8611 on Dutch and 0.8592 on English were obtained on Twitter corpus for gender identification (Rangel et al., 2015). These results clearly demonstrate that the performance of author profiling methods is different for different genres and languages.

Fig. 2 exhibits the comparison among the best results for age on multilingual and translated corpora. The highest accuracy 0.750 is obtained on translated corpus for word 2–gram. This indicates that machine translation is helpful in improving the performance for age group identification. Once more when we compare our results with the PAN ones, the

[17] Interestingly all the PAN Author Profiling corpora are monolingual.

Table 12
Comparison of results obtained using content based methods for multilingual and translated corpora.

Feature	Gender						Age					
	Multilingual			Translated			Multilingual			Translated		
	FM/CL MCC	ACC 0.680	AUC	FM/CL MCC	ACC 0.680	AUC	FM/CL MCC	ACC 0.460	AUC	FM/CL MCC	ACC 0.460	AUC
						Word N–Grams						
1gram	**IG/RF**	**0.875**	**0.847**	CS,GR, IG/SVM	0.729	0.639	IG/RF	0.688	0.762	CS,GR, IG/SVM,J48	0.583	0.689
2gram	CS,GR, IG/RF,SVM	0.813	0.736	CS,GR, IG/NB	0.750	0.729	CS,GR, IG/NB	0.708	0.747	**CS,GR, IG/NB**	**0.750**	**0.809**
3gram	CS,GR, IG/SVM	0.854	0.785	CS,GR/ SVM	0.813	0.700	GR/RF	0.625	0.699	CS,IG/ RF	0.625	0.717
						Character N–Gram						
2gram	IG/RF	0.833	0.729	CS,GR, IG/J48	0.708	0.540	IG/RF	0.620	0.778	CS,GR, IG/NB,J48,RF	0.521	0.559
3gram	**GR/RF**	**0.875**	**0.799**	GR/RF	0.745	0.779	CS,GR, IG/RF	0.646	0.816	CS,GR, IG/SVM	0.563	0.608
4gram	CS,GR, IG/RF, J48	0.833	0.810	CS/RF	0.708	0.627	GR/RF	0.708	0.817	CS,GR, IG/J48	0.563	0.666
5gram	CS,GR, IG/RF	0.854	0.806	GR/RF	0.750	0.635	CS,GR, IG/NB,RF	0.688	0.734	CS,GR, IG/NB,RF	0.625	0.622
6gram	CS,IG/ RF	0.854	0.802	CS,GR, IG/NB,SVM	0.729	0.682	**IG/RF**	**0.729**	**0.789**	CS,GR, IG/NB,SVM,RF	0.542	0.662
7gram	CS,GR, IG/RF,J48	0.854	0.818	CS,GR, IG/RF,SVM	0.723	0.708	CS/RF	0.708	0.845	CS,GR, IG/SVM	0.604	0.657
8gram	**GR/RF**	**0.875**	**0.862**	CS,IG/ J48	0.729	0.641	IG,GR/ RF	0.688	0.788	GR/SVM	0.542	0.678
9gram	CS,GR, IG/RF,SVM	0.854	0.803	CS,GR, IG/NB,SVM	0.729	0.733	IG/RF	0.708	0.824	CS,GR, IG/NB,J48	0.563	0.623
10gram	GR/RF	0.833	0.824	CS,IG/ J48	0.771	0.622	GR/RF	0.708	0.834	CS,GR, IG/SVM	0.479	0.512

results vary as the genre and language of the dataset changes. For instance, in PAN 2014 Author Profiling Competition (Rangel et al., 2014), the highest results obtained on Social Media corpora for age were: (*Accuracy* = 0.4262) for Spanish and (*Accuracy* = 0.3728) for English language. On Twitter dataset, the best accuracy 0.692 and 0.633 were obtained on Spanish and English corpora respectively. In PAN 2015 Author Profiling Competition (Rangel et al., 2015), for age the highest accuracy 0.838 and 0.795 were obtained on English and Spanish Twitter datasets.

To conclude, inclusively the character based features (both stylistic and content) outperform other features for both age and gender identification tasks. The possible reason is, communication on Facebook is usually very informal and character based features are likely to capture more discriminative information from the text.

7. Conclusion and future work

In this paper, we presented a benchmark multilingual (Roman Urdu and English) Facebook author profiles corpus (RUEN–AP–17) for the development and evaluation of author profiling methods. The proposed corpus contains 479 profiles (having both private and public comments/posts of a Facebook user) along with demographic information (age, gender, native language, native area, personality (type) and occupation). This study also contributes a manually generated bilingual dictionary of 7749 entries to translate Roman Urdu words into English (to investigate the impact of machine translation on the performance of author profiling methods).

To automatically identify an author's age and gender on our proposed multilingual translated corpora, for gender identification task, the best results were obtained on multilingual corpus (*Accuracy* = 0.875) with word unigram, character 3 and 8 gram content based approach and for age identification task, the highest accuracy 0.750 was obtained on translated corpus with word bigram content based approach.

The potential avenues of future research work are: increase the size of our proposed corpus by collecting more author profiles, applying more novel techniques on the proposed corpus, explore the behavior of language dependent features (e.g: Part Of Speech), increase the size of bilingual dictionary and identification of author traits other than age and gender.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2017.03.005

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems, 20*(5), 67–75.

Afzal, H., & Mehmood, K. (2016). Spam filtering of bi-lingual tweets using machine learning. In *2016 18th International conference on advanced communication technology (ICACT)* (pp. 710–714). PyeongChang, Korea: IEEE.

Alowibdi, J. S., Buy, U. A., & 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'13)* (pp. 739–743). Niagara Falls, Canada: ACM.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication, 20*(2), 204–220.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321–346.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123.

Barasa, S. N. (2010). *Language, mobile phones and internet: a study of SMS texting, email, IM and SNS chats in computer mediated communication (CMC) in Kenya*. Netherlands Graduate School of Linguistics Phd dissertation.

Bilal, M., Israr, H., Shahid, M., & Khan, A. (2015). Sentiment classification of Roman-Urdu opinions using Naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University-Computer and Information Sciences*.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 1301–1309). Edinburgh, United Kingdom: Association for Computational Linguistics.

Caplan, J. (2013). Social media and politics: Twitter use in the second congressional district of virginia. *Elon Journal of Undergraduate Research in Communications, 4*(1), 5–14.

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

Ciot, M., Sonderegger, M., & Ruths, D. (2013). Gender inference of Twitter users in non-english contexts. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)* (pp. 1136–1145). Seattle, USA: Association for Computational Linguistics.

Cvijikj, I. P., & Michahelles, F. (2011). Monitoring trends on Facebook. In *The 9th IEEE international conference on dependable, autonomic and secure computing (DASC 2011)* (pp. 895–902). Sydney, Australia: IEEE.

Daud, M., Khan, R., Daud, A., et al. (2014). Roman Urdu opinion mining system (Ruomis). *Computer Science & Engineering, 4*(5/6), 1.

De-Arteaga, M., Jimenez, S., Mancera, S., & Baquero, J. (2013). Author profiling using corpus statistics, lexicons and stylistic features—Notebook for PAN at CLEF-2013. *Clef 2013 evaluation labs and workshop – Working notes papers. Valencia, Spain*.

Duong, D. T., Pham, S. B., & Tan, H. (2016). Using content-based features for author profiling of Vietnamese forum posts. In *Recent developments in intelligent information and database systems* (pp. 287–296). Springer.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). TAT: An author profiling tool with application to arabic emails. In *Proceedings of the Australasian language technology workshop, Melbourne, Australia* (pp. 21–30).

Fairon, C., & Paumier, S. (2006). A translated corpus of 30,000 French SMS. In *Proceedings of the fifth international conference on language resources and evaluation (LREC 2006)* (pp. 351–354). Genoa, Italy: European Language Resources Association (ELRA).

Farahbakhsh, R., Han, X., Cuevas, A., & Crespi, N. (2013). Analysis of publicly disclosed information in Facebook profiles. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM'13)* (pp. 699–705). Niagara Falls, Canada: ACM.

Flekova, L., Ungar, L., & 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*.

Gencosman, B. C., Ozmutlu, H. C., & Ozmutlu, S. (2014). Character n-gram application for automatic new topic identification. *Information Processing & Management, 50*(6), 821–856.

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers age and gender. In *Proceedings of the third international AAAI conference of weblogs and social media (ICWSM'09)* (pp. 214–217). San Jose, California: AAAI Press.

How, Y., & Kan, M.-Y. (2005). Optimizing predictive text entry for short message service on mobile phones. In *MobileHCI '05: Proceedings of the seventh international conference on human-computer interaction with mobile devices and services: 5*. Salzburg, Austria: ACM.
Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers, 4*(8), 966–974.
Javed, I., & Afzal, H. (2013). Opinion analysis of bi-lingual event data from social networks.. In *Proceedings of the first international workshop on emotion and sentiment in social and expressive media: approaches and perspectives from ai (ESSEM 2013) a workshop of the XIII international conference of the italian association for artificial intelligence (AI*IA 2013)* (pp. 164–172). Torino, Italy: Citeseer.
Javed, I., Afzal, H., Majeed, A., & Khan, B. (2014). Towards creation of linguistic resources for bilingual sentiment analysis of Twitter data. In *19th international conference on application of natural language to information systems* (pp. 232–236). Montpellier, France: Springer.
Juola, P. (2015). Industrial uses for authorship analysis. In *Mathematics and computers in sciences and industry* (pp. 21–25). INASE.
Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management, 2*(2), 271–277.
Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the pacific association for computational linguistics (PACLING'03): 3* (pp. 255–264). Halifax, Nova Scotia, Canada: Computer Science Deptartment at Dalhousie University.
Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401–412.
Lin, J. (2007). *Automatic author profiling of online chat logs*. Naval Postgraduate School, Monterey, California mathesis.
Litvinova, T. A. (2014). Profiling the author of a written text in russian. *Journal of Language and Literature, 5*(4), 210–216.
Mikros, G. K. (2012). Authorship attribution and gender identification in Greek blogs. In *Methods and applications of quantitative linguistics* (pp. 21–32). University of Belgrade, Serbia.
Mikros, G. K., & Perifanos, K. (2013). Authorship attribution in Greek tweets using author's multilevel N-gram profiles. *AAAI 2013 spring symposium: Analyzing microtext (sam2013)*. Stanford, USA: AAAI Press.
Mukund, S., & 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* (pp. 1–8). Association for Computational Linguistics.
Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media (ICWSM'13)* (pp. 439–448). Cambridge, Massachusetts, USA: AAAI Press.
Nguyen, D., Smith, N. A., & Rosé, C. P. (2011). Author age prediction from text using linear regression. In *LaTeCH '11: Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 115–123). Portland, Oregon: Association for Computational Linguistics.
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934–952.
Peersman, C., Daelemans, W., & 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'11)* (pp. 37–44). Glasgow, Scotland, UK: ACM.
Pervaz, I., Ameer, I., Sittar, A., & Nawab, R. M. A. (2015). Identification of author personality traits using stylistic features—Notebook for PAN at CLEF 2015. *Clef 2015 evaluation labs and workshop – Working notes papers*. Toulouse, France: CEUR-WS.org.
Pham, D. D., Tran, G. B., & Pham, S. B. (2009). Author profiling for Vietnamese blogs. In *2009 International conference on asian language processing (IALP)* (pp. 190–194). Singapore: IEEE Computer Society.
Poulston, A., Stevenson, M., & Bontcheva, K. (2015). Topic models and n-gram language models for author profiling—Notebook for PAN at CLEF 2015. *Clef 2015 evaluation labs and workshop – Working notes papers*. Toulouse, France: CEUR-WS.org.
Przybyła, P., & 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. *Clef 2015 evaluation labs and workshop – Working notes papers*. Toulouse, France: CEUR-WS.org.
Rangel, F. (2013). Author profile in social media: Identifying information about gender , age , emotions and beyond. In *Fifth BCS-IRSG symposium on future directions in information access (FDIA 2013), Granada, Spain* (pp. 58–60).
Rangel, F., & Rosso, P. (2013). On the identification of emotions and authors' gender in Facebook comments on the basis of their writing style. In *Proceedings of the first international workshop on emotion and sentiment in social and expressive media: Approaches and perspectives from AI (ESSEM 2013): 1096* (pp. 34–46). Turin, Italy: CEUR Workshop Proceedings.
Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing & Management, 52*(1), 73–92.
Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at PAN 2013. *Clef 2013 evaluation labs and workshop – Working notes papers. Valencia, Spain*.
Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. *Clef 2015 evaluation labs and workshop – Working notes papers*. Toulouse, France: CEUR-WS.org.
Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., et al. (2014). Overview of the 2nd author profiling task at PAN 2014. *Clef 2014 evaluation labs and workshop – Working notes papers*. Sheffield, UK: CEUR-WS.org.
Reddy, T. R., Vardhan, B. V., & Reddy, P. V. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research, 11*(5), 3092–3102.
Rosenthal, S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 763–772). Portland, Oregon: Association for Computational Linguistics.
Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: Predicting age and gender from blogs—Notebook for PAN at CLEF 2013. *Clef 2013 evaluation labs and workshop – Working notes papers. Valencia, Spain*.
Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., et al. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1146–1151). Doha, Qatar: Association for Computational Linguistics.
Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI spring symposium on computational approaches to analyzing weblogs: 6* (pp. 199–205). Palo Alto, California: AAAI Press.
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE, 8*(9), e73791.
Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on youtube. *Information Processing & Management, 52*(1), 46–60.
Shrestha, P., Rey-Villamizar, N., Sadeque, F., Pedersen, T., Bethard, S., & Solorio, T. (2016). Age and gender prediction on health forum data. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Portoro, Slovenia: European Language Resources Association (ELRA).
Soler, J., & Wanner, L. (2016). A semi-supervised approach for gender identification. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Portoro, Slovenia: European Language Resources Association (ELRA).
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information, 60*(3), 538–556.
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471–495.
Türkoğlu, F., Diri, B., & Amasyalı, M. F. (2007). Author attribution of Turkish texts by feature mining. In *International conference on intelligent computing* (pp. 1086–1093). Qingdao, China: Springer.
Van de Loo, J., De Pauw, G., & Daelemans, W. (2016). Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics (IJCSDF), 5*(1), 46–60.
Verhoeven, B., Daelemans, W., & Plank, B. (2016). TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Portoro, Slovenia: European Language Resources Association (ELRA).

Villegas, M. P., Garciarena Ucelay, M. J., Errecalde, M. L., & Cagnina, L. (2014). A Spanish text corpus for the author profiling task. *CACIC 2014 XX congreso argentino de ciencias de la computación. San Justo, Buenos Aires, Argentina*.

Volkova, S., & Yarowsky, D. (2014). Improving gender prediction of social media users via weighted annotator rationales. *NIPS 2014 workshop on personalization: Methods and applications. Montreal, Canada*.

Wanner, L. (2015). Multiple language gender identification for blog posts. In *Proceedings of the 37th annual meeting of the cognitive science society, Pasadena, California* (pp. 2248–2251).

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

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management, 48*(4), 741–754.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the fourteenth international conference on machine learning (ICML 1997), Nashville, TN, USA* (pp. 412–420). Morgan Kaufmann Publishers Inc..

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on human language technology (HLT'01) research* (pp. 1–8). San Diego, California: Association for Computational Linguistics.

Zhang, W., Caines, A., Alikaniotis, D., & Buttery, P. (2016). Predicting author age from weibo microblog posts. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Portoro, Slovenia: European Language Resources Association (ELRA).

Zielinski, A., Bügel, U., Middleton, L., Middleton, S., Tokarchuk, L., Watson, K., et al. (2012). Multilingual analysis of twitter news in support of mass emergency events. In *Proceedings of the 9th international conference on information systems for crisis response and management (ISCRAM 2012). Vancouver, Canada*.

Zimmer, M. (2010). But the data is already public: On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313–325.