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.