Multimodal Author Profiling for Twitter

Notebook for PAN at CLEF 2018

Braja Gopal Patra[1], Kumar Gourav Das[2], and Dipankar Das[3]

[1]Department of Biostatistics and Data Science, School of Public Health,
University of Texas Health Science Center (UTHealth), Houston, TX, USA
[2]Department of Computer Science & Engineering,
Future Institute of Engineering & Management, Kolkata, India
[3]Department of Computer Science & Engineering, Jadavpur University, Kolkata, India
{brajagopal.cse,kumargouravdas18,dipankar.dipnil2005}@gmail.com

Abstract Author profiling is gaining the interest of people in both academia and outside it. Author profiling/analysis deals with the identification of author information from text based on stylistic choices. It helps in identifying author related information such as gender, age, native language, personality, demographics, etc. Thus, author profiling is both challenging and important. This paper describes the systems submitted to author profiling task at PAN-2018 using multimodal (textual and image) Twitter datasets provided by the organizers and the aim is to identify the author's gender. An image captioning system was used to extract captions from images. Mainly latent semantic analysis, word embeddings, and stylistic features were extracted from tweets as well as captions. The proposed multimodal author profiling systems obtained classification accuracies of 0.7680, 0.7737, and 0.7709 for Arabic, English and Spanish languages, respectively using support vector machine.

Keywords: Author profiling, gender detection, latent semantic analysis, latent dirichlet allocation, word embeddings.

1 Introduction

Social media has become an integral part of human life. People often spend a lot of time on social media. Further, they also input text data which is prone to noise elements like typos, and grammatical mistakes. Thus it is both challenging and necessary to uncover various characteristics of the author from such noisy social media text. Author profiling (AP) is essential in several areas including marketing, forensic science, and security. For example, from a marketing perspective, it is always useful to know details about authors of text in blogs and reviews, so that relevant recommendations can be provided to users. The linguistic profile of an author of abusive message would be helpful from a forensic linguistics viewpoint.

AP gained importance as a research area since the last decade [17]. Initially, AP was only based on text data generated by authors [8,13,17,21,25]. In PAN-2018, a new research trend in AP was started, as both text and image data were made available to

be used for AP. PAN is a series of scientific events and shared tasks on digital text forensics[1].

In this paper, we perform gender identification from multimodal Twitter data, provided by the organizers of AP task[2] at PAN-2018. The major focus is on social media text as we are interested in how everyday language reflects on social and personal choices. The organizers provided tweets and photos of users using either of three languages namely, Arabic, English, and Spanish. The training dataset consists of data obtained from 3000 users each of English and Spanish languages, while there are only 1500 users of Arabic language.

For English dataset, we identified several important textual features including words embeddings and stylistic features. An image captioning system was used to extract captions from images, and then the above textual features were identified from the captions. In contrast, a language-independent approach was used for Arabic and Spanish datasets. We collected term frequency-inverse document frequency (TF-IDF) of unigrams, then singular value decomposition (SVD) was implemented on TF-IDF vectors to reduce sparsity. Finally, latent semantic analysis (LSA) was used on the reduced vectors to get the final feature vectors. Support vector machine (SVM) was implemented for classification purpose.

Rest of the paper is organized in the following manner. Section 2 discusses related work briefly. Section 3 provides an overview of data, features, system architecture, and techniques used in the experiments. Section 4 describes a detailed analysis of results. Finally, conclusions and future directions are listed in Section 5.

2 Related Work

AP focuses on the prediction of demographics and psychometric traits (age, gender, native language, personality, religion) of an author using stylistic and content-based features. AP has many applications in academic research, marketing, security and forensic analysis. Initially, research on AP was conducted on English language [17,25] and later gained popularity in other languages like Dutch [13], Greek [8], Italian [21], Spanish [13,25], Vietnamese [19], and so on.

There has been much research on AP from blogs as well as social media texts. Bayot and Gonçalves [6] performed age and gender classification on PAN-2016 AP datasets using TF-IDF scores and word embeddings. Classification was performed using support vector machine (SVM) and results showed that TF-IDF worked better than `word2vec` for age classification while `word2vec` performed better for gender classification. Akhtyamova et al. [1] used word embeddings with logistic regression for AP task at PAN-2017. On the other hand, Arroju et al. [4], Bartoli et al. [5], and Marguardt et al. [14] used LIWC [27] for AP at PAN-2017.

Schler et al. [11] tried to identify age and gender from the writing style in blogs. The authors used non-dictionary words, parts-of-speech (POS), function words, hyperlinks, combined with content features like unigram with the highest information gain for AP

[1] https://pan.webis.de/index.html

[2] https://pan.webis.de/clef18/pan18-web/author-identification.html

task. Argamon et al. [3] documented how the variation of linguistic characteristics was responsible for identifying authors age and gender. The authors mainly focused on the functional words with POS features for gender prediction. Holmes et al. [10] and Burger et al. [7] performed similar studies by focusing on the extraction of age and gender information from formal text.

Exhaustive studies performed by Rangel and Rosso [22] shows that age and gender depend on the use of language. They used stylistic features like frequency, punctuation marks, POS, emoticons, and obtained the best result by SVM classifier on PAN-2013 AP dataset. Another notable work mentioned by the same authors which took emotions into account for AP task on tweets [23]. They have used EmoGraph, Graph based approaches for identifying gender and age on PAN-2013 AP dataset. In another work, Weren et al. [28] used information retrieval based features such as information gain and cosine similarity with each category for age and gender identification on PAN-2013 AP dataset.

The above survey reveals that a variety of features can be used for AP. Many experiments in AP were performed using content-based features like slang words, happy-emotion words, sad-emotion words, sentiment words [23]. In contrast, stylistic features, such as frequency, punctuation, POS, and other different statistics were also used for AP in [9]. More recently, word embeddings like `word2vec` and document embeddings like `Doc2Vec` were used features for AP in addition to bag-of-words and TF-IDF [12,15].

AP task at PAN started in 2013 and it focused on age and gender classification for two languages (English, Spanish). AP task at PAN-2014 targeted on the same language sets with four different genres of corpora: social media, blogs, Twitter, and hotel reviews, though hotel reviews dataset was only available for English. This task focuses on age and gender classification of authors. AP task at PAN-2015 extended to four different languages (Dutch, English, Italian, and Spanish) and datasets were collected only from Twitter. AP task at PAN-2016 focused on gender and age classification, and the corpora contain tweets, reviews, blogs and other social media for three different languages (Dutch, English, Spanish). AP task at PAN-2017 focused on Twitter datasets in four different languages (Arabic, English, Spanish, Portuguese) with gender and language variety identification. This time, the AP task at PAN-2018 is performed on three different languages (English, Spanish, Arabic), and the datasets contain tweets and images from Twitter.

3 Methodology

3.1 Dataset

The AP task at PAN-2018 focused on users' gender detection using their tweets and photos shared on Twitter. The organizers provided a training dataset each for the three languages (Arabic, English, and Spanish). Both English and Spanish datasets contain 3000 users' information each (1500: Male, 1500: Female) while Arabic dataset contains 1500 users' information (750: Male and 750: Female). For each user, there are 100 tweets along with 10 images. The organizers also provided the test dataset and the details can be found in the overview paper of AP task [24].

3.2 Features

This section describes several features used in our experiments. Feature selection plays an important role in any machine learning framework and depends upon the dataset used for experiments. The features are as follows:

Stylistic Features: This is an important feature and has been extensively used in AP tasks [16,17,22]. The number of stop words, punctuations, happy and sad smilies, tweets or retweets, hashtags, and slangs were considered in the present study. The stop word lists for all three languages were collected from `nltk` corpus[3]. The slang word list was manually prepared only for English.

Word Embeddings based Features: Recently, word embeddings gained popularity in text mining and information retrieval, and it has been used in several tasks including AP [12,15]. For the present study, word vector representations were obtained using the `word2vec` model, `GloVe` [18] (global vector for word representation). There are many advantages of `GloVe` over the traditional `word2vec` model. First, it is trained on 2 billion tweets, and second, it provides a flexible dimension of feature space. `GloVe` delivers a single feature vector for each of the words in a tweet and those word vectors were converted to tweet vectors ($\overrightarrow{t_i}$) using equation 1. Finally, tweet vectors ($\overrightarrow{t_i}$) were added together to create a single user vector as in equation 2.

$$\overrightarrow{t_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \overrightarrow{w_{ij}} \tag{1}$$

$$\overrightarrow{U_k} = \sum_{i=1}^{100} \overrightarrow{t_i} \tag{2}$$

where $\overrightarrow{t_i}$ is tweet vector for i^{th} tweet, w_{ij} is the j^{th} word in i^{th} tweet; N_i is the number of words present in `GloVe` for i^{th} tweet, and $\overrightarrow{U_k}$ is the k^{th} user vector.

The word vectors of dimensions 100 and 200 were used for image captions and tweets. We used word embeddings only for English dataset due to the availability of pre-trained models on tweets.

Latent Semantic Analysis: LSA is a technique for creating a vector representation of a document, similar to document embeddings or `Doc2Vec`. It has been successfully used for several applications in Natural Language Processing (NLP) including AP [2]. The steps for implementing LSA on a set of tweets belonging to a user is as follows. Initially, we calculated TF-IDF vector for tweets of each user and then implemented singular value decomposition (SVD) on TF-IDF vectors to reduce dimensionality. Finally, we implemented LSA to get final vectors for each user on 100 tweets. This feature was used to convert tweets to feature vectors for Arabic and Spanish languages, and to convert hashtags to feature vectors for all three languages.

Topic Words: It is useful to collect topic words which describe the whole document in few words. We used Latent Dirichlet Allocation (LDA) to collect all the important words for a single user and LDA implemented using `gensim`[4] was used in the experiments. For a single user, we collected three topics containing 10 words each from 100

[3] https://www.nltk.org/book/ch02.html

[4] https://radimrehurek.com/gensim/

tweets. Topic words were converted to feature vectors either using word embeddings (GloVe) or LSA. This feature was used for all three training datasets.

Hashtags: Hashtags are informative on microblogs such as Twitter. Total of 1777, 48292, and 35018 number of unique hashtags were present in training datasets of Arabic, English, and Spanish languages, respectively. Thus, the extensive use of hashtags in training datasets (except Arabic) motivated us to use it as a feature. We used LSA to get a feature vector from all hashtags used by a single user. We used this feature for all three languages.

Image Captions: Several state-of-the-art image captioning systems using deep learning are available nowadays. We used an existing image captioning system by Tsutsui and Crandall [26] to extract information present in images. This image caption generation system provides a detailed image captioning for all images. It also provides captions in Chinese, English, and Japanese. For the present task, we only considered captions in English language. LDA was used to identify topic words from image captions. The image captions and topic words were converted to feature vector using either LSA or word embeddings (GloVe).

3.3 System Architecture

Figure 1 shows the detailed architecture of text-based AP system for English language. AP system for English used mainly word embeddings, GloVe. It was used to convert topic words and tweet tokens into separate feature vectors. We used different dimensions for different modules of word embeddings. We performed 10-fold cross-validation on training dataset with different vector sizes of GloVe and the maximum accuracy was obtained for feature dimensions of 200 and 100 for tweets and topic words, respectively. Thus, we used similar settings for all experiments. Further, hashtags were also converted to feature vectors by sequentially using TF-IDF, SVD, and LSA as described in section 3.2. We generated 25-dimensional feature vector from all hashtags of a single user.

Figure 2 describes the detailed architecture of text-based AP system for both Arabic and Spanish. For Arabic and Spanish, no pre-trained word embeddings were available for tweet dataset. Thus, using word embeddings was not an option for both of the languages. We calculated TF-IDF for unigrams and reduced vector size using SVD and then used LSA to get the final feature vectors as described in section 3.2. We identified

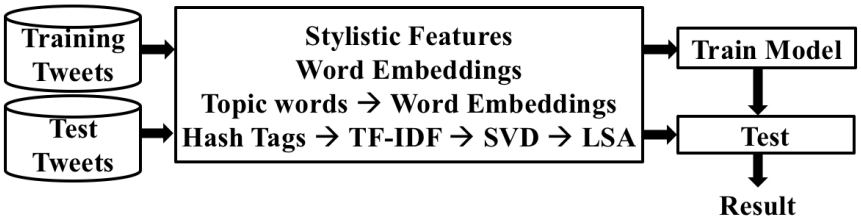

Figure 1. Architecture of text-based AP system for English dataset

topic words from tweets of a single user, then implemented a similar method to get feature vectors from topic words. The feature vector from hashtags was extracted using the same method which was used for English language. We generated 200-dimensional vector from tweets and 100-dimensional feature vector from topic words. We also generated 25-dimensional feature vector for hashtags.

The image captioning system generates captions in English. Figure 3 describes the architecture of image-based AP system for English while Figure 4 describes the architecture of image-based AP systems for Arabic and Spanish. For English AP system, we collected captions for all images of a single user. We converted all the words (except stop words) into word vectors using `GloVe`. We identified topic words using LDA and then converted each topic words into word vectors using `GloVe`. Each word vectors are summed together using equation 1 to get a single vector for a single user. The im-

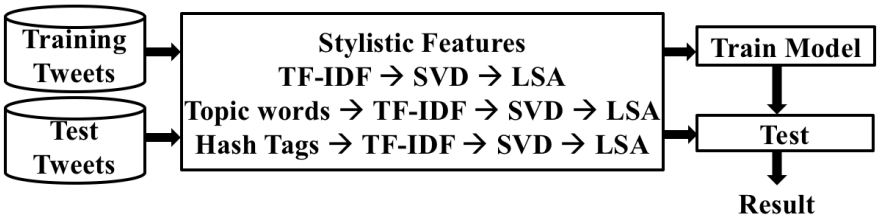

Figure 2. Architecture of text-based AP systems for Arabic and Spanish datasets

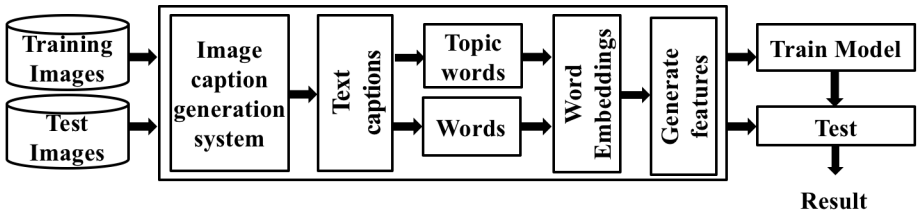

Figure 3. Architecture of image-based AP system for English dataset

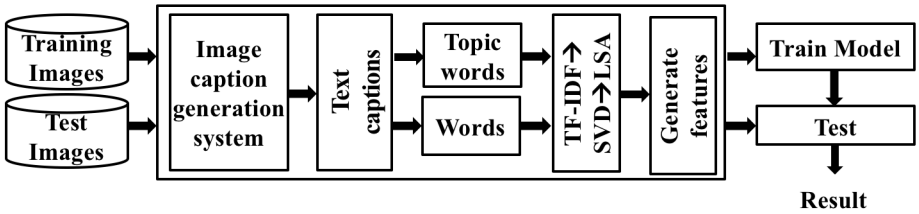

Figure 4. Architecture of image-based AP systems for Arabic and Spanish datasets

age caption vector and image caption topic vector resulted in a 200-dimensional feature vector for each user.

For Arabic and Spanish AP systems, we collected captions for all images of a single user and then, collected all the words (except stop words). We also identified topic words from the above captions. Finally converted all words and topic words into two 100-dimensional feature vector using LSA as described in Section 3.2. The image captions are in English language; we could have implemented `GloVe` for all three languages. We wanted AP systems for Arabic and Spanish to be language-independent; thus, a different method was implemented for extracting features from captions to that of English AP system. For image captions, a 200-dimensional feature vector was generated from both captions and topic words. We also developed three multimodal AP systems for three datasets. For the multimodal systems, text and image features were combined together.

4 Results and Discussion

Initially, several classifiers such as Decision Tree, Random Forest, SVM implemented in `scikit-learn`[5] were used for 10-fold cross-validation on the training datasets. It was observed that SVM classifier outperformed all other classifiers. Thus, we used SVM for developing all AP systems using text, image, and combination of both. We used the linear kernel for all the experiments. All AP systems are evaluated based on accuracies.

We submitted trained models and feature extraction codes in the virtual machine, TIRA[6] [20]. TIRA provides a means for evaluation as a service[7]. The system performances were calculated on the test dataset in TIRA and the organizers provided the accuracies of the systems.

4.1 Results

Initially, text and image features were separately used for classification. Later, multimodal systems were developed using a combination of text and image features. The accuracies of text, image and multimodal AP systems for all three languages are presented in Table 1. The maximum accuracy of 0.7586 was obtained for the textual based AP system for Spanish language among all three languages. The maximum accuracy of 0.6918 was obtained for image-based AP system for Spanish language among all three languages. Though the multimodal AP system did not perform well for Spanish language and the main reason may be the curse of dimensionality. The maximum accuracy of 0.7737 was obtained for multimodal AP system for English of all three languages.

[5] http://scikit-learn.org/stable/

[6] http://www.tira.io/

[7] https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/research/activities-by-field/tira/#c41469

Table 1. Accuracies of AP systems developed on English, Spanish and Arabic languages using different feature categories

Languages	Text Features	Image Features	Combined Features
Arabic	.7430	.6570	.7680
English	.7558	.6747	.7737
Spanish	.7586	.6918	.7709

4.2 Discussion

The multimodal systems for all three languages outperformed unimodal systems developed using either text or image. This shows the superiority of multimodal dataset over the traditional unimodal dataset. The Arabic language dataset contains 1500 users' data as compared to 3000 users' data for other languages and this may be the main reason for low accuracy of AP systems for Arabic language among all languages. There were no words found in `GloVe` for many tweets from English dataset and that resulted in a zero vector for those tweets. These tweets contain mostly emoticons or hashtags or miss spelled words. This may be one of the reasons for low accuracy of text-based AP system for English dataset.

Our system ranked 12^{th} among 23 participants in AP task at PAN with the average accuracy of 0.7709 for all three multimodal AP systems. The highest accuracy of 0.8198 was obtained by *takahashi18* team across all multimodal AP systems. Our AP systems for Arabic and Spanish achieved 10^{th} rank and for English, it achieved 16^{th} rank. The multimodal systems obtained the maximum accuracies of 0.8180, 0.8584, and 0.8200 for Arabic, English and Spanish languages by *miranda18*, *takahashi18* and *daneshvar18* teams, respectively.

5 Conclusion

We presented AP systems to identify the gender of users from Arabic, English, and Spanish multimodal datasets. Among three languages, the multimodal AP system for English outperformed other two languages.

LSA worked well in the case of AP systems for Arabic and Spanish languages; it will be interesting to implement LSA on English dataset. In the future, we will perform several experiments with different word and document embeddings on all datasets. Several other language-independent approaches such as n-grams can be implemented later. We are also planning to implement different deep learning models for gender detection.

References

1. Akhtyamova, L., Cardiff, J., Ignatov, A.: Twitter author profiling using word embeddings and logistic regression - notebook for PAN at CLEF 2017. In: Working Notes for CLEF 2017 Conference, Dublin, Ireland (2017)

2. Alvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villasenor-Pineda, L., Escalante, H.J.: INAOE's participation at PAN'15: Author profiling task - notebook for PAN at CLEF 2015 (2015)
3. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text - Interdisciplinary Journal for the Study of Discourse 23(3), 321–346 (2006)
4. Arroju, M., Hassan, A., Farnadi, G.: Age, gender and personality recognition using tweets in a multilingual setting - notebook for PAN at CLEF 2015. In: Working Notes for CLEF 2015 Conference, Toulouse, France (2015)
5. Bartoli, A., De Lorenzo, A., Laderchi, A., Medvet, E., Tarlao, F.: An author profiling approach based on language-dependent content and stylometric features - notebook for PAN at CLEF 2015. In: Working Notes for CLEF 2015 Conference, Toulouse, France (2015)
6. Bayot, R., Gonçalves, T.: Multilingual author profiling using word embedding averages and svms. In: 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA). pp. 382–386. IEEE (2016)
7. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Conference on empirical methods in natural language processing (EMNLP). pp. 1301–1309. Association for Computational Linguistics (2011)
8. Dang Duc, P., Giang Binh, T., Son Bao, P.: Authorship attribution and gender identification in greek blogs. In: 8th International Conference on Quantitative Linguistics (QUALICO) (2012)
9. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for english emails. In: 10th Conference of the Pacific Association for Computational Linguistics. pp. 263–272 (2007)
10. Holmes, J., Meyerhoff, M.: The handbook of language and gender, vol. 25. John Wiley & Sons (2008)
11. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the Association for Information Science and Technology 60(1), 9–26 (2009)
12. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196 (2014)
13. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.F.: Adapting cross-genre author profiling to language and corpus - notebook for PAN at CLEF 2016. In: Working Notes for CLEF 2016 Conference, Évora, Portugal. pp. 947–955 (2016)
14. Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.F., Davalos, S., Teredesai, A., De Cock, M.: Age and gender identification in social media. In: CLEF 2014 Evaluation Labs. pp. 1129–1136 (2014)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
16. Patra, B.G., Banerjee, S., Das, D., Bandyopadhyay, S.: Feeling may separate two authors: Incorporating sentiment in authorship identification task. In: International Conference on Natural Language Processing. pp. 121–126 (2013)
17. Patra, B.G., Banerjee, S., Das, D., Saikh, T., Bandyopadhyay, S.: Automatic author profiling based on linguistic and stylistic features - notebook for PAN at CLEF 2013. In: Working Notes for CLEF 2013 Conference, Valencia, Spain (2013)
18. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
19. Pham, D.D., Tran, G.B., Pham, S.B.: Author profiling for vietnamese blogs. In: International Conference on Asian Language Processing. pp. 190–194. IEEE (2009)

20. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of pans shared tasks. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 268–299. Springer (2014)
21. Poulston, A., Stevenson, M., Bontcheva, K.: Topic models and n–gram language models for author profiling - notebook for PAN at CLEF 2015. In: Working Notes for CLEF 2015 Conference, Toulouse, France (2015)
22. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. In: 10th International Workshop on Natural Language Processing and Cognitive Science. pp. 177–186 (2013)
23. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information processing & management 52(1), 73–92 (2016)
24. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in twitter. In: CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2018)
25. Sapkota, U., Solorio, T., Montes-y Gómez, M., Ramírez-de-la Rosa, G.: Author profiling for english and spanish text - notebook for PAN at CLEF 2013. In: Working Notes for CLEF 2013 Conference, Valencia, Spain (2013)
26. Satoshi Tsutsui, D.C.: Using Artificial Tokens to Control Languages for Multilingual Image Caption Generation. In: CVPR Language and Vision Workshop (2017)
27. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)
28. Weren, E.R., Kauer, A.U., Mizusaki, L., Moreira, V.P., de Oliveira, J.P.M., Wives, L.K.: Examining multiple features for author profiling. Journal of information and data management 5(3), 266–279 (2014)