Arabic Author Profiling and Deception Detection using Traditional Learning Methodologies with Word Embedding

Haritha Ananthakrishnan, Akshaya Ranganathan, Thenmozhi D, and Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai
{haritha16038,akshaya16009}@cse.ssn.edu.in {theni_d, aravindanc}@ssn.edu.in

Abstract. With the ubiquity of social media, although one's thoughts and opinions can be expressed through virtual platforms effortlessly, there have been numerous cases of posts that threaten the security of a certain community, caste, or religion or spread false propaganda against a certain group of people. Developments in the fields of Natural Language Processing and Machine Learning have paved the way to the concept of author profiling, which helps identify an author's age, demographics, and gender details. The Author Profiling in Arabic Tweets task of FIRE 2019 aims to monitor Arabic Twitter posts and profile their authors concerning their age, gender, and language variety using learning concepts. The task of Deception Detection in Arabic texts focuses on monitoring Twitter and News headlines and detect deceptive texts: Posts that are drafted to seem authentic but suggest other ulterior motives. We have adopted the concept of SGD Optimized Support Vector Machine classification with AraVec word embedding for both the tasks and have achieved a joint F-1 score of **0.3403 for Author Profiling** and an average score of **0.7598 for the Deception detection** task.

Keywords: Author profiling · Deception Detection · Natural Language Processing · Machine Learning · Arabic Tweets · News · Support Vector Machines · Stochaistic gradient descent

[1]

1 Introduction and Related works

The APDA task of FIRE 2019 funded by ARAP Qatar aims to enhance cybersecurity using Machine Learning. Social media such as Twitter, Facebook, Instagram, etc. have gained popularity over the past decade where users can post

images, and text online, no questions asked. Many news magazines and papers have shifted their base to virtual media as well. Although these practices have uncovered many societal goods, they have engendered many drawbacks such as the rapid propagation of offensive posts and unreliable news that threaten the security of certain communities or individuals. Author profiling is the process of predicting the chatacterestics an author (gender, age, and location), based on their *stylistic writing features*. This not only forms a security layer, but also provides interesting linguistic information for targeted advertising, and marketing. The second task, Deception detection aimed at classifying news headlines and snippets as deceptive or non-deceptive, based on their language and word usage. This was done in [8] through the analysis of linguistic cues or leakages such as frequencies and patterns of word usage. The Text Attribution Tool (TAT) was developed for author profiling in a variety of languages [2]. Arabic poses a challenge as extensive data pre-processing is required like tokenization, character set normalization, informal spelling normalization, etc. Buckwalter scheme has been used for character set normalization. A variety of algorithms have been tried in the past and Bagging and SVM based SMO algorithms proved to have considerably better performance. Generally, traditional machine learning algorithms like SVM had better results [7]. For deception detection, Credibility Analysis of Arabic Content on Twitter (CAT)[8] which used user's timeline features like the number of retweets, user's activity, etc.

2 Dataset Analysis and Data cleaning

2.1 Dataset Analysis

The APDA task of FIRE 2019 was divided into two individual tasks, Author Profiling of Twitter posts and Deception Detection of news headlines and tweets. The corpora of the Author Profiling task consisted of a total of **2,250** users released over five days as groups of 450 users per day, which had even distribution across all classes. Each user's posts were collectively classified into three of the following classes: Gender, Age and language variety.

- **Gender** - Male or Female
- **Age** - Under 25, Between 25 and 34, Above 34
- **Language variety** - Algeria, Egypt, Iraq, Kuwait, Lebanon-Syria, Lybia, Morocco, Oman, Palestine-Jordan, Qatar, Saudi Arabia, Sudan, Tunisia, UAE, Yemen.

The training dataset of the Deception detection in Arabic texts task was divided into Twitter headlines and News headlines, each of which had two classifications based on whether or not the sentence was deceptive – **Truth or Lie.**

2.2 Data Cleaning

Data cleaning was a major task for the Arabic data set, as the language was written from right to left, and had different rules of sentence termination, all

Table 1. Distribution of corpora for both tasks

Type	Training users	Test users
Twitter	532	241
News	1443	370
Author Profiling	2250	720

of which were indecipherable to those who do not know Arabic. The following processes were implemented in python to flatten out data discrepancies.

- **Links** : Twitter posts contain lots of hyperlinks containing "Http://", "www.", ".com", etc. All of these were adding noise to the textual data and hence were removed using regular expressions.
- **Hashtags**: English hashtags were removed in their entirety, whereas Arabic hashtags were maintained with only the removal of the hashtag to prevent loss of useful linguistic information.
- **Non - Arabic Words**: As it was impractical to run language models for both Arabic and English due to the sparse density of English posts, all English words were removed from the text.
- **Twitter handles**: Social Media handles starting with "@" were eliminated from the text
- **Special Characters**: To further level the contents of the text, all special characters, erroneous blank spaces, numbers, and empty strings were removed
- **Emojis:** A major challenge for data cleanup for both tasks, was the removal of emojis, which had characters outside of the basic multilingual plane. These characters were extracted using Unicode conversion and used regular expressions to remove them.
- **Stop Words**: To remove Arabic stopwords, the **NLTK platform's Arabic stopword list** was taken, against which every sentence was filtered.

3 Methodology and Implementation

3.1 Word Embedding

Word embedding [4] is one of the most recent developments in the field of Natural Language Processing that facilitates the identification of the context of a word, where words are represented as vectors in a continuous space, capturing syntactic and semantic relationships between them. AraVec [6] is a powerful, pre-trained word embedding tool developed solely for Arabic NLP research, which is built upon Twitter, World Wide Web, and Wikipedia Arabic pages. The pre-processing step of AraVec included the removal of *tashkeel*, an Arabic symbol added after a word to distinguish it from others, which did not contribute to the overall meaning of the sentence. AraVec is built over the gensim[2] Word2Vec

[2] https://radimrehurek.com/gensim/about.html

model[3] which is a two-layer neural network used to identify the right context of words using CBOW and Skip-Gram techniques. AraVec was used to pre-process our text, as the generated word vectors could assign appropriate weights to the semantics of words.

3.2 Machine Learning model

We have implemented an SGD optimized SVM classifier with Aravec word embedding for our model. In Stochastic Gradient Descent optimization[1], the gradient of the loss is estimated one sample at a time, which is randomly shuffled for performing the iteration. In our model, SGD Classifier of sklearn performs Stochastic Gradient Descent Optimization on a **linear SVM Classification Model** whose training accuracy was 86% for Author profiling and an average of 92% for Deception detection of Twitter And News.

4 Results Analysis

The performance of our model for deception detection was remarkably better than that for author profiling. The F1 scores were **0.34 for author profiling and 0.76 for deception detection** [9]. SGD optimized SVM worked better for deception detection. This can be attributed to the comparatively low sizes of the news datasets, leading to better F1 scores. For author profiling, the model was able to predict well when predicting the gender and location. F1 scores were **0.76 and 0.83** respectively. The model performed poorly when it came to age with the F1 score being **0.55**. Therefore, the joint performance was lower in comparison to the best performer who scored **0.4556**. The results can be explained by pondering into the structural differences in Arabic. Research on Language and Gender Differences in Jordanian Spoken Arabic [11] shows that there are significant differences in the Arabic speech of men and women. Similarly, [12] aims at exploring the regional variations in Arabic. However, such differences in written Arabic amongst different age groups are unfathomable. Through word embedding, vectors are generated by grouping words that belong to similar contexts. As the differences of Arabic amongst age groups were not remarkable, the vectoriser could have given similar scores which ultimately led to the poor performance. The same logic can be used to explain the comparable performances in deception detection and prediction of gender and location.

5 Conclusion and Future works

Author profiling and deception detection have numerous applications given the amount of data exchanges over the internet each day. Although numerous NLP models are available for the English language, these need to be extended to languages such as Arabic, which is spoken by a vast majority of the world, which is what APDA@FIRE aims to achieve. This paper aims to predict the personality

of authors based on their style of word usage and to ensure the credibility of news by classifying the snippets as true or false. We have used word embedding and a traditional learning model to implement the same. Future scope includes comparing the accuracy of different types of traditional models such as Decision trees, Random Forrest, and Bayesian classifiers, and studying the choice of other word embedding tools that can capture the minute differences in written Arabic amongst age groups.

References

1. Robbins H, Monro S. A stochastic approximation method. The annals of mathematical statistics. 1951 Sep 1:400-7.
2. Estival, D., Gaustad, T., Hutchinson, B., Pham, S.B. and Radford, W., 2008. Author profiling for English and Arabic emails.
3. R. Rehurek and P. Sojka, "Software framework for topic modeling with large corpora," in In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.
4. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
5. Conroy NJ, Rubin VL, Chen Y. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology. 2015;52(1):1-4.
6. Abu Bakr Soliman, Kareem Eisa, and Samhaa R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP", in proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, UAE, 2017.
7. 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. Working Notes Papers of the CLEF. 2018.
8. Rangel F, Charfi A, Rosso P, Zaghouani W. Detecting Deceptive Tweets in Arabic for Cyber-Security.
9. Overview of the Track on Author Profiling and Deception Detection in Arabic. Francisco Rangel, Paolo Rosso, Anis Charfi, Wajdi Zaghouani, Bilal Ghanem, Javier Sánchez-Junquera. In: Mehtha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org, Kolkata, India, December 12-15.
10. `http://arap.qatar.cmu.edu/`
11. Al-Harahsheh, A.M.A., 2014. Language and gender differences in Jordanian spoken Arabic: a sociolinguistics perspective. Theory and Practice in Language Studies, 4(5), p.872.
12. Ibrahim, Z., 2009. Beyond lexical variation in modern standard Arabic: Egypt. Lebanon and.