A New Approach for Author Profiling and Identification of Deception in Texts

Hamada A. Nayel
Faculty of Computers and Artificial Intelligence
Benha University, Egypt
{hamada.ali@fci.bu.edu.eg}

ABSTRACT: *In this paper, we describe the methods and experiments that have been used in development of our system for Author Profiling and Deception Detection in Arabic shared task. There are two tasks, Author Profiling in Arabic Tweets and Deception Detection in Arabic Texts. We have submitted three runs for each task. The proposed system depends on classical machine learning approaches namely Linear Classier, Support Vector Machine and Multilayer Perceptron Classier. Bag-of-Word with range of n-grams model has been used for feature extraction. Our submissions for the first task achieved the second, seventh and third ranks. For the second task, one of our submissions outperformed all other submissions developed by other teams.*

Keywords: Arabic NLP, Author Profiling, Deception Detection

Received: 3 October 2019, Revised 24 January 2020, Accepted 5 February 2020

DOI: 10.6025/jcl/2020/11/2/73-79

1. Introduction

The tremendous usage of social platforms makes analysing shared contents a crucial task. One of the key tasks is Author Profiling (AP), which aims at predicting author attributes such as native language, gender, or political attitude [14]. AP has gained a lot of interest, due to it's applications in different areas such as E- commerce, Cyber-Security and forensics. In E-commerce, companies may analyze online reviews to improve targeted advertising. Analysing online reviews helps companies to improve their marketing strategy by knowing the demographics of people (gender and age) whose liked or disliked their products [14]. In Cyber- Security, AP can be used for detection of different crimes such as phishing, Cyber-blackmailing and Cyber-bullying. In forensics, profile of authors could be used as valuable additional evidence in criminal investigations [18].

Arabic is an important language having a huge number of native and nonnative speakers. The research in Natural Language Processing for Arabic language is continuously increasing. Applying NLP tasks for Arabic is a challenge due to different aspects of Arabic such as orthography, morphology, dialects, short vowels and word order [1]. AP has been studied for for English, Spanish and Arabic in [14], Indian languages [6] and Russian [5].

In this paper, we describe the model submitted for Author Profiling and Deception Detection in Arabic (APDA) shared task [16]. Shared task comprises of two tasks AP in Arabic Tweets and Deception Detection in Arabic texts. The first task identifies three attributes of Arabic Twitter users namely, age, gender and language variety. The second task detects the deception in Arabic texts.

2. Related Work

Author profiling is an important task that involves a lot of challenges and hitches. Many research works have been done on author profiling in different languages. The following are a brief about some of these works in recent years.

Different research areas such as psychology, linguistics and NLP have studied the relation between linguistics features and profile of the corresponding authors. The relation between language use and the personality traits has been studied by Pennebaker et al. [15]. They studied how variations of linguistic features in a text can provide information regarding the profile of its author. Author profiling task at PAN 2013 aimed at identifying age and gender of the author [17]. A large corpus collected from social media both in English and Spanish has been used for PAN 2013. In PAN 2014 [13], a compiled a corpus of four different genres, namely social media, blogs, Twitter, and hotel reviews has been used. Rangel et al. [12], organized the third author profiling task for age, gender and personality prediction. English, Spanish, Dutch and Italian languages were considered in this task. Different features have been used by participants in model design such as BoW, n-grams, frequencies and punctuations.

Nayel and Shashirekha [9, 10] have been designed a model for Native Language Identification for Indian languages. They used Term Frequency/Inverse Document Frequency (TF/IDF) with range of n-grams as feature extraction approach. They investigated different classification algorithms such as SVM, multinomial Naive Bayes, ANN-based classier and ensemble based classifier.

The research works that have been done for age and gender identification in the Arabic are rare [19]. Estival et al. [4] studied the age and gender identification problem as well as level of education in English and Arabic emails. For Arabic, they collected 8,028 emails from 1,030 native speakers of Egyptian Arabic. Several classifiers, such as SVM, KNN and decision trees combined with chi-square and information gain, have been tested to develop the Text Attribution Tool (TAT). They achieved accuracies of 72.10% and 81.15% for gender and age identification respectively.

Alsmearat et al [2] investigated gender identification in 500 articles collected from well-known Arabic newsletters. Articles written in Modern Standard Arabic (MSA) have been collected from writers with similar academic profiles and experience in journalistic writings. They applied different classification algorithms using BoW, sentiments and emotions as a feature set to train the classifiers.

In this work, we applied an effective ML-based approach using a simple TF/IDF features for the APDA shared task.

3. Task Description and Corpora

APDA shared task consists of two main tasks, the following are the descriptions of both tasks,

Task 1. Author Profiling in Arabic Tweets which aims at identifying author personality such as gender, age and language variety of Arabic Twitter users. In this subtask given a twitter written in Arabic, the system predict the age range, gender and language variety of twitter writer. There are three categories of age, under 25, between 25, and 35 and above 35. For language variation, 15 Arabic varieties have been considered namely Algeria (AL), Egypt (EG), Iraq (IR), Kuwait (KW), Lebanon-Syria (LS), Libya (LI), Morocco (MO), Oman (OM), Palestine-Jordan (PJ), Qatar (QA), Saudi Arabia (SA), Sudan (SU), Tunisia (TU), United Arab Emirates (UAE), Yemen (YE).

The training corpus for this task consists of tweets in Arabic, labeled with age, gender and language variety. This corpus is divided into five sub-corpora: dz-ag-iq (for AL, EG and IR), kw-lbsy-ly (for KW, LS and LI), ma-om-psjo (for MO, OM and PJ), qa-sa-sd (for QA, SA and SU) and tn-uae-ye (for TU, UAE and YE).

Task 2. Deception Detection in Arabic Texts which detects the deception in Arabic. The text is annotated with credible or non-credible label. There are two genre of data Twitter and news headlines.

The training corpus consists of this task contains two different genres. The first one is Twitter, a set of tweets written in Arabic collected and annotated with credible and non-credible labels. The second genre is news headlines, some news headlines are collected from news agencies and labeled with credible and non-credible labels.

4. Approaches

A detailed description of our model and the classification algorithms have been used are given in this section.

4.1 Problem Formulation

Given a set of segments of text such as a tweet, comment or news headline $S = \{s_1, s_2, ..., s_n\}$ and each segment is composite of a set of tokens or words 4 Hamada A. Nayel $s_i = \{w_1, w_2, ..., w_k\}$. Consider a set of 15 language varieties as described above $L = \{AL, EG, IR, KU, LS, LI, MO, OM, PJ, QA, SA, SU, TU, UAE, YE\}$.

Assume that, the set $A = \{UN, BE, AB\}$ represents the age categories under 25 (UN), between 25, and 35 (BE) and above 35 (AB). In addition, the set $G = \{M, F\}$ represents male and female respectively. Then, we can formalize each subtask as follows:

Task 1. It has been formalized as a multi-label classification problem. A multi-label classification is a classification problem where, the instance can be assigned with multiple class labels. Given an instance $s_k \in S$, we have to assign the triple $< g, a, l >$ such that, $g \in G$, $a \in A$ and $l \in L$.

Task 2. It has been formalized as a simple binary classification problem. Given a text, the model should decide whether this text is deception or not.

4.2 Model

Our model consists of the following steps:

4.2.1 Preprocessing

Preprocessing is a key step in building models for Arabic language. In this step, each tweet s_k has been tokenized into a set of words or tokens to get n-gram bag of words. the following processes have been implemented to each tweet:

Punctuation Elimination We removed punctuation marks such as {'+', '–', '#', '$'..}, which are increasing the dimension of feature space with redundant features. Example of redundancy, the following tokens {#الزمالك , الزمالك} pronounced "Al Zamalek (a famous football team in Egypt)", are the same with extra # which produces redundant features.

Tweet Cleaning Twitter users usually do not follow the standard rules of the language especially Arabic language. A common manner of users is to repeat a specific letter in a word. Cleaning the tokens from this redundant letters helps in feature space reduction. In our experiments, the letter is assumed to be redundant if it is repeated more than two times. For example the words "هههههه" ("hahahah" i.e. giggles) and "عاااااجل" (i.e. "urgent") containing redundant letter and will be reduced to "هه" and "عاجل" respectively.

4.2.2 Feature Extraction

TF/IDF with range of n-grams has been used to represent tweets as vectors. If $<w_1, w_2, ..., w_k>$ are the tokenized words in a tweet T_j, the vector associated to the tweet T_j will be represented as $<v_{j1}, v_{j2}, ..., v_{jk}>$ where v_{ji} is the weight of the token w_i in tweet T_j which is calculated as:-

$$v_{ji} = tf_i * log\left(\frac{N+1}{df_i+1}\right)$$

where tf_{ji} is the total number of occurrences of token w_i in the tweet T_j, df_i is the number of tweets in which the token w_i occurs and N is the total number Author Profiling and Deception Detection in Arabic Texts 5 of tweets. We used range of 2-grams model, i.e. unigram and bigram. For example sentence "السيتى مايستحقوش الخساره", which means the City does not worth loss (the City

refers to Manchester City football team)) has follow ing set of features {"السيتى", "مايستحقوش", "الخساره", "السيتى مايستحقوش", "مايستحقوش الخساره"}.

4.2.3 Training the Classier

Three classifiers have been trained for our model namely Linear classier, SVM and Multilayer Perceptron [20]. Linear classifier uses a set of linear discriminant functions to distinguish between different classes [20]. Linear classier is a simple and computationally effective approach. SVM is a linear classier which uses training samples or vectors close to the boundaries of classes as support vectors. SVM implemented for different NLP tasks effectively [11, 8]. Multilayer Perceptron (MLP) is a feed-forward ANN that characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses back propagation for training the network. MLP is a deep learning method.

4.3 Performance Evaluation

1. Task 1.

Accuracy is used to evaluate the performance of systems for this task. Individual accuracies will be calculated for each subtask (age, gender, language variety). Systems will be ranked by the joint accuracy (when all subtasks are properly identified together).

2. Task 2.

The performance of systems developed for this task will be evaluated using the macro-averaged measures (precision, recall and F1-score) and systems will be ranked by F1-score.

5. Experiments and Results

Classi er			Dataset				
			dz-ag-iq	kw-lbsy-ly	ma-om-psjo	qa-sa-sd	tn-uae-ye
Linear Classi er	Gender	Mean STD	79.78% 3.94%	78.22% 4.19%	82.22% 4.39%	84.00% 2.29%	72.00% 3.54%
	Age	Mean STD	58.89% 3.06%	64.22% 5.68%	53.33% 5.44%	64.67% 3.68%	51.33% 3.25%
	Country	Mean STD	99.33% 0.89%	99.55% 0.54%	98.44% 1.66%	96.67% 2.43%	97.33% 1.66%
MLP	Gender	Mean STD	74.67 % 4.30 %	73.78% 5.19 %	76.22% 4.07 %	82.00% 4.30%	70.89% 5.42%
	Age	Mean STD	57.11 % 3.56 %	64.00% 5.38 %	52.89% 2.49 %	62.22% 3.30%	52.89% 3.34%
	Country	Mean STD	98.22 % 1.13 %	98.89 % 0.70 %	97.56 % 1.78 %	94.00% 2.86%	94.89% 2.59%
SVM	Gender	Mean STD	79.33% 4.48 %	76.67% 5.07%	80.00% 4.77%	83.78% 2.59%	72.22% 3.14 %
	Age	Mean STD	56.00% 2.78%	63.78 % 6.19%	52.67% 2.68 %	64.22% 4.00%	52.00 % 4.00%
	Country	Mean STD	99.33% 0.89%	99.56% 0.54%	97.78% 2.11%	96.00% 2.39%	97.33% 1.66%

Table 1. 5-fold Cross-Validation accuracies for all classifiers for task 1

We have designed a model and used different classification algorithms for the different submissions. The classification algorithms are linear classifier, SVM and MLP classiers. Stochastic gradient descent optimization algorithm has been used for linear classier. Linear kernel has been applied for SVM kernel. The number of neurons in hidden layer is 20 neurons and the logistic function has been used as activation function for MLP classifier.

Separate models have been trained for each subtask (age, gender and language variety), then the output have been combined. While training the classifiers, 5-fold cross-validation technique has been used. The cross validation accuracies of all classification approaches for task 1 and task 2 are given in Table 1 and Table 2 respectively.

In Table 1, we highlighted the best reported accuracies for each sub-corpus and subtask. It is clear that, linear classier reported the best accuracies for the majority of subtasks and sub-corpora. While, MLP gives best accuracy for the age subtask of tn-uae-ye corpus.

Among 28 submissions of task 1, our submissions achieved 2nd, 3rd and 7th ranks as shown in Table 3. It is clear that linear classier reported the best performance among all of our submissions. There are 25 submissions for task 2 and our submissions achieved 1st, 2nd and 6th ranks as shown in Table 3. Our submission based on SVM approach outperforms all 25 submissions of all teams.

		NEWS	**TWITTER**
Linear Classifier	Mean STD	74.70% 2.95%	75.94% 2.17%
MLP	Mean STD	74.36% 2.89%	77.63% 2.74%
SVM	Mean STD	74.49% 3.37%	78.94% 3.51%

Table 2. 5-fold Cross-Validation accuracies for all classifiers for task 2

	Task 1				
Classi cation Algorithm	**Rank**	**Gender**	**Age**	**Variety**	**Joint**
Linear Classifier	2	81.53%	57.08%	97.50%	44.86%
SVM	3	80.14%	57.92%	97.08%	44.86%
MLP	7	76.67%	57.64%	95.97%	41.94%
	Task 2				
	Rank	**NEWS**	**TWITTER**	**AVERAGE**	
SVM	1	75.42%	84.64%	80.03%	
Linear Classifier	2	74.17%	84.63%	79.40%	
MLP	6	71.33%	83.37%	77.35%	

Table 3. Performance Evaluation of our Model for Test Data

References

[1] Alayba, A. M., Palade, V., England, M., Iqbal, R. (2018). Improving sentiment analysis in arabic using word representation. 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR) (March 2018). https://doi.org/10.1109/asar.2018.8480191

[2] Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., Kanaan, G. (2015). Emotion analysis of arabic articles and its impact on identifying the author's gender. *In*: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA). p. 1-6 (Nov 2015). https://doi.org/10.1109/AICCSA.2015.7507196

[3] Cappellato, L., Ferro, N., Jones, G. J. F., SanJuan, E. (2015). Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015, CEUR Workshop Proceedings, vol. 1391. CEUR-WS.org (2015), http://ceur-ws.org/Vol-1391

[4] Estival, D., Gaustad, T., Pham, S. B., Radford, W. (2007). Profiling for english emails.

[5] Litvinova, T., Gudovskikh, D., Sboev, A., Seredin, P., Litvinova, O., Pisarevskaya, D., Rosso, P. (2017). Author gender prediction in russian social media texts. *In*: van der Aalst, W. M. P., Khachay, M. Y., Kuznetsov, S. O., Lempitsky, V. S., Lomazova, I. A., Loukachevitch, N. V., Napoli, A., Panchenko, A., Pardalos, P. M., Savchencko, A. V., Wasserman, S., Ignatov, D. I. (eds.) Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017), Moscow, Russia, July 27 - 29, 2017. CEURWorkshop Proceedings, vol. 1975, p. 105-110. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1975/paper12.pdf

[6] Majumder, P., Mitra, M., Mehta, P., Sankhavara, J. (2017). Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017, CEUR Workshop Proceedings, vol. 2036. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2036

[7] Mehta, P., Rosso, P., Majumder, P., Mitra, M. (2018). Working Notes of FIRE 2018 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 6-9, 8 Hamada A. Nayel 2018, CEURWorkshop Proceedings, vol. 2266. CEUR-WS.org (2018), http://ceurws.org/Vol-2266

[8] Nayel, H. A., Shashirekha, H. L. (2017). Improving NER for clinical texts by ensemble approach using segment representations. *In*: Bandyopadhyay, S. (ed.) Proceedings of the 14th International Conference on Natural Language Processing, ICON 2017, Kolkata, India, December 18-21, 2017. p. 197-204. NLP Association of India (2017), https://aclweb.org/anthology/papers/W/W17/W17-7525/

[9] Nayel, H. A., Shashirekha, H. L. (2017). Mangalore-university@inli-re-2017: Indian native language identification using support vector machines and ensemble approach. *In*: Majumder et al. [6], p. 106-109, http://ceur-ws.org/Vol-2036/T4-2.pdf

[10] Nayel, H. A., Shashirekha, H. L. (2018). Mangalore university inli@re2018: Artificial neural network and ensemble based models for INLI. *In*: Mehta et al. [7], p. 110-118, http://ceur-ws.org/Vol-2266/T2-10.pdf

[11] Nayel, H. A., Shashirekha, H. L., Shindo, H., Matsumoto, Y. (2019). Improving multi- word entity recognition for biomedical texts. CoRR abs/1908.05691 (2019), http://arxiv.org/abs/1908.05691

[12] Pardo, F. M. R., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. *In*: Cappellato et al. [3], http://ceur-ws.org/Vol-1391/inv-pap12-CR.pdf

[13] Pardo, F. M. R., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Ver- hoeven, B., Daelemans, W. (2014). Overview of the author proling task at PAN 2014. *In*: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes for CLEF 2014 Conference, Sheeld, UK, September 15-18, 2014. CEUR Workshop Proceedings, vol. 1180, p. 898-927. CEUR-WS.org (2014), http://ceur-ws.org/Vol 1180/CLEF2014wn-Pan-RangelEt2014.pdf

[14] Pardo, F. M. R., Rosso, P., Montes-y-G–omez, M., Potthast, M., Stein, B. (2018). Overview of the 6th author profiling task at PAN 2018: Multimodal gender identication in twitter. *In*: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2125/invited paper 15.pdf

[15] Pennebaker, J. W., Mehl, M. R., Niederhoer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54 (1) 547-577.

[16] Rangel, F., Rosso, P., Char, A., Zaghouani, W., Ghanem, B., Sanchez-Junquera, J. (2019). Overview of the track on author

profiling and deception detection in arabic. *In*: Working Notes of Forum for Information Retrieval Evaluation FIRE 2019, Kolkata, India, December 12-15, 2019. CEURWorkshop Proceedings, CEUR-WS.org.

[17] Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. (2013). Overview of the author profiling task at pan 2013. *In*: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. p. 352-365. CELCT.

[18] Rosso, P., Pardo, F. M. R., Ghanem, B., Char, A. (2018). ARAP: arabic author profiling project for cyber-security. Procesamiento del Lenguaje Natural 61, 135-138 (2018), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5654

[19] Rosso, P., Rangel, F., Far–as, I. H., Cagnina, L., Zaghouani, W., Char, A. (2018). A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass* 12 (4), e12275 (2018), https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12275

[20] Theodoridis, S., Koutroumbas, K. (2009). Chapter 3 - linear classifiers. *In*: Pattern Recognition (Fourth Edition), p. 91 - 150. Academic Press, Boston, fourth edition edn. (2009), http://www.sciencedirect.com/science/article/pii/B9781597492720500050