Detecting Deceptive Tweets in Arabic for Cyber-Security

Francisco Rangel
Universitat Politècnica de València, Spain
fraranpa@prhlt.upv.es
Anis Charfi
Carnegie Mellon University, Qatar
acharfi@andrew.cmu.edu

Paolo Rosso
Universitat Politècnica de València, Spain
prosso@dsic.upv.es
Wajdi Zaghouani
Hamad Bin Khalifa University, Qatar
wzaghouani@hbku.edu.qa

***Abstract*—In the framework of the QNRF project on Arabic Author Profiling for Cyber-Security, we addressed deception detection in Arabic in order to discard those messages that do not really represent potential threats. We have applied the Low Dimensionality Statistical Embedding (LDSE) method to several corpora for Arabic including the Arabic credibility corpus and two new corpora that we created: the Qatar Twitter corpus and the Qatar News corpus. We achieved a performance of 0.797 Macro F-measure on the Arabic Credibility corpus. The obtained results with two well-known distributed representations, namely Continuous Bag of Words and Skip Grams, showed the competitiveness of our approach. The LDSE approach gave similar results on the two corpora that we created. We evaluated our work in a cross-genre scenario, showing the robustness of LDSE when there are enough data about similar topics.**

***Index Terms*—deception detection, Arabic, cyber-security, Twitter**

I. Introduction

Social media allow people to communicate beyond their environment and geographical boundaries. According to the Rheingold's [7] theory of the smart mobs, thanks and through social media, thousands (or even millions) of people can agree on concrete issues and carry out coordinated actions. The power of social media promotes the democratization of information and influence. Both have been transferred from a few (traditional) media such as newspapers or television, to the citizens. Social media are especially important in countries that are under censorship since the anonymity may empower the citizens' journalism and activism. Nevertheless, social media anonymity may induce people with not so lawful interests to hide behind their devices. These users can promote vandalism, spread threatening messages, or even terrorist propaganda.

In the framework of the project Arabic Author Profiling for Cyber-Security (ARAP) [1], we aim at preventing cyber-threats using machine learning (Figure 1). To this end, we monitor social media to early detect threatening messages and, in such a case, to profile the authors behind. Profiling potential terrorists [10] from messages shared in social media may allow detecting communities whose aim is to undermine the security of others. Nonetheless, we must be aware of false positives, i.e., potential threatening messages that are actually deceptive, ironic or humorous[2]. In this work we focus on detecting deceptive messages in Arabic.

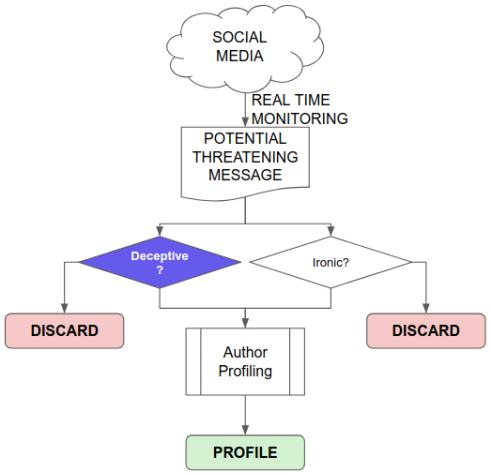

Fig. 1. Workflow of the Arabic author profiling for cyber-security project.

II. Related work

We can consider that a message is deceptive when it is intentionally written trying to sound authentic. Originally, the main focus of deception detection was to detect spam in opinion reviews [8]. Nowadays, the interest is shifting towards fake news detection, for example, in the context of fact check shared task[3] at CLEF[4] on automatic identification and verification of claims in political debates [5].

Deception detection research in Arabic is very limited [9]. Despite the fact that the aforementioned shared task also included Arabic, the contents were translated from English. Since the claims corresponded to US politics, they are not representative of the idiosyncrasy of Arabs. In this sense, the authors in [1] collected a corpus in Arabic from 600 tweets and 179 news articles. They automatically annotated

[1] arap.qatar.cmu.edu

[2] https://qz.com/1107023/the-inside-story-of-the-hack-that-nearly-started-another-middle-

[3] http://alt.qcri.org/clef2018-factcheck

[4] http://clef2018.clef-initiative.eu/

the credibility by measuring the cosine similarity between the tweets and the news articles. The authors in [2] complained about the automatic generation of the annotation and they collected and manually annotated two corpora from Twitter and Blogs. Regarding Twitter, they retrieved over 36 million tweets about four topics: *i)* The forces of the Syrian government; *ii)* Syrian revolution; *iii)* Syrian problems and concerns related to the Syrian revolution; and *iv)* The election of the Lebanese president. The annotation process was carried out by five annotators. According to the authors in [12] the obtained inter-annotator agreement (Fleiss' kappa 0.43) was moderate. The authors also proposed a method to approach the credibility analysis of Twitter contents. The Credibility Analysis of Arabic Content on Twitter (CAT) [3] relies mainly on features obtained from the user who tweeted the content to be analysed. For example, the authors retrieved the user's timeline and extracted features such as the number of retweets, the user's activity, or the user's expertise in the topic being discussed. They compared their approach with several baselines and showed a significant improvement.

III. Low Dimensionality Statistical Embedding

In the framework of our ARAP project we aim at early detecting potentially threatening messages and at profiling their authors. This implies real time retrieval and analysis of such messages and their authors, i.e., our system has to work in a big data environment. Therefore, we must bear in mind several considerations. For example, most of the time, the textual content (e.g., the tweet) is the only available data, and due to time restrictions, it is not feasible to retrieve the users' history (e.g., the timeline) and/or other complementary data. Furthermore, due to the lack of linguistic resources for the Arabic language, we opted for a language-independent approach.

We proposed the Low Dimensionality Statistical Embedding (LDSE) [6] to represent documents on the basis of the different use of the words depending on the class, in our case, when the user lies or not. The key concept is a weight that represents the probability of a term to belong to one of the two classes: credible or non-credible. Hence, the distribution of weights for a given document should be closer to the weights of its corresponding class. Formally, we represent the documents following the next three steps:

Step 1. We calculate the *tf-idf* weights for the terms in the training set *D* and build the matrix Δ, where each row represents a document , each column a vocabulary term *t*, and each cell represents the *tf-idf* weight w_{ij} for each term in each document. Finally, $\delta(d_i)$ represents the assigned class *c* to the document .

$$\Delta = \begin{bmatrix} w_{11} & w_{12} & ... & w_{1m} & \delta(d_1) \\ w_{21} & w_{22} & ... & w_{2m} & \delta(d_2) \\ ... & ... & ... & ... & \\ w_{n1} & w_{n2} & ... & w_{nm} & \delta(d_n) \end{bmatrix}, \quad (1)$$

Step 2. Term weights $W(t,c)$ are obtained as the ratio between the weights of the documents belonging to the class c and the total distribution of weights for that term (as shown in Eq. 2).

$$W(t,c) = \frac{\sum_{d \in D/c=\delta(d)} w_{dt}}{\sum_{d \in D} w_{dt}}, \forall d \in D, c \in C \quad (2)$$

Step 3. We represent the documents with the obtained weights $W(t,c)$ following Eq. 3:

$$d = \{F(c_1), F(c_2), ..., F(c_n)\} \sim \forall c \in C, \quad (3)$$

Each $F(c_i)$ contains the set of features shown in Eq. 4, with the following meaning: *i)* average value of the document term weights; *ii)* standard deviation of the document term weights; *iii)* minimum value of the weights in the document; *iv)* maximum value of the weights in the document; *v)* overall weight of a document as the sum of weights divided by the total number of terms of the document; and *vi)* proportion between the number of vocabulary terms of the document and the total number of terms of the document.

$$F(c_i) = \{avg, std, min, max, prob, prop\} \quad (4)$$

This representation statistically embeds the distribution of weights of the document terms. The dimensionality reduction is drastic to only six features per class. In our two-class problem, the dimensionality is reduced to only twelve features. Thus, representing a document is very fast, which makes LDSE suitable to process big data.

IV. Methodology

In this section, we describe the methodology followed to verify that LDSE is suitable for deception detection in Arabic and competitive with state-of-the-art approaches. Firstly, we present the different corpora we used. Then, we describe the alternative methods that we compared with LDSE.

A. Corpora

Besides the Credibility corpus [3], we have created two new corpora in the context of the ARAP project: the Qatar Twitter corpus and the Qatar News corpus.

1) The Credibility corpus: The Credibility corpus was collected from Twitter and annotated with the help of five annotators. The authors retrieved more than 36 million tweets from several queries related to the Syrian political situation in 2010. Statistics about that corpus are shown in Table I. Topic wise, the corpus contains tweets about two main topics: the Syrian government and the Syrian revolution. For both topics, there is a considerable imbalance between the number of credible and non-credible tweets.

Topic	Credible	Non-Credible	Total
Syrian Government	1,131	510	1,641
Syrian Revolution	439	628	1,067
Combined	1,570	1,138	2,708

TABLE I
DISTRIBUTION OF CREDIBLE VS. NON-CREDIBLE TWEETS FOR EACH TOPIC IN THE CREDIBILITY CORPUS.

2) The Qatar Twitter corpus: In the context of the ARAP project, we created the Qatar Twitter corpus by retrieving during 2017 and annotating[5] tweets referring to the Qatar Blockade and the Qatar World Cup. Statistics about this corpus are shown in Table II. The number of tweets for the blockade topic is completely balanced between credible and non-credible classes. For the World Cup topic the corpus is almost balanced, with a slightly smaller amount of credible tweets (48% / 52%).

Corpus	Topic	Credible	Non Credible	Total
Qatar Twitter	Blockade	115	115	230
	World Cup	262	281	543
	Total	377	396	773
Qatar News		889	999	1,888

TABLE II
DISTRIBUTION OF CREDIBLE AND NON-CREDIBLE TWEETS PER TOPIC IN QATAR TWITTER CORPUS AND QATAR NEWS CORPUS.

3) The Qatar News corpus: We also created the Qatar News corpus by retrieving and annotating short contents such as headlines and/or excerpts from well-known Arabic newsletters. Statistics on this second corpus are shown in Table II. The number of documents is almost balanced, with a slightly smaller amount of credible news (47% / 53%).

B. Methods

The authors of the deception corpus [3] proposed the Credibility of Arabic Tweet (CAT) method to address the task. The authors retrieved the timeline of the user who created the tweet. Then, they obtained 22 user-based features such as the expertise of the Twitter user in the discussed topic or her/his activity. The authors combined also 26 content-based features such as the number of retweets, the number of URLs, and the tweet sentiment. They tested several machine learning algorithms and reported that the best results were obtained with Random Forest. The authors also compared CAT with three baselines: *i)* the stratified baseline that randomly predicts the credibility according to the distribution of the classes in the training set; *ii)* the uniform baseline that randomly predicts the credibility following a uniform distribution; *iii)* the majority baseline that predicts all the tweets to belong to the majority class in the training set.

For our work, we also compare LDSE with two state-of-the-art approaches based on Arabic word embeddings [11]: Continuous Bag of Words (CBOW) and Skip-Grams (SG)[6]. For the three methods we have experimented with several machine learning algorithms and will report in the following the best performing one in each case. In most cases, SVM and Multilayer Perceptron gave the best results.

V. EXPERIMENTS AND DISCUSSION

In this section, we report and discuss the obtained results. For each corpus, we compare LDSE with the presented alternative methods and obtain the statistical significance of their different performance. To evaluate the results, we use the macro-averaged measures (precision, recall and F1-score) for two reasons: *i)* due to the imbalance of the data, they gave the same importance to the different classes no matter their size; and *ii)* in order to ensure comparability to previous investigations.

A. Results on the Credibility Corpus

Figure 2 shows the results we obtained when using the Credibility corpus. It can be observed that LDSE gives lower results than CBOW and SG for both topics, although it outperforms them on the full corpus. All three methods outperform CAT and they also provide much better results than the baselines.

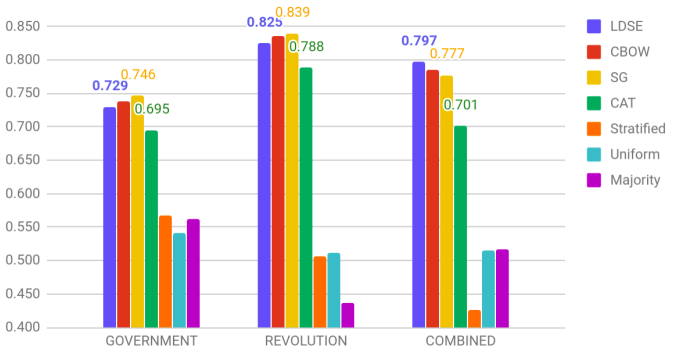

Fig. 2. Results in terms of Macro-F measure for the different methods and the different topics in the Credibility corpus.

Topic	Method	Precision	Recall	F
Syrian Government	LDSE	0.739	0.728	0.729
	CBWO	0.757	0.759	0.738
	SG	**0.768**	**0.767**	**0.746**
	CAT	0.696	0.717	0.695
	Stratified	0.566	0.568	0.567
	Uniform	0.590	0.523	0.541
	Majority	0.475	0.689	0.562
Syrian Revolution	LDSE	0.840	0.817	0.825
	CBWO	0.835	0.835	0.835
	SG	**0.846**	**0.842**	**0.839**
	CAT	0.790	0.791	0.788
	Stratified	0.505	0.511	0.507
	Uniform	0.524	0.508	0.512
	Majority	0.346	0.589	0.436
Combined	LDSE	**0.800**	**0.796**	**0.797**
	CBWO	0.789	0.788	0.785
	SG	0.781	0.780	0.777
	CAT	0.701	0.703	0.701
	Stratified	0.336	0.580	0.426
	Uniform	0.523	0.512	0.515
	Majority	0.516	0.518	0.517

TABLE III
RESULTS IN TERMS OF MACRO-PRECISION, MACRO-RECALL AND MACRO-F MEASURE FOR THE DIFFERENT METHODS AND THE DIFFERENT TOPICS OF THE CREDIBILITY CORPUS. IN BOLD ARE THE BEST RESULTS FOR EACH MEASURE.

The detailed results for the credibility corpus are shown in Table III. The obtained results on the Syrian revolution topic are much higher than the ones obtained on the Syrian Government topic. Concretely, 0.078, 0.075 and 0.093 respectively for

[5]For both the Qatar Twitter and Qatar News corpora, the annotators were 20 students at Hamad Bin Khalifa University, representing various Arabic countries. The inter-annotator agreement was about 80%.

[6]Word embedding are averaged to obtain the text embeddings.

precision, recall and F-measure in case of the best performing approach (SG). In case of LDSE, these differences are even greater, concretely, 0.101, 0.089, 0.096 respectively for the three measures. This may be due, notwithstanding there are more data in case of the Syrian Government, to the greater imbalance of the data for the topic Syrian revolution (69% / 31% vs. 41% / 59%).

On the two topics, SG outperforms LDSE and CBOW with respect to all measures. In case of the Syrian Government, these results are significantly higher for precision and recall, but they are lower for F-measure. In case of the Syrian Revolution, although the results are better for SG, they are not statistically significant with respect to LDSE (Table IV).

Topic	Method	Precision	Recall	F
Syrian Government	CBWO	-1.1876	-1.8331**	-0.5831
	SG	-1.9275*	-2.5714*	-1.1067
Syrian Revolution	CBWO	-0.3809	-1.0967	-0.6149
	SG	0.3131	-1.5355	-0.8649
Combined	CBWO	1.0017	0.7253	1.0860
	SG	1.7180**	1.4405	1.7975**

TABLE IV
SIGNIFICANCE (P-VALUES) WHEN COMPARING LDSE RESULTS WITH THE CONTINUOUS BAG OF WORDS AND SKIP-GRAMS REPRESENTATIONS IN THE CREDIBILITY CORPUS (*0.05; **0.01).

LDSE outperforms the two embeddings when using the whole corpus, albeit in most measures without statistical significance. Therefore, we can infer that the performance of LDSE depends on: *i)* the quantity of data to learn properly the representation weights; and *ii)* the imbalance of the data. In case of the full corpus, the proportion between credible and non-credible classes is 58% / 42%. In conclusion, in this corpus LDSE is competitive with the other two embedding-based representations and significantly outperforms previous works such as CAT.

B. Results on the Qatar Twitter Corpus

Figure 3 shows the results in terms of Macro F-measure on the Qatar Twitter corpus[7]. LDSE outperforms the embedding-based methods in case of the Qatar Blockade topic as well as for the full corpus, but it gives the worst results on the Qatar World Cup topic. The overall performance of LDSE in terms of F-measure is 0.797. The difference compared to the baselines is 0.297 for both the stratified and the uniform baselines, and it increases up to 0.449 for the majority baseline.

In Table V the detailed results for the different methods are shown. In case of the Qatar Blockade, LDSE provides the best results in precision and F-measure, although without statistical significance with respect to those obtained with the CBOW and SG embeddings. However, the SG method outperforms both LDSE and CBOW with respect to recall.

[7]We could not reproduce the CAT method, hence its results are not represented in the next experiments.

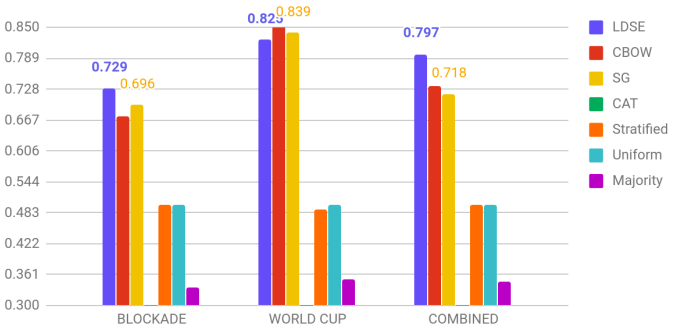

Fig. 3. Results in terms of Macro F-measure for the different methods on the Qatar Twitter corpus.

Topic	Method	Precision	Recall	F
Blockade	LDSE	**0.739**	0.728	**0.729**
	CBWO	0.675	0.674	0.674
	SG	0.696	**0.969**	0.696
	CAT	-	-	-
	Stratified	0.500	0.500	0.500
	Uniform	0.500	0.500	0.500
	Majority	0.500	0.250	0.335
World Cup	LDSE	0.840	0.817	0.825
	CBWO	**0.872**	**0.870**	**0.869**
	SG	0.841	0.839	0.839
	CAT	-	-	-
	Stratified	0.485	0.485	0.490
	Uniform	0.500	0.500	0.500
	Majority	0.517	0.269	0.352
Combined	LDSE	**0.800**	**0.796**	**0.797**
	CBWO	0.734	0.734	0.733
	SG	0.718	0.718	0.718
	CAT	-	-	-
	Stratified	0.500	0.500	0.500
	Uniform	0.500	0.500	0.500
	Majority	0.512	0.261	0.348

TABLE V
RESULTS IN TERMS OF MACRO-PRECISION, MACRO-RECALL AND MACRO-F MEASURE FOR THE DIFFERENT METHODS AND THE DIFFERENT TOPICS ON THE QATAR TWITTER CORPUS. IN BOLD ARE THE BEST RESULTS FOR EACH MEASURE. THERE ARE NO RESULTS FOR THE CAT METHOD.

It is worth to mention that these results are much lower than those obtained in case of the World Cup topic. For example, LDSE provides better results by 0.101, 0.089 and 0.096 on the World Cup topic respectively for each measure, and the difference between the best results is even higher in case of precision (0.133) and F-measure (0.140). These differences are difficult to understand taking into account the corpus size and balance. On the one hand, the Blockade topic is completely balanced, although the World Cup topic is only slightly imbalanced (48% / 52%). On the other hand, the size for the Qatar Blockade is approximately 42% of the World Cup topic size. Our intuition is that the different use of the lexicon makes the classes more separable in case of the Qatar Blockade.

Table VI shows the p-values when comparing LDSE with the other two embedding-based representations. As can be seen, there is no statistical significance for most of the results obtained for the two topics. However, LDSE outperforms the other two embedding-based methods on the combined corpus

with statistical significance. We can conclude that in this corpus, LDSE is competitive with the other two embedding-based representations, outperforming them with statistical significance when there are enough data.

Topic	Method	Precision	Recall	F
Blockade	CBWO	1.5079	1.2649	1.2889
	SG	1.0242	-7.2083*	0.7819
World Cup	CBWO	-1.5018	-2,4036*	-2.0140*
	SG	-0.0450	-0.9606	-0.6170
Combined	CBWO	3.0693*	2.8748*	2.9675*
	SG	3.7693*	3.5753*	3.6237*

TABLE VI

SIGNIFICANCE (P-VALUES) WHEN COMPARING LDSE RESULTS WITH THE CONTINUOUS BAG OF WORDS AND SKIP-GRAMS REPRESENTATIONS ON THE QATAR TWITTER CORPUS (*0.05; **0.01).

C. Results on the Qatar News Corpus

The results in terms of Macro F-measure on the Qatar News corpus can be seen in Figure 4. The overall performance obtained by LDSE and SG is 0.707, followed by CBOW with 0.706. The difference to the baselines is substantial. LDSE outperforms the stratified and uniform baselines by 0.205 and 0.211 respectively, and this difference increases up to 0.342 for the majority baseline.

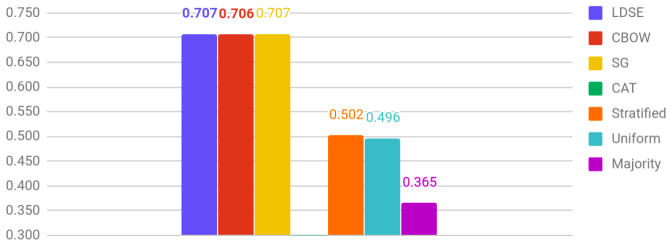

Fig. 4. Results in terms of Macro-F measure for the different methods on the Qatar News corpus.

In Table VII the detailed results for the three measures are provided. The best results for all the measures are obtained by LDSE and SG. LDSE provides the best result in precision (0.715) although without statistical significance over CBOW (0.707) and SG (0.710). Regarding recall and F-measure, LDSE ties with SG (0.709 and 0.707 respectively), which slightly outperform CBOW (0.707 and 0.706 respectively), without statistical significance.

Method	Precision	Recall	F
LDSE	**0.715**	**0.709**	**0.707**
CBWO	0.707	0.707	0.706
SG	0.710	**0.709**	**0.707**
CAT	-	-	-
Stratified	0.502	0.502	0.502
Uniform	0.484	0.516	0.496
Majority	0.529	0.280	0.365

TABLE VII

RESULTS IN TERMS OF MACRO-PRECISION, MACRO-RECALL AND MACRO-F MEASURE FOR THE DIFFERENT METHODS ON THE QATAR NEWS CORPUS. IN BOLD ARE THE BEST RESULTS FOR EACH MEASURE. THERE ARE NO RESULTS FOR THE CAT METHOD.

Table VIII shows the p-values when comparing LDSE with the other two embedding-based representations. None of them is statistically significant, with the highest difference in case of precision with respect to CBOW (0.5422). We can conclude that, for this corpus, LDSE is competitive when compared to the other two embedding-based approaches.

	Precision	Recall	F
CBWO	0.5422	0.1351	0.0675
SG	0.3394	0.0000	0.0000

TABLE VIII

SIGNIFICANCE (P-VALUES) WHEN COMPARING LDSE RESULTS WITH THE CONTINUOUS BAG OF WORDS AND SKIP-GRAMS REPRESENTATIONS ON THE QATAR NEWS CORPUS (*0.05; **0.01).

D. Results on the Cross-genre Deception Detection

In the following, we present the results obtained in a cross-genre scenario, that is, when training on one corpus and evaluating on a different one. This is a realistic scenario since in some cases we do not have (enough) labelled data in one genre to train our models (e.g., WhatsApp) while there may be (enough) labelled data available in another genre (e.g., Twitter). Therefore, we consider it important to investigate the robustness of LDSE also from a cross-genre perspective.

Train/Test	Method	Credibility	News	Tweets
Credibility	LDSE	**0.797**	**0.594**	**0.619**
	CBOW	0.785	0.583	0.503
	SG	0.777	0.584	0.527
News	LDSE	0.481	**0.797**	**0.661**
	CBOW	0.695	0.733	0.579
	SG	**0.737**	0.718	0.585
Tweets	LDSE	0.500	0.595	**0.707**
	CBOW	**0.578**	**0.609**	0.706
	SG	0.575	0.608	**0.707**

TABLE IX

RESULTS IN TERMS OF MACRO-F MEASURE IN A CROSS-GENRE SCENARIO. THE TRAINING CORPORA ARE INDICATED IN THE FIRST COLUMN AND THE TEST CORPORA IN THE FIRST ROW. IN BOLD ARE THE BEST RESULTS FOR EACH MEASURE.

The results are shown in Table IX. Each cell represents the Macro F-measure obtained when trained with the corpus mentioned in the first column (left) and tested on the corpus mentioned in the first row (up). When trained with the Credibility corpus, LDSE provides the best results on the Qatar News and Twitter corpora with high statistical significance (Table X). Similarly, when training on Qatar News and evaluating on Qatar Twitter, LDSE significantly outperforms the other two embedding-based representations. When the training corpus is the Qatar News and the approaches are evaluated on the Qatar Twitter, all of them perform with no statistical significance. Nonetheless, when the testing corpus is the Credibility one, no matter the training corpus, LDSE performs significantly worse than the other two methods. Our intuition is that the Credibility corpus and the two Qatar corpora are significantly different regarding terminology. Basically, the Credibility corpus refers mainly to Syrian affairs, whereas the other corpora refer to Qatar affairs.

Train/Test	Method	Credibility	News	Tweets
Credibility	CBOW	1.0860	0.6868	4.5954*
	SG	1.7975**	0.6245	3.6565*
News	CBOW	-15.9985*	4.6377*	3.3212*
	SG	-19.3042*	5.6633*	3.0830*
Tweets	CBOW	-5,7578*	-0.8788	0.0432
	SG	-5.5351*	-0.8158	0.0000

TABLE X
SIGNIFICANCE (P-VALUES) WHEN COMPARING LDSE RESULTS WITH THE CONTINUOUS BAG OF WORDS AND SKIP-GRAMS REPRESENTATIONS ON THE CROSS-GENRE SCENARIO (*0.05; **0.01).

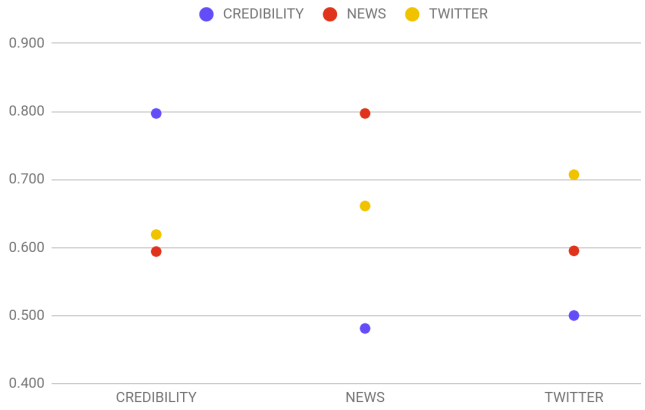

Fig. 5. Results of LDSE in terms of Macro F-measure in the cross-genre scenario. The Column lists the training corpus and color depicts the test corpus.

In Figure 5 we represent the obtained results by LDSE on the trhee corpora. In each column we represent the training corpus and the circle represents the result achieved on the evaluation corpus. Therefore, we can see the performance of LDSE on the different corpora, as well as in a cross-genre setting. When evaluating on the same corpus, the worst result is obtained on the Qatar Twitter corpus, with a Macro F-measure 0.09 lower. When evaluating on a different corpus, the performance decreases in all the cases. The worst results are obtained when evaluating on the Credibility corpus, with the highest decrease when training on the Qatar News corpus (0.316). When the training corpus is the Credibility one, the performance is very similar on the Qatar News corpus (0.619) and the Qatar Twitter corpus (0.594). The difference when training on the Qatar News corpus and evaluating on the Qatar Twitter corpus (0.136) is very similar to the opposite case (0.112). This may indicate some stability and robustness of LDSE in cross-genre scenarios when dealing with similar domains.

VI. CONCLUSIONS

In this work, we have addressed the deception detection in Arabic as part of the QNRF funded research project Arabic Author Profiling for Cyber-Security (ARAP).

Firstly, we used several Arabic corpora and two of them were created in the context of the ARAP project. Based on that, we have compared our LDSE approach with several state-of-the-art approaches including the distributed representations based on word embeddings: Continuous Bag of Words and Skip Grams. The obtained results on the Credibility corpus (0.797) showed the competitiveness of LDSE. We also experimented with two corpora that we created: the Qatar Twitter corpus and the Qatar News corpus. The obtained results show that LDSE provides better results when it works on enough data. Furthermore, LDSE is especially suitable for big data as it represents texts using only six features per class, i.e., twelve features in deception detection. Furthermore, LDSE is language independent and it does not need linguistic resources.

Finally, we have evaluated the methods in a cross-genre scenario. That is, we trained the models on one corpus and evaluated them on the other corpora. The obtained results show that LDSE significantly outperforms the embedding-based approaches in case of similar domains.

As future work, we will focus more on the cross-genre scenario due to its importance especially in case of cyber-security. Our objective is to investigate how to make LDSE more robust against changes in the domain.

ACKNOWLEDGMENT

This publication was made possible by NPRP 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

REFERENCES

[1] R.M.B. Al-Eidan, H.S. Al-Khalifa, A.S. Al-Salman. Measuring the Credibility of Arabic Text Content in Twitter. In 2010 Fifth International Conference on Digital Information Management (ICDIM), 2010

[2] A. Al Zaatari, R. El Ballouli, S. Elbassuoni, W. El-Hajj, H.M. Hajj, K.B. Shaban, N. Habash, E. Yahya. Arabic Corpora for Credibility Analysis. In: Language Resources and Evaluation Conference (LREC), 2016

[3] R. El Ballouli, W. El-Hajj, A. Ghandour, S. Elbassuoni, H. Hajj, K. Shaban. CAT: Credibility Analysis of Arabic Content on Twitter. In Proc. of the Third Arabic Natural Language Processing Workshop, 2017

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Jeff. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems, 2013

[5] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, and G. Da San Martino. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In: International Conference of the Cross-Language Evaluation Forum for European Languages, 2018.

[6] F. Rangel, P. Rosso, M. Franco. A Low Dimensionality Representation for Language Variety Identification. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing16), Springer-Verlag, LNCS(9624), pp. 156-169, 2018

[7] H. Rheingold. Smart mobs: the next social revolution. Basic books, 2007

[8] L. Cagnina, P. Rosso. Detecting Deceptive Opinions: Intra and Cross-Domain Classification Using an Efficient Representation. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 25, Suppl. 2, pp. 151–174, World Scientific, 2017

[9] P. Rosso, F. Rangel, I. Hernndez Farias, L. Cagnina, W. Zaghouani, A. Charfi. A Survey on Author Profiling, Deception, and Irony Detection for the Arabic Language. In: Language and Linguistics Compass, vol. 12 (4), 2018

[10] C. Russell, B. Miller. Profile of a Terrorist. Studies in Conflict & Terrorism, vol. 1 (1), pp. 17–34, Taylor & Francis, 1977

[11] A.B. Soliman, K. Eisa, S.R. El-Beltagy, AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. In: 3rd International Conference on Arabic Computational Linguistics (ACLing), 2017.

[12] A.J. Viera, J.M. Garrett. Understanding Interobserver Agreement: the Kappa Statistic. Fam med journal, 37(5), pp.360-363. 2005