An Arabic Corpus of Fake News: Collection, Analysis and Classification

Maysoon Alkhair, Karima Meftouh, Nouha Othman, Kamel Smaïli

An Arabic Corpus of Fake News: Collection, Analysis and Classification

Maysoon Alkhair[1], Karima Meftouh[2], Nouha Othman[3], and Kamel Smaili[4]

[1] Sudan University of Science and Technology
[2] University Badji Mokhtar, Annba, Algeria
[3] LARODEC, University of Tunis, Tunisia
[4] LORIA, University of Lorraine, F-54600, France
maysoonalkhier111@gmail.com, k.meftouh@gmail.com, othmannouha@gmail.com, smaili@loria.fr

Abstract. Over the last years, with the explosive growth of social media, huge amounts of rumors have been rapidly spread on the internet. Indeed, the proliferation of malicious misinformation and nasty rumors in social media can have harmful effects on individuals and society. In this paper, we investigate the content of the fake news in the Arabic world through the information posted on YouTube. Our contribution is threefold. First, we introduce a novel Arab corpus for the task of fake news analysis, covering the topics most concerned by rumors. We describe the corpus and the data collection process in detail. Second, we present several exploratory analysis on the harvested data in order to retrieve some useful knowledge about the transmission of rumors for the studied topics. Third, we test the possibility of discrimination between rumor and no rumor comments using three machine learning classifiers namely, Support Vector Machine (SVM), Decision Tree (DT) and Multinomial Naïve Bayes (MNB).

Keywords: Rumors, classifiers, Fake news corpus, Text analysis

1 Introduction

Social networks such as Facebook, Twitter, Google+ and YouTube have become popular channels of communication where people can express different attitudes and opinions [5]. Consequently, a vast volume of reviews and comments has been created in the last years in social networks. Obviously, anyone can express his opinion and related information, which leads to accumulation of a huge amount of unverified information [2]. This issue was widely studied by the NLP community with a view to differentiating between a rumor (or fake news) and a proven information.
Researchers proposed automated or semi-automated approaches which can effectively help in handling and analyzing the tremendous amount of social network data. Recently, there has been much focus on the veracity of the information by studying and proposing algorithms in order to automatically detect rumors in social networks. However, most works analyze and measure the rumor only after its diffusion. The issue is that there is an important gap between its diffusion and its streaming detection. This can lead to a damaging effect on the social or political events of a country or even the world. The

speed at which the breaking news is growing on the Internet does not allow enough time to check the information [8]. In order to analyse the rumors, the data are often extracted from Twitter, Facebook or YouTube [18, 3]. In fact, it is easier to spread a rumor in social networks since almost everything could be published.
Unlike most existing works which focus on identifying the rumors when they arise, in this paper, we investigate the content of the fake news in the Arabic world through the information posted on YouTube. The main objective of this work is to crawl Arabic rumors in order to build a corpus that we will share with the international community. We focused on three proven fake news concerning the death of personalities. We selected the death rumors of the following Arab celebrities: the dancer Fifi Abdou, the president Bouteflika and the comedian Adel Imam.
The remainder of this paper is structured as follows: in Section 2, we present related work on rumors analysis and rumors extraction. Then we give an overview of the rumors we collected and the way we categorize them. Thereafter, we give details about the collected corpus in Section 3. Several statistical analysis are described in section 4. In Section 5, we present some results of machine learning classification algorithms and finally we conclude and outline some possible future works.

2 Related Work

In this section, we provide an overview of research into social media rumours with the focus on two crucial tasks namely, rumors extraction and analysis.

The comparability methods were widely used to identify similar data related to same rumors when the dataset is collected.
Authors in [11] investigated how rumors are arising, spreading in different ways and broadcasting quickly to a large number of audiences. In [9], the authors proposed a statistical approach that uses 3 features extracted from the microblogs, the Hashtags and URLs. They showed the effectiveness of these features in identifying disinformers and those who believe and spread the rumors. They annotated a dataset of 10K tweets collected on 5 different controversial topics.
The authors in [1] proposed methods for assessing the credibility of certain tweets. They analyzed microblog posts and classified them as credible or not credible, based on some features extracted from the tweets. An example of the used feature is the number of retweeting performed by a user. They evaluated their methods subjectively and remarked that credible news are propagated through authors that have previously written a large number of messages.
The authors in [4] suggested determining whether or not a given text is a rumor by using web mining algorithms and linguistic rules. They evaluated their approach on customer reviews which constitutes a good framework of possible disinformation.
In [6], an approach was proposed to capture the temporal evolution of the features of the microblogs based on the time series that model the social context information. The approach showed significant performance and has proven to be able to detect rumors at early stage after their initial broadcast.
Tolosi et al. [15] studied the challenges concerning the detection of the tweets that are likely to become rumors. In their work, the classifier used several features such as

the user id, the user profile, the text style and the URL domains. The given classifier achieved an F1-score of 65%.
The authors in [17] introduced a novel approach to detect rumors that takes advantage of the sequential dynamics of publishing information during breakthroughs in social media. They employed Twitter datasets collected from five news stories. The classifier was based on Conditional Random Fields and exploited the context learned in a rumor detection event, which they compared to the rumor detection system at the same time.

3 Corpus

In the following, we will describe the methodology we followed to collect the necessary data for this research work. The Figure 1 illustrates the overall steps. The details concerning each of them will be given in the next subsections.

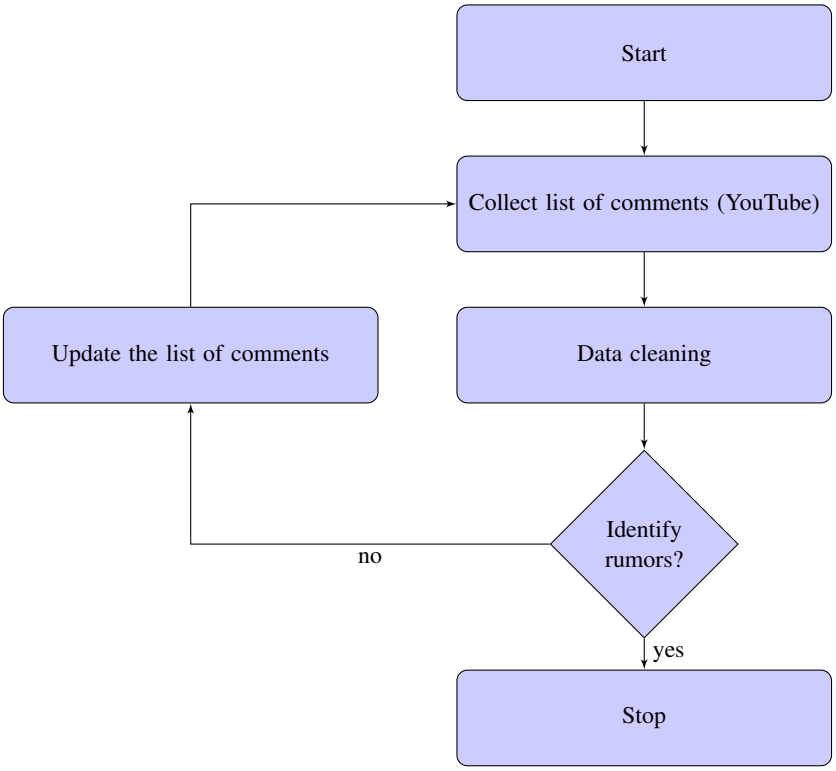


Fig. 1. Overview of the approach of collecting the the rumors dataset

3.1 Data collection

To build the corpus, we harvested the data using the YouTube API which allows to search for all the videos that match certain criteria and retrieve all the related comments. In order to increase our chance to get data in which we get fake news, we selected the topic of *Personalities death*. In fact, a lot of rumors in Internet concern the death of singers, actors, presidents, etc. That is why, in this work we selected three famous people in the Arab world who are mostly concerned by rumors: Fifi Abdu (an Egyptian dancer), Abdelaziz Bouteflika (the former Algerian president) and Adel Imam (an Egyptian comedian). Obviously, retrieving comments from YouTube by using Hashtags related to these three personalities will capture comments corresponding to rumors and no rumors. Therefore, when these data were collected, we used a set of relevant keywords concerning rumors (Fifi died, Allah yarhemak,True news, Algerian president dies, Bouteflika death, yes death, adel imam dies, Allah yerhamo, Adel die). If any comment contains one of these keywords, it will be considered as a rumor comment and it is saved in a rumor dataset, otherwise, it will be saved in the no-rumor subset. Table 1 shows some statistics of the harvested data, where $|C|$ indicates the number of comments for each topic.

Table 1. The collected stories related to Fifi abdo, Bouteflika and Adel Imam.

Topics	$\|c\|$
Fifi Abdo	2,363
Bouteflika	1,216
Adel Imam	500

3.2 Data Cleaning

In order to have a relevant analysis and develop a robust classifier, we first need to clean the data. Data cleaning is an important step in major NLP tasks to improve the quality of text data and ensure the reliability of the statistical analysis. Our cleaning step aims to filter the rumors and extract the useful terms. To this end, we removed from the collected data the special characters such as: $\{*, @, \%, \&...\}$. We also removed URL links, words in foreign languages, duplicated comments, etc. Table 3.2 gives the updated statistics about the collected corpus. It shows that the total size of the dataset has been reduced by around 20% after the cleaning process.

4 Data Analysis

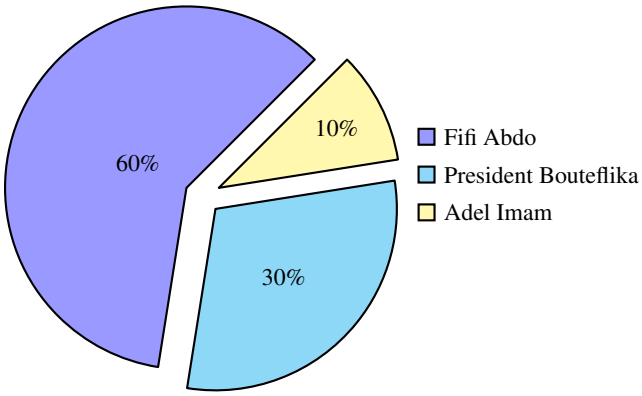
In this section, we will analyze our dataset in order to retrieve some knowledge about the transmission of rumors for the three studied topics. In Figure 2, we give the distribution of the vocabulary of this dataset in accordance to the three topics.

We remark that, with this approach of collecting data related to rumors, we harvested more data concerning the death of Fifi Abdo than for the two others even, if the

Table 2. The collected stories related to Fifi abdo, Bouteflika and Adel Imam after the data cleaning step

Topics	$\|c\|$
Fifi Abdo	2,145
Bouteflika	964
Adel Imam	326

Fig. 2. Distribution of the vocabulary



second personality was the President of a country. This is probably due to the fact that the dancer *Fifi Abdo* interests more people than the President Bouteflika and more than the famous actor *Adel Imam*.
Table 3 and Table 4 show that the Internet users posted more comments on the topic of *Fifi Abdo* whether rumors or no rumors which confirms that people are more interested by this personality than the two others. In Table 4, we remark that the number of comments about *Fifi Abdo* which are not supposed to be rumors are twice as much as for the topic *President Bouteflika*.

Table 3. Statistics corresponding to the rumors dataset.

Topic	$\|C\|$	$\|W\|$
Fifi Abdo	187	1605
President Bouteflika	106	3507
Adel Imam	50	508

Where $|c|$ represents the number of comments and $|W|$ the number of words.
In Figure 3, we give the distribution of the rumors through the period of the data collection. Even if this corpus is small, we can mention that a rumor can subsist for several years such as for the one concerning *Fifi Abdo* or several months such as those concerning *Bouteflika* or *Adel Imam*. It would be interesting in a future work, to find the

correlation between the spreading of the rumor and external events that induce the rumor.

Fig. 3. Distribution of the rumors

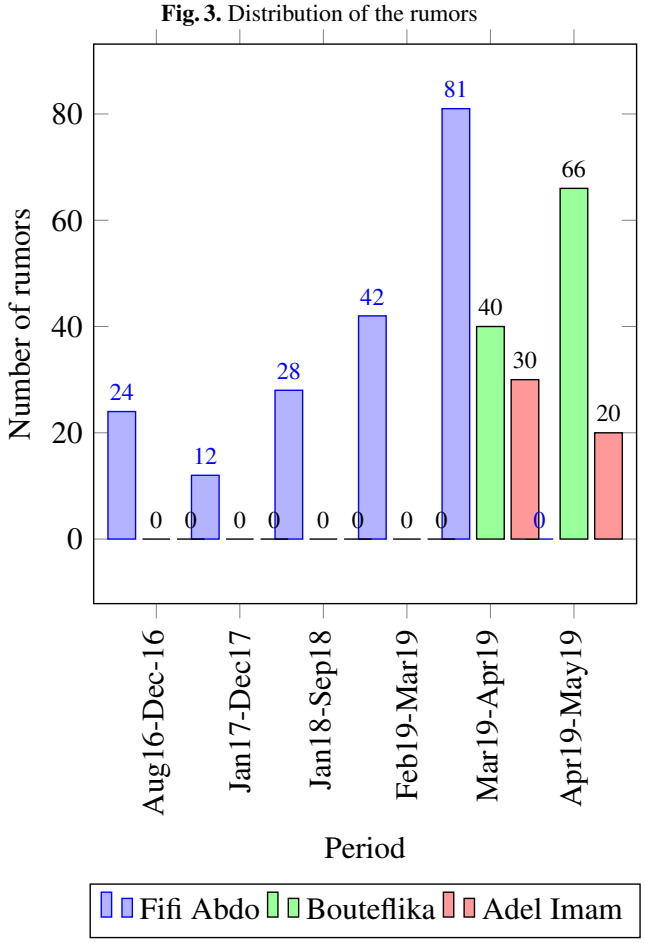


Table 4. Statistics corresponding to the No rumors dataset

Topic	$\|C\|$	$\|W\|$
Fifi Abdo	1958	22708
President Bouteflika	858	11917
Adel Imam	276	3085

We analyze the collected data to learn and understand what characterizes the messages conveying rumors. In Table 5, we give some samples from the corpus we collected automatically. For each rumor or no rumor sentence, we give its translation. These samples show clearly that the collected data concern rumors. In Table 6, we listed the most significant words corresponding to the dataset of rumors. The most used word is الله (*God*). As these are rumors about death, Muslims come back to God and beg forgiveness for the deceased. This explains the existence of this word in a significant way. The words related to the death are in the top list (مات, ماتت). As these rumors become truths for certain people, some people believe in them and even ask God to be merciful with the dead, which explains why we found the terms: يرحمو, يرحمها, يرحمك.

Table 5. Some examples of the collected data

Topic	**Examples**	
	Rumor	**Non Rumor**
Fifi Abdo	فيفي عبدو ماتت (Fifi abdo is died)	خبر كاذب (Fake news)
	الله يرحمك يا فوفو (May God have mercy on you Fufu)	فيفي عبدو عايشة زي القردة (Fifi is alive as a monkey)
Bouteflika	بوتفليقة مات في 2015 وصرحوا بها في سويسرا (Bouteflika died since 2015 they announced it in Switzerland)	كذابين بوتفليقة ماماتش ناس تزرع في الفتنة Bouteflika is not dead people want to spread sedition)
	الله يرحم رئسنا (May God have mercy on our President)	كذابة راهو حي (You are a liar he is alive)
Adel Imam	الله يرحمو مات عادل (May God have mercy on him Adel died)	عادل امام عايش كذاب انت (Adel Imam is alive you are a liar)
	صحيح مات بعملية بواسر (True he died after an operation of hemorrhoids)	الفنان عادل امام مامتش دول بيعملوا كذا عاشان يلمو ليكات (The artist Adel Imam is not dead they do this to get "like(s)")

Table 6. Most frequent words in the rumor dataset

Word	الله	مات	ماتت	يرحمو	بوتفليقة	ربي	يرحمها	فيفي	يرحمك	ربنا	عادل
Trans	God	Died	she died	have mercy	Bouteflika	My God	mercy on her	Fifi	mercy on your soul	our God	Adel
Count	201	75	64	63	48	42	31	24	20	19	9

If we analyze the corpus of no rumors (see Table 7), we also find the reference to the word God. Consequently, in the Arab world, this word could not be discriminating to identify rumor texts. However, we find numerous proper names corresponding to the studied topics: فيفي, عبدو and بوتفليقة. Negation terms alone or agglutinated to verbs are present in the top list of words. They invalidate an event, namely the death of the personalities, this is the case of the words: مش and مامات. We also found the antonym of the word *death*: حي, which indicates that the person is alive. The corpus also contains the word *lie* that indicates that the message or the event we talked about is fake. The above mentioned words seem to be discriminating for these topics.

Table 7. Most frequent words in the non-rumor dataset

Word	الله	ربنا	فيفي	عبدو	بوتفليقة	مش	يشفيها	حي	مامات	عادل	كذب
Trans	God	Our God	Fifi	Abdo	Bouteflika	not	heals her	alive	he is not died	Adel	lie
Count	914	305	207	118	114	108	89	89	64	31	34

5 Classification

In order to test the possibility of discrimination between rumor and no rumor comments, three data classification methods have been conducted in this work: Decision Tree (DT), Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM).

The Decision Tree classifier is a supervised machine learning technique where the data is recursively split according to the different attributes of the dataset. The leaves constitute the decisions and the nodes correspond to the area where data are split [10].
The principle of SVM [16] consists in looking for the optimal linear separating hyperplane that separates the data of one class from the other. SVMs aim to define the optimal boundary separating classes in feature space. The best hyperplane is the one that maximizes the distance between classes. The classification of new data is based on which side of the boundary the data is placed. In our case, we picked out a linear kernel for the separation.
Naïve Bayes classifiers are widely used in different applications in natural language processing and particularly in text classification[7][12][14] due to their efficiency and their acceptable predictive performance. MNB estimates the conditional probability of a particular term given a class as the relative frequency of the term t in all documents belonging to the class C. To train the MNB classifier, we used 1-gram, 2-gram and 3-gram of words as features supported by a TFIDF vector scores [13].
We performed few experiments and evaluated the classifiers with the most widely used measures in Information retrieval namely, Recall, Precision and Accuracy. Their corre-

sponding formulas are recalled respectively in 1,2 and 3:

$$Recall = \frac{tp}{tp+fn} \tag{1}$$

$$Precision = \frac{tp}{tp+fp} \tag{2}$$

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \tag{3}$$

where tp, tn, fp and fn are True Positive [5], True Negative[6], False Positive [7] and False Negative [8] respectively.

In Table 8, we reported the results of the three classifiers. The training was done on 70% of the data and the test on the remaining subset of the corpus. We conducted two kinds of tests. The first one has been done on each rumor topic and the second one, on the mixture of all the topic rumors. We observed that the achieved performance

Table 8. Performance on detecting rumors

Topic	SVM			D. Tree			MNB		
	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
Fifi Abdo	**95.35**	**87.72**	82.16	93.59	79.94	**82.8**	92.63	78.01	73.42
Bouteflika	94.2	92.69	78.18	**95.56**	**94.09**	**83.9**	93.86	90.7	77.99
Adel Imam	**93.68**	**85.2**	78.82	89.47	73.15	**80.88**	90.53	74.87	72.65
Combi	**95.35**	**92.77**	83.12	93.47	84.07	**83.56**	92.38	82.76	76.94

varies depending on the rumor topic and the used classifier. The best accuracy and the best Precision for Fifi Abdo are obtained by the SVM classifier while the best recall is achieved by the decision tree. For the rumors concerning the president Bouteflika, the best results whatever the measure are produced by the decision Tree. For the third rumor topic, the best accuracy and the best precision are achieved by the SVM, while the best recall is the one of the decision tree. When all the rumor topics are mixed, the best results in terms of accuracy and precision are obtained by the SVM and the best recall is achieved by the decision tree. Overall, for this dataset, the best classifier is the SVM one, while the MNB has not succeeded to outperform the other classifiers, in spite of its effectiveness in other classification applications, for any of the topics. This is probably due the fact that the MNB requires the use of more detailed features.

[5] case was positive and predicted positive
[6] case was negative and predicted negative
[7] case was negative but predicted positive
[8] case was positive but predicted negative

6 Conclusion

In this paper, we introduced a new Arabic corpus of fake news that we will make publicly available for research purposes. We detailed the collection process and gave important details about the harvested data on the subject of the death of three Arab celebrities. An exploratory analysis was carried out on the collected fake news to learn some features which characterize the messages conveying rumors such as, the frequent use of certain words. The classification task was performed using three classification methods namely Support Vector Machine (SVM) Decision Tree (DT) and Multinomial Naïve Bayes (MNB) to test the possibility of discrimination between rumor and no rumor comments. We witnessed that the achieved performance varies depending on the rumor topic and the used classifier. In the future, we look forward investigating the performance of other classification methods and also envisage enlarging our corpus by collecting more examples in various topics and performing a deep analysis on the data. One of our objective is to tackle the issue of detecting the rumors or the source of the rumors as soon as they arise.

Bibliography

[1] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web. pp. 675–684. ACM (2011)

[2] Chomsky, N., Herman, E.: A propaganda model. Manufacturing Consent: the Political Economy of the Mass Media. 2d ed. New York: Pantheon Books pp. 1–35 (2002)

[3] Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J.: Rumor cascades. In: ICWSM (2014)

[4] Galitsky, B.: Detecting rumor and disinformation by web mining. In: 2015 AAAI Spring Symposium Series (2015)

[5] Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Life in the network: the coming age of computational social science. Science (New York, NY) **323**(5915), 721 (2009)

[6] Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.F.: Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1751–1754. ACM (2015)

[7] McCallum, A., Nigam, K., et al.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. vol. 752, pp. 41–48. Citeseer (1998)

[8] Procter, R., Crump, J., Karstedt, S., Voss, A., Cantijoch, M.: Reading the riots: What were the police doing on twitter? Policing and society **23**(4), 413–436 (2013)

[9] Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1589–1599. Association for Computational Linguistics (2011)

[10] Quinlan, J.R.: Learning decision tree classifiers. ACM Comput. Surv. **28**(1), 71–72 (Mar 1996)

[11] Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Detecting and tracking the spread of astroturf memes in microblog streams. arXiv preprint arXiv:1011.3768 (2010)

[12] Rish, I., et al.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol. 3, pp. 41–46 (2001)

[13] Spärck Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation **28**, 11–21 (1972)

[14] Su, J., Shirab, J.S., Matwin, S.: Large scale text classification using semi-supervised multinomial naive bayes. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 97–104. Citeseer (2011)

[15] Tolosi, L., Tagarev, A., Georgiev, G.: An analysis of event-agnostic features for rumour classification in twitter. In: Tenth International AAAI Conference on Web and Social Media (2016)
[16] Vapnik, V.: Statistical learning theory. Wiley (1998)
[17] Zubiaga, A., Liakata, M., Procter, R.: Learning reporting dynamics during breaking news for rumour detection in social media. arXiv preprint arXiv:1610.07363 (2016)
[18] Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one **11**(3), e0150989 (2016)