Supervised Learning Approach for Twitter Credibility Detection

Noha Y. Hassan
Computer Science Dept.
Beni-Suef University
Beni-Suef, Egypt
noha.yehia@fcis.bsu.edu.eg

Wael H. Gomaa
Computer Science Dept.
Beni-Suef University
Beni-Suef, Egypt
wael.goma@fcis.bsu.edu.eg

Ghada A. Khoriba
Computer Science Dept.
Helwan University
Cairo, Egypt
ghada_khoriba@fci.helwan.edu.eg

Mohammed H. Haggag
Computer Science Dept.
Helwan University
Cairo, Egypt
mohamed.haggag@fci.helwan.edu.eg

***Abstract* - Twitter is the most popular micro-blogging medium that allows users to exchange short messages, provides a platform for public people to share the news. Nowadays, Twitter counts with an average of 328 million monthly active users and is growing rapidly. Detecting the credibility of shared information on Twitter becomes a necessity, especially during high impact events. In this paper a classification model based on supervised machine learning techniques is proposed to detect credibility. The proposed model uses an extensive set of features including both content-based and source-based features. The research compares the performance of five different machine learning classifiers using three feature sets: content based, source based and a combination of both sets. The best performance is achieved when using a combined set of features and applying Random Forests as a classifier with accuracy 78.4%, precision 79.6%, recall 91.6% and f1-measure 85.2%. Experiments also revealed that the proposed model achieves improvement of 22% when compared to CRF which applies the same approach in terms of F1-measure. Feature analysis is presented to highlight the importance of the source-based features compared with the content-based features as deciders for credibility.**

Keywords—Twitter ; credibility; machine learning; content-based; source-based.

I. INTRODUCTION

Micro-blogging mediums such as Facebook and Twitter are used for sharing news, opinions and experiences among people all over the world. They are growing very fast in popularity and are now replacing traditional media as a source for obtaining news and information [1]. Twitter allows users to post and exchange short messages or "tweets". Tweets are shared with the author's followers and can be easily disseminated through "re-tweet". Recently, Twitter has been considered as the most micro-blogging platform used as news source [2,3]. News on Twitter comes from different sources most of them from public users. The absence of supervision makes Twitter a suitable environment for spreading rumors. Many researches revealed that a lot of content on Twitter may be incredible especially in high impact events [4-6]. Gupta et al. [7] studied the spreading of rumors on Twitter during Hurricane Sandy and discovered that about 86% of the fake tweets were re-tweets. Therefore, detecting credible or trustworthy information in Twitter becomes a necessity as more people depend on social media to obtain news.

In fact, it is hard to determine the credible tweets manually. Recently, several approaches have been proposed to handle this challenge. In this paper, a model that automatically classify tweets as credible or non-credible is proposed. The proposed model is based on a set of features, particularly 32 features, including both content-based features and source-based features. To train and test the model, a dataset of 5802 annotated tweets collected during five high impact events was used. The performance of five different supervised classifiers: Random Forests (RF), Support Vector Machines (SVM), Logistic Regression (LR), Naïve Bayes (NB) and K-Nearest Neighbor (KNN) was compared. The proposed model achieved an accuracy rate of 78.4% in predicting the credibility of tweet messages using Random Forests classifier. The research also presents a feature analysis to identify the most prominent features based on our outcomes.

This paper is organized as follows: Section 2 discusses related works in credibility assessment using different techniques. Section 3 includes the proposed model then the results of our findings are presented in section 4. Conclusions and future work are presented in section 5.

II. RELATED WORK

Most of the research aiming at determining the credibility of Twitter messages are classification-based approaches. These approaches classify tweets to credible and non-credible using supervised machine learning techniques [8-14]. A ground truth that contains a collection of annotated tweets with the features related to them is used to build automatic classifiers that can accurately determine the credibility of a given tweet. The accuracy of the annotation process is an important factor affecting the efficiency of the prediction [15]. Another key factor is the relevance of the extracted features. Some research considers the content of the tweet itself [16] while others focus on the author as the source of the tweet [17]. In this section, some of the research that is most related to this area is reviewed.

Castillo et al. were the first to work on the Twitter credibility problem in a structured way [8,9]. The research focused on tweets related to trending topics and developed a supervised machine learning model to predict their credibility. They defined different types of features, some of them are related to the content of the tweet while others focus on the author of the tweet or were aggregated from the related topic. The labeling process of the dataset included two rounds: the first-round separates posts which contain information about news events (labeled as NEWS), from personal opinions (labeled as CHAT). The second round focuses on the tweets labeled as NEWS and classify them into credible/non-credible. The extracted features and the annotated data are used to train many classifiers such as SVM, decision trees, decision rules, and Bayesian networks on the annotated data, but best results were achieved by J48

decision tree. Gupta et al. [10] concluded that measuring the credibility of Twitter messages can be automated accurately based on Twitter features. The authors used supervised machine learning and relevance feedback approach to rank tweets into seven levels of credibility. The research identified some prominent features based on the content and source-based features. Number of followers, number of unique characters and swear words were the most effective features. To evaluate their model, the dataset was collected using the Streaming API related to fourteen high impact events. Experiments indicated that about 30% of tweets related to an event include information about the event while 14% was spam and only 17% include credible information about the event.

Another research that aimed to assess the credibility of individual tweets is the one by Zubiaga et al. [11]. The authors developed a credibility detection system that enables flagging and warns users of unverified posts. The dataset consists of 5802 tweets collected using Twitter streaming API during five breaking news stories. Regarding features extraction process, their feature set contains: word vectors, word count, use of question mark and capital ratio, listed count, follow to friend ratio and age of the user. The authors used the Conditional Random Fields (CRF) as a sequential classifier and compared its performance with three more classifiers: Naive Bayes (NB), Support Vector Machines (SVM), and Random Forests (RF). The experimental results showed that CRF outperformed the other classifiers achieving approximately 61% F1-measure.

III. PROPOSED MODEL

In this section, the proposed model for tweets credibility detection is illustrated. As shown in figure 1, the model consists of four modules : 1) Feature extraction, 2) Feature scaling, 3) Training and 4) Testing and evaluation. The components of the model are explained in detail in the following sections.

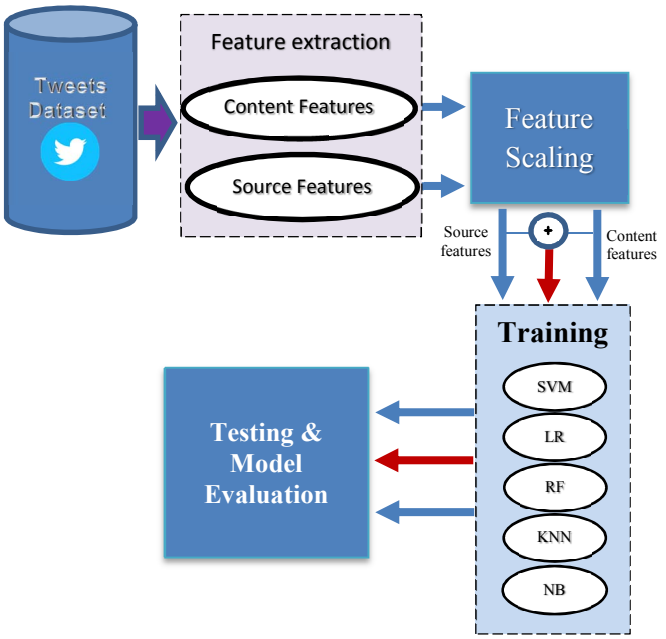

Fig 1: The proposed model architecture

A. Feature Extraction

Feature extraction is the process of obtaining the most relevant information that can distinguish between classes to use them in building the learning model. There are several useful features introduced by the previous research in credibility assessment [12,14,18]. Most of these studies rely on Twitter features and categorize them into:

- **Content-based features:** features that focus on the content of the message such as the length of the tweet, retweet count, the presence of hashtags or user mentions.
- **Source-based features:** consider characteristics of the author as the source of the tweet such as number of followers and if the user has description.

To predict the credibility of a tweet, the content features extracted from the tweet itself must be considered as an important factor. Also, the source features are indication of the author's experience and reputation. Table 1 listed our feature set which contains 32 features comprised of 17 content features and 15 source features. Some of the extracted features are computed like followers/friends ratio which indicates the popularity of the author. Other features are extracted from the author's previous tweet posts such as average number of URLs and retweet fraction.

Table 1. Selected content and source features for credibility assessment.

Content-based features	Source-based features
Retweet count	Followers count
Length of the tweet in characters	Friends count
Number of words	Listed count
The tweet has URL?	Has description?
The tweet has user mentions?	Length of description
The tweet has hashtags?	Length of screen name
The tweet has Question mark?	User has URL?
The tweet has exclamation?	Is verified account?
The tweet has special characters?	Has default profile picture?
The tweet has emoticons?	Followers/friends ratio
URL count	Average number of hashtags
User mentions count	Average number of URLs
Hashtags count	Average number of mentions
Question mark count	Average tweet length
Exclamation count	Retweet fraction
Special characters count	
Emoticons count	

B. Feature scaling

Our feature set contains features highly varying in range such as followers count, followers/friends ratio and retweet count. Table 2 presents the range (maximum value – minimum value) of some of the selected features. Some of the machine learning depends on calculating distance between data points. The features with high range will weight in a lot more in distance calculations than features with low range. After the feature extracting process, features must be normalized or rescaled to standardize the range of features before using the classifiers. For this task, min-max scaling is used to rescale the range of features in [0, 1].

The general formula is given as:

$$X_{scaled} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where: X_{scaled} is the normalized value and X is the original value.

Table 2 : Range value of some extracted features

Feature	Range
Retweet count	99499
Tweet length	135
url count	2
Hashtag count	8
Followers count	25303073
Listed count	2275623
Description length	160
Followers/friends ratio	4712390
Mean URLs	0.75
Mean Hashtags	1.63
Mean RT	45223.14

C. Training

The annotated tweets and the extracted features are used to train a set of machine learning classifiers namely, Random Forests, Naïve Bayes, SVM, Logistic Regression and KNN. The expirements revealed that Random Forests outperformed the rest of classifiers.

Random Forests algorithm is a supervised machine learning algorithm used in many researches to solve the credibility problem [13,14]. The classifier is an ensemble algorithm which builds multiple decision trees and combines them together to produce a more effective classifier. The decision tree algorithm uses the entropy and information gain to select the most discriminant feature and use it for branching. One of the decision trees disadvantages is overfitting specially when the tree is deep, RF classifier can limit overfitting if there are enough trees [19]. Figure 2 illustrates the algorithm.

Algorithm 1: Random forests classifier

1. *Select randomly M features from the feature set.*
2. *For each x in M*
   a. *calculate the Information Gain*

$$Gain(t,x) = E(t) - E(t,x)$$
$$E(t) = \sum_{i=1}^{c} - P_i \log_2 P_i$$
$$E(t,x) = \sum_{c \in X} P(c)E(c)$$

   Where E(t) is the entropy of the two classes, E(t,x) is the entropy of feature x.
   b. *select the node d which has the highest information gain*
   c. *split the node into sub-nodes*
   d. *repeat steps a, b and c to construct the tree until reach minimum number of samples required to split*
3. *Repeat steps 1 and 2 for N times to build forest of N trees*

Fig. 2. Random Forests classifier algorithm

To predict the credibility of any given tweet, the *N* decision trees are used to predict the class label of the tweet then majority vote is done to decide on the label.

The implementation of Random Forests in scikitlearn python library[1] is used setting the number of generated trees to 500. The algorithm can be used to measure the feature importance by indicating how much the tree nodes maximize the information gain across all trees. Figure 3 shows the feature importance matrix for the selected features. It is clear that source-based features such as listed count and followers count are more discriminant than content-based features. The aggregated features, from the user's history, such as retweet average (Mean_RT), URLs average (Mean_URLs) and hashtag average (Mean_hashtag) were proved to be important features and have high impact on the efficiency of the prediction.

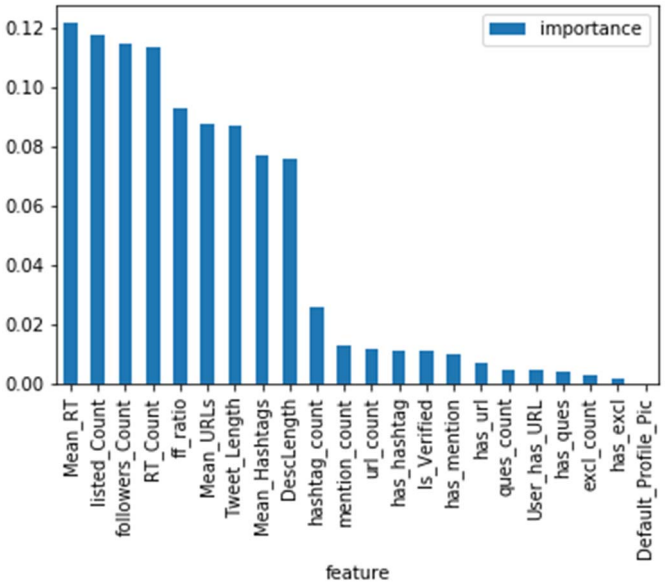

Fig 3: The feature importance.

D. Dataset

In fact, few public datasets are available. The dataset collected in the works by Zubiaga et al. [11] is primarily used in the experiments. The dataset was collected using Twitter streaming API during five breaking news events that could highly prompt the initiation and propagation of rumors. The events were widely reported in Twitter at the time of occurrence namely: Charlie Hebdo, Sydney Siege, Ottawa Shooting, Germanwings-Crash and Ferguson Shooting. The collected tweets were sampled by picking tweets that have a high number of retweets. The annotation of the tweets was performed and reviewed by a team of journalists. The final dataset consists of 5802 tweets was manually annotated to 3830 (66%) credible and 1792 (34%) non-credible tweets.

E. Feature Analysis

A simple statistical analysis is conducted to identify the most prominent features based on our outcomes. As described earlier, the experiments revealed that the source-based features are more important in the credibility task than content-based features. Four source-based features are found to be highly effective in credibility prediction, namely followers count, listed count, average of retweets and average of URLs.

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

Approximately 76% of the tweets whose authors has followers count less than 10,000 are credible while 24% are non-credible. Moreover, tweets with large numbers in followers count and listed count are likely to be non-credible. Average of retweet and average of URLs are features extracted from the author's history. In terms of the retweet average, about 75% of the tweets whose author has more than 300 are classified as credible tweets. Moreover, 80% of the tweets whose author has URLs average less than 0.15 are classified as credible tweets. Regarding the content-based features, the importance of the URL inclusion feature is captured. About 77% of the tweets which do not include URL are credible tweets while only 52% of the tweets which include URL are credible tweets.

Sentiment features as the tweet sentiment, whether positive, negative or neutral have been proved to be an indicator of credibility [8,14]. It is important to find the correlation between the sentiment and the credibility of the tweet. The results shown in table 3 indicates that there is not significant difference between credible and not credible tweets. NLTK sentiment analyzer[2] was used for the classification of tweets.

Table 3. sentiment analysis for credible and non-credible tweets.

Credibility	Positive	Neutral	Negative
Credible	35.5%	1.1%	63.4%
Non-credible	34.0%	1.0%	65.0%

F. Model Evaluation

To evaluate the proposed model, three experiments were conducted: training the classifiers using content feature set only, training the classifiers using source features set only and in the third experiment both content and source feature sets were used. The experiments included training five different classifiers namely, Random Forests, Naïve Bayes, SVM, Logistic Regression and KNN. A 10-fold cross validation was applied on the entire dataset. Different performance measurements are used to evaluate the results:

$$Accuracy = (TP+TN)/(TP+FP+TN+FN) \quad (2)$$

$$Precision = TP/(TP+FP) \quad (3)$$

$$Recall = TP/(TP+FN) \quad (4)$$

$$F1\text{-}measure = 2 * (Precision*Recall)/(Precision+Recall) \quad (5)$$

Where:
TP is the number of tweets correctly identified as credible, FP is the number of tweets incorrectly identified as credible, TN is the number of tweets correctly identified as non-credible and FN is the number of tweets incorrectly identified as non-credible.

IV. Experiments and Results

In this section, we study and compare the performance of the proposed model when training the classifiers using different feature sets. The target is to determine whether content features only or source features only can be good indicators for credibility. The five classifiers were trained using content-based features only, then the experiment was repeated using source-based features only. Tables 4 and 5 present the accuracy, Precision, Recall and F1-measure of the two experiments. As shown in table 4, Logistic Regression achieved the best accuracy rate 67.1%, Random Forests achieves the best precision 69% while the best of both recall and F1-measure are achieved by Naïve Bayes 99% and 79.2% respectively when using content-based features only.

Table 4. The performance of the proposed model using content-based features

Classifier	Accuracy	Precision	Recall	F1-measure
Random Forests	0.616	**0.690**	0.762	0.724
KNN	0.621	0.684	0.792	0.734
SVM	0.665	0.679	0.936	0.787
Logistic Regression	**0.671**	0.679	0.931	0.785
Naïve Bayes	0.660	0.661	**0.99**	**0.792**

As shown in table 5, an accuracy of 77.8% was achieved by Random Forests classifier which is 16% more than the accuracy of 61.6% achieved by the same classifier when content-based features were used alone. Using the source-based features only, Random Forests outperformed the rest of classifiers in terms of accuracy, precision and F1-measure achieving 77.8%, 79.5% and 83.8% respectively while Naïve Bayes achieves the best recall 99%.

Table 5. The performance of the proposed model using source-based features

Classifier	Accuracy	Precision	Recall	F1-measure
Random Forests	**0.778**	**0.795**	0.887	**0.838**
KNN	0.709	0.757	0.826	0.789
SVM	0.660	0.669	0.964	0.789
Logistic Regression	0.661	0.672	0.951	0.787
Naïve Bayes	0.664	0.666	**0.990**	0.796

The experimental results indicated that the source-based features are more powerful than the content-based features as deciders for credibility. This observation is proven to be true when compared with the features' importance that was measured by Random Forests classifier. The important matrix, shown in figure 3, indicates that most of the important features are source based-features such as listed count and followers count. The most important content features are retweet count and length of the tweet. However, the classifier achieved 78.4% accuracy,

Table 6. The performance of the proposed model using both content-based and source-based features

Classifier	Accuracy	Precision	Recall	F1-measure
Random Forests	**0.784**	**0.796**	0.916	**0.852**
KNN	0.662	0.725	0.789	0.755
SVM	0.670	0.687	0.919	0.786
Logistic Regression	0.669	0.688	0.912	0.784
Naïve Bayes	0.667	0.677	**0.948**	0.790

[2]https://www.nltk.org/api/nltk.sentiment.html

79.6% precision, 91.6% recall and 85.2% F1-measure when the two feature sets are combined as shown in table 6. As a result, we can rely on source-based features to decide on tweet credibility but combining the two feature sets improves the performance of the learning model.

The performance of the five classifiers in terms of accuracy in the three experiments is compared and presented in figure 4. Random Forests achieves higher accuracy rates when using both source and combined feature sets while Logistic Regression is the classifier that best exploits content-based features.

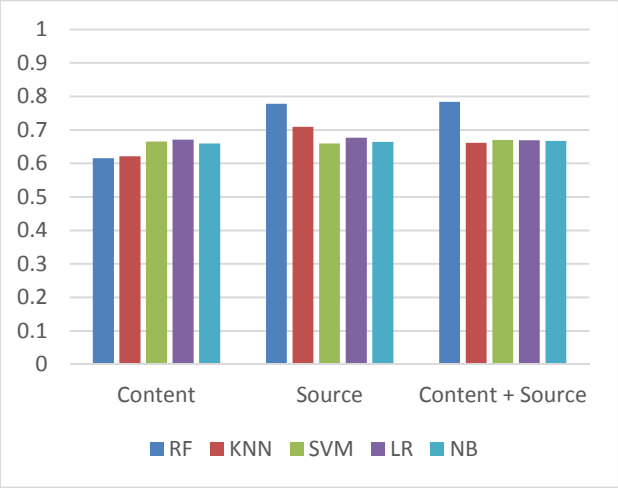

Fig 4: Comparison between the accuracy rates of the used classifiers with different feature sets.

Regarding the effect of the feature scaling process, experimental results revealed that feature scaling has a big impact on the quality of some classifiers while it has a small or no impact on others. Table 7 shows the accuracy results of all classifiers using content features before and after scaling. Classifiers that exploit distances between samples are more sensitive to feature ranges such as KNN and SVM. KNN recorded the largest improvement in accuracy rate, approximately 19% since it just looks at the Euclidean distance between samples.

Table 7: Accuracy results of the feature-based model using content features before and after scaling.

Classifier	Accuracy before scaling	Accuracy after scaling
Random Forests	0.610	0.609
KNN	0.435	0.620
SVM	0.530	0.659
Logistic Regression	0.664	0.670
Naïve Bayes	0.652	0.660

Moreover, the proposed model was compared with a similar approach existing in the literature. The approach introduced by Zubiaga et al. [11] relies in its classification on content and source-based features. They focused on the textual features that can be extracted from the tweet such as word vectors and Part of speech tags. Their model relies on applying CRF Conditional Random Fields as a classifier. Five-fold cross validation was applied on the same dataset used by Zubiaga et al.[11] in order to achieve fair comparison. Table 8 depicts the Precision, Recall, and the F1-measure of both CRF and the proposed model using the same dataset. Our intuition is that the proposed model outperforms CRF because the inclusion of features extracted from user's timeline which is proved to be high discriminant features. The comparison in table 8 shows an improvement of 18% in terms of percentage in F1-measure over CRF when using content-based features while the improvement is 49% when using source-based features. The proposed model outperformed CRF by 11%, 33%, 22% in terms of percentage increase in precision, recall and F1-measure respectively when using both sets.

Table 8. Comparison between the proposed model and Zubiaga et al [11].

Content features			
	Precision	**Recall**	**F1-measure**
CRF	0.683	0.545	0.606
Proposed model	**0.679**	**0.929**	**0.785**
Source features			
	Precision	**Recall**	**F1-measure**
CRF	0.462	0.268	0.339
Proposed model	**0.790**	**0.872**	**0.829**
Combined features			
	Precision	**Recall**	**F1-measure**
CRF	0.667	0.556	0.607
Proposed model	**0.778**	**0.890**	**0.831**

V. CONCLUSION

In this paper, a credibility detection model based on machine learning techniques was introduced making use of a large set of features. Some of the features are source based and some are content based that are extracted from the text of the tweet. The selected features include some features extracted from the author's history. To test the performance of the proposed model, a dataset contains 5802 English tweets was used to train a set of classifiers and compare the performance of them. The Random Forests classifier achieved the best results and outperformed Zubiaga et al. approach in terms of precision, recall and F1-measure. For future work, the textual features can be studied to examine their impact on credibility prediction.

REFERENCES

[1] Stocker, Alexander, Alexander Richter, and Kai Riemer. "A Review of Microblogging in the Enterprise." it-Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik 54.5 (2012): 205-211.

[2] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[3] Cuesta, Álvaro, David F. Barrero, and María D. R-Moreno. "A Descriptive Analysis of Twitter Activity in Spanish around Boston Terror Attacks." International Conference on Computational Collective Intelligence. Springer, Berlin, Heidelberg, 2013.

[4] Gupta, Aditi, Hemank Lamba, and Ponnurangam Kumaraguru. "$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter." eCrime Researchers Summit (eCRS), 2013. IEEE, 2013.

[5] Lu, Xin, and Christa Brelsford. "Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami." Scientific reports 4 (2014): 6773..

[6] Ashktorab, Zahra, et al. "Tweedr: Mining twitter to inform disaster response." ISCRAM. 2014.

[7] Gupta, Aditi, et al. "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy." Proceedings of the 22nd international conference on World Wide Web. ACM, 2013.

[8] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[9] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Predicting information credibility in time-sensitive social media." Internet Research 23.5 (2013): 560-588.

[10] Gupta, Aditi, and Ponnurangam Kumaraguru. "Credibility ranking of tweets during high impact events." Proceedings of the 1st workshop on privacy and security in online social media. ACM, 2012.

[11] Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. "Learning reporting dynamics during breaking news for rumour detection in social media." arXiv preprint arXiv:1610.07363 (2016).

[12] Gupta, Aditi, and Ponnurangam Kumaraguru. "Credibility ranking of tweets during high impact events." Proceedings of the 1st workshop on privacy and security in online social media. ACM, 2012.

[13] Lorek, Krzysztof, et al. "Automated credibility assessment on Twitter." Computer Science 16.2) (2015): 157-168.

[14] El Ballouli, Rim, et al. "CAT: Credibility Analysis of Arabic Content on Twitter." WANLP 2017 (co-located with EACL 2017) (2017): 62.

[15] Madlberger, Lisa, and Amai Almansour. "Predictions based on Twitter—A critical view on the research process." Data and Software Engineering (ICODSE), 2014 International Conference on. IEEE, 2014.

[16] Al-Khalifa, Hend S., and Rasha M. Al-Eidan. "An experimental system for measuring the credibility of news content in Twitter." International Journal of Web Information Systems 7.2 (2011): 130-151.

[17] Alrubaian, Majed, et al. "Reputation□based credibility analysis of Twitter social network users." Concurrency and Computation: Practice and Experience 29.7 (2017): e3873.

[18] Almansour, Amal. Credibility assessment for Arabic micro-blogs using noisy labels. Diss. King's College London, 2016.

[19] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.