CAT: Credibility Analysis of Arabic Content on Twitter

Rim El Ballouli,[1] **Wassim El-Hajj,**[2] **Ahmad Ghandour,**[2] **Shady Elbassuoni,**[2]
Hazem Hajj[3] **and Khaled Shaban**[4]
[1] Department of Computer Science, University Joseph Fourier - Grenoble 1
[2] Department of Computer Science, American University of Beirut
[3] Department of Electrical and Computer Engineering, American University of Beirut
[4] Department of Computer Science, Qatar University
rim.el-ballouli@univ-grenoble-alpes.fr
{we07,arg06,se58,hh63}@aub.edu.lb, khaled.shaban@qu.edu.qa

Abstract

Data generated on Twitter has become a rich source for various data mining tasks. Those data analysis tasks that are dependent on the tweet semantics, such as sentiment analysis, emotion mining, and rumor detection among others, suffer considerably if the tweet is not credible, not real, or spam. In this paper, we perform an extensive analysis on credibility of Arabic content on Twitter. We also build a classification model (CAT) to automatically predict the credibility of a given Arabic tweet. Of particular originality is the inclusion of features extracted directly or indirectly from the author's profile and timeline. To train and test CAT, we annotated for credibility a data set of $9,000$ Arabic tweets that are topic independent. CAT achieved consistent improvements in predicting the credibility of the tweets when compared to several baselines and when compared to the state-of-the-art approach with an improvement of 21% in weighted average F-measure. We also conducted experiments to highlight the importance of the user-based features as opposed to the content-based features. We conclude our work with a feature reduction experiment that highlights the best indicative features of credibility.

1 Introduction

The Web has become a treasured source of opinions, news and information about current events. Twitter, Facebook, Instagram, and others play a vital role in publishing such information. This immense data has become a vital and rich source for tasks such as popularity index, elections, opinion mining, pro/con classification, emotion recognition, rumor detection, etc.

With the large scale of data generated on these outlets, it is inevitable that the credibility of the generated information would highly vary. This would in turn influence the opinions of the readers and the accuracy of the tasks performed on such data. A recent study by (Allcott and Gentzkow, 2017) indicated that fake news published on social media during and before the American presidential elections in November 2016, did have an effect on voters, but was not the reason behind the victory of Trump. Others suggest otherwise and confirm that fake news made Trump president. Take for instance an interview with the Washington Post, when fake news writer and promoter Paul Horner said that "I think Trump is in the White House because of me", hinting that his fake news were believed by the voters and even adopted and shared [1].

In this paper, we focus on tweets, being a main source of news and opinions, and propose a model, called CAT, that best classifies tweets as credible or not. We adopt the Merriam Webster definition of credibility that states: credibility is the quality of being believed or accepted as true, real or honest. CAT uses a binary classifier that classifies a given tweet as either credible or not. CAT is built on top of an exhaustive set of features which includes both content based and user-based features. Content-based features are features extracted from the tweet itself, for instance, sentiment, language, and text cues, whereas user-based features are extracted from the tweet author, for instance, exper-

[1] https://goo.gl/3txvTd

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 62–71,
Valencia, Spain, April 3, 2017.

tise of the user generating the tweet, and the number of followers. In particular, we use 26 content-based features and 22 user-based features.

To train and test our classifier, we extracted over $9,000$ Arabic tweets and annotated them with the help of six well-paid human judges using a custom crowd-sourcing platform. The $9,000$ tweets were divided among the judges to obtain three annotations for each tweet. The judges annotated each tweet as either "credible", "non-credible" or "can't decide". To assist the annotators in accurately assessing the credibility of a given tweet, they were provided with useful cues such as the tweet itself and its author. While we based our experiments on Arabic content, our credibility model is general enough to predict the credibility of tweets in any language provided that the necessary resources for extracting some of the language dependent features such as sentiment are available.

Predicting the credibility of tweets has been previously studied to some extent. However, to the best of our knowledge, none of the previous work considered features from timeline or profile-picture face detection to assess the credibility of a given tweet. For example, (AlMansour, 2016) relies on some features including the presence of a profile picture to perform credibility assessment. However, in our approach, we do not only evaluate the presence of a profile picture, we also take this feature one step further by using Google cloud vision API to perform face detection and extract textual information that might be available in the picture. We compared CAT to several baselines and to a recent state of-the-art approach, namely, TweetCred (Gupta et al., 2014). CAT consistently surpassed the accuracy of the baseline approaches. It also outperformed TweetCred with an improvement of 16.7% in Weighted Average F-measure. While TweetCred relies in its classification on real-time features only, CAT utilizes the authors history for any clues that might be helpful in deciding on the credibility of the tweet.

Finally, most of the previous work on predicting the credibility of tweets have been based on annotated English tweets. In this paper, we propose a robust credibility classifier (CAT) that can work for tweets in any language and we test it on a relatively big data set of Arabic tweets. Our annotated data set of $9,000$ Arabic tweets is made public to act as a valuable resource for future research in this area. Another credibility data set exists, but it is smaller and topic dependent (Al Zaatari et al., 2016).

2 Related Work

We broadly classified research on credibility into the work done on Arabic content and that done on English content. Credibility of Arabic content has not received profound attention from researchers and as such, this area has a lot of room for improvement. For English content, some researchers tackled the problem of judging the credibility of tweets. Others tackled the problem of judging the credibility of tweet clusters, and others built classifiers to judge the credibility of tweet authors instead of tweets. We overview each line of research next.

Credibility of Arabic tweets: In (Sultan et al., 2010), the authors propose a model to identify credible Arabic news on Twitter. Their model relies heavily on the similarity of the tweet content with collected news from reputable sources. They collected both tweets and news articles on trendy topics. After text processing, they represented both the tweet and the articles as TF-IDF vectors. They relied on the cosine similarity measure between the tweet and the articles to determine the tweet's credibility. The model is able to predict credibility of previously discussed topics on the web. Yet, it fails to assign credibility values for tweets discussing breaking events.

Credibility of English Tweets: In (Gupta et al., 2014), the authors developed the first real time credibility analyzer through a semi supervised ranking model (TweetCred). They extracted a total of 45 features, all of which can be extracted in real time. Their feature set did not include features related to a group of tweets. Neither did it include user-based features that are dependent on the previous tweet posts of a user. Next, the feature vectors for all the annotated tweets were given as input to SVM-Ranking algorithm as training data set. They used the trained model as a backend for their system. When a new Twitter feed comes in real-time, the rank of the tweet is predicted using the learned model and displayed to the user on a scale of 1 (low credibility) to 7 (high credibility). TweetCred relies in its classification on real-time features only, such as, count of re-tweets, and count of friends. CAT, however, utilizes the tweeter's history for any clues that might be helpful in deciding on the credibility of the tweet.

In (Landis and Koch, 1977), the authors tackled the same problem from another perspective where they proposed an automated ranking scheme to present the user with a ranked output of tweets according to credibility. They used SVM ranking scheme to rank the results according to the perceived credibility of the information contained in the tweet.

Credibility of tweet authors: In (Canini et al., 2011), the authors designed an automatic tool to rank social network users based on their credibility and relevance to the query. They defined information credibility of a source by expertise and topical relevance of the source discussion topic. Expertise is measured by calculating the proportion of ones followers who are likely to be in the search results. Relevance is measured using LDA topic modeling. This work though fails when determining the expertise of the author, since some authors with high social network status may be non-credible in general, or non-credible when discussing certain topics that they are not experts in, or biased with respect to the topic being discussed. Add to that and as will be concluded in our work, relying solely on the author is not sufficient to decide on credibility.

Credibility of tweet clusters: In (Castillo et al., 2011), the authors built an automatic tool to assess the level of credibility of news topics. Their credibility classifier relies on topic-based features i.e. features extracted from a group of tweets and not from individual tweets. In turn, the classifier classifies a cluster of tweets or a topic as credible or non credible While, this model is useful for detecting rumor topics, it cannot detect non-credible tweets within a credible topic. Next we discuss our approach.

3 Methodology

The process of creating CAT and testing it passed through multiple steps that include: data set collection, data set annotation, feature engineering, sentiment extraction, experimental evaluation, and finally feature analysis. The following sections explain the details of every step.

3.1 Data Set Collection

The data collection started by querying twitter API while specifying two conditions: the tweet should be written in Arabic, and it should include a hashtag. We collected around 17 million tweets in a period of two weeks. The next step was to perform data cleaning on the 17 million tweets. The following data cleaning steps were performed: 1) all tweets composed *only* of religious quotations and versus were removed. To do this we used trigram matching with dictionaries we created to remove tweets that have words such as القرآن ، الله (*Alqr|n* - The holy book of Muslims, *Allh* - God) or words matching with Ahadeeth and Athkar found in المكتبة الشاملة - "Maktabah Alshamelah" [2]. 2) all tweets that are ads were removed, 3) all tweets that are composed of *only* emoticons or love/hate words were removed, and 4) all tweets that were sexual or composed of *only* badmouth words were removed. These data cleaning steps were mainly performed using regular expressions and also utilizing a self-created list of words and emoticons. We also removed all tweets that contain a hashtag without any text appended to it and all retweets to avoid duplication. Hence, a big portion of the collected tweets were retweets. After data cleaning, we grouped the tweets by the hashtags obtaining tweet clusters, each cluster containing related tweets. To ensure the topic independence of our data set, we randomly selected 10% of the tweets in every cluster and grouped them to obtain a data set of $9,000$ tweets that includes tweets addressing a wide variety of topics. Besides retrieving tweet text, metadata about the tweet and the tweet author were collected as well. We next describe the annotation process.

3.2 Annotation

To facilitate the annotation process we developed an in-house platform to collect annotations. While there exits other crowd-sourcing platforms such as Mechanical Turk and CrowdFlower, we relied on our own platform due to limitations imposed by existing platforms when dealing with Arabic data. For each tweet to be annotated, we provided the annotators with two URL links. The first link provided the annotator with the tweet text as displayed on Twitter. This option provided annotators with cues such as count of retweets, favorites that the tweet received, and the authors screen name. The second link provided the complete author profile as found on Twitter. The author profile is rich with cues that annotators can use to make their decisions. These cues include the follower count, previous tweet posts, author's profile image, and in some cases a brief description about

[2] http://shamela.ws/

the author. These two URLs (the tweet itself and the tweet's author profile) provided the annotators with rich information that can aid them in deciding whether a tweet is credible or not. Annotators were asked to either label a tweet as "credible" or "non-credible". They were also given the option to select "can't decide" when they felt confused or unsure. We also added the option of "deleted" since some authors delete their tweets after posting them or Twitter blocks the account in which case annotators will not be able to view the tweet. Each tweet received three annotations from three different annotators. Tweets that where labeled as "can't decide" by at least two annotators have been discarded. A majority vote was used to decide on the final labels of the tweets. In total, 60% of the $9,000$ tweets were annotated as credible and 40% were annotated as non credible.

To ensure good annotation quality, seven annotators were exposed to a tutorial session discussing the annotation task before starting the annotation. Each was given a sample set to annotate before being recruited to complete the full annotation task. The sample set annotation was used to check the quality of the annotation task. The sample set included gold tweets which allowed us to test how annotators performed on this task. Two groups, each having three annotators were recruited to complete the full task and received monetary compensation for their annotations. During the full annotation task, we also injected gold tweets to assess the quality of annotations. Moreover, the full annotation task included repeated tweets, which were used as an additional way to assess the quality of annotations i.e. certain tweets are repeated twice in the data set and later we verified whether the annotator annotated the same tweet similarly. All annotators passed our gold tweets and were generally consistent with their annotation across repeated tweets. Each annotator had 10% of his assigned tweets as repeated tweets. These tweets were discarded form our data set to avoid confusion on the classifier's side.

To measure the inter-rater agreement per group, we computed Fleiss' kappa, which is used to measure agreement when there are more than two raters. The kappa score between the three annotators was 0.48. While there is no precise rule for interpreting kappa scores, the work in (Landis and Koch, 1977) suggests that such a kappa score translates to having a moderate agreement between the annotators. Substantial agreement is achieved with a kappa score greater than or equal 0.61. The achieved inter-rater agreement highlights the difficulty and subjectivity of this task. Take for example the sample tweets in Table 1. Example (a) is presenting the opinion of the tweet author and since the opinion has no bad words and is not very biased, the annotators considered it credible. Example (b) is also presenting the opinion of the tweet author, but the author said that he has a proof and did not present it; consequently, the annotators considered it as not credible. Example (c) presents the tweet author's point of view that is against the Syrian regime, which some annotators who were subjective agreed with and annotated it as credible, while others did not and annotated the tweet as not credible. It is certainly difficult to achieve higher agreement in tasks that are very subjective and are affected by the annotators background.

When grouping tweets by the day of creation, we found that on average 40% of the tweets generated per day were non-credible tweets. This highlights the importance of building a credibility model for tweets.

3.3 Credibility Features

In this section, we discuss the content-based and user-based features that were extracted from the tweets. Our feature-set is composed of 48 features broadly categorized into content-based and user-based features. Content-based features are features extracted from the tweet itself, whereas user-based features are extracted from the tweet author. Content-based features are composed of 26 features. These features are further grouped into four subcategories, which are sentiment, social, meta, and textual features. The sentiment category is composed of the tweet sentiment, whether positive, negative or neutral. Sentiment has been previously shown to be an indicator of credibility (Castillo et al., 2011; ODonovan et al., 2012; Kang et al., 2012) and hence included in the feature-set. The social category captures the social aspects of the tweet such as the count of user mentions, the number of retweets, etc. which can be all indicators of credibility. For instance, a tweet with many retweets might be more credible than ones with few or zero retweets. The meta category is composed of a single feature which is the day at which the tweet is posted, which might affect credibility

(a)	وقوف العالم متفرجاً امام القضيه السوريه ماهو إلا خوف من ثورة شعوبهم ولكن تركوا سوريا على ماهي عليه لتكون درساً لشعوبهم التي تفكر بالثوره *wqwf AlEAlm mtfrjAF AmAm AlqDyh Alswryh mAhw <lA xwf mn vwrp $Ewbhm wlkn trkwA swryA ElY mAhy Elyh ltkwn drsAF l$Ewbhm Alty tfkr bAlvwrh* 'the world not reacting to the Syrian war is because leaders are afraid from an uprising but their people, but the leaders left Syria in its crisis so their people will think twice before doing an uprising themselves'
(b)	فعلاً الثورة السورية ثورة فتنة وعندي الدليل القاطع *fElAF Alvwrp Alswryp vwrp ftnp wEndy Aldlyl AlqATE* 'the Syrian revolution is surely a sedition and i have the proof'
(c)	نظام الاسد قوة نزعت منهم الرحمه و الأخلاق و الدين . يارب انصر أهلنا في سوريا *nZAm AlAsd qwp nzEt mnhm AlrHmh w Al>xlAq w Aldyn . yArb AnSr >hlnA fy swryA* 'The Syrian regime is deprived from ethics, mercy, and faith. May god bring victory to the Syrian people'

Table 1: (a) a credible tweet, (b) a non-credible tweet, (c) confusing tweet that can be credible or non-credible

(weekday vs. weekend). Finally, the textual category includes features such as the count of exclamation marks and the count of unique characters.

Here is the list of all the content-based features: positive sentiment, negative sentiment, objectivity, count of mentions, has user mention, count of retweets, tweet is a retweet, tweet is a reply, retweeted, day of week, length of tweet in words, count chars and count words, count of urls, length of tweet in chars, count of hashtags, count of unique words, count of unique chars, has hashtag, has url, count of ?, count of !, has !, has ?, count of ellipses, has stock symbol, count of special symbols ($!), used url shortner.

User-based features are composed of 22 features. These features are further grouped into three subcategories, which are network, meta, and timeline. Network features include features that capture the connectivity between the tweet author and other twitter users. For instance, the counts of followers and friends, highlight the popularity of the tweet author. Meta features include registration age of the author, profile picture, whether she is a verified twitter user, etc. Timeline features are features that are extracted from the author's previous tweet posts, for instance, the rate of activity of the tweet author. Here is the list of all the user-based features: count of followers, count of friends, fo/fe, fe/fo, is verified, has description, length of description, has url, has default image, does the image hold a face, length of screen name, registration age, listed count, status count, favorites count, tweet time spacing, status retweet count, retweet fraction, average tweet length , average urls/mentions ratio in tweets, average number of hashtags, average tweet length, focus of user on topic.

We extract such an exhaustive set of features to study their actual impact on credibility assessment. The extraction of most features requires simple computations, with the exception of sentiment which is more complex (discussed next).

3.4 Sentiment Extraction

To extract the sentiment of a given tweet, we used ArSenL (Badaro et al., 2014) which is an Arabic sentiment lexicon. Four existing resources were used in the creation of ArSenL: English WordNet (EWN) (Miller et al., 1990), Arabic WordNet (AWN) (Black et al., 2006), English SentiWordNet (ESWN) (Esuli and Sebastiani, 2006) and the Standard Arabic Morphological Analyzer (SAMA) (Maamouri et al., 2010). For each tweet, we removed all non-Arabic tokens such as URLs, user mentions, and hashtags. Next, a tweet was tokenized and fed into MADAMIRA (Pasha et al., 2014), a morphological analysis tool for Arabic text. Finally using the lemma for each word in the tweet, we extracted its corresponding positive, negative and objective scores from the ArSenL lexicon. To compute the positive score of the whole tweet we compute the average of all the words' positive sentiment in the tweet. The same method is used to obtain the whole tweet's negative and objective scores. Other more complex methods can be used to find the tweet sentiment

[(Hobeica et al., 2011), (Al Sallab et al., 2015), (Badaro et al., 2015), (Baly et al., 2016), (Al Sallab et al., in press 2017)], but we resorted to this method for simplicity.

4 Performance Evaluation

In this section, we present the credibility classifier (CAT) and evaluate it vs. multiple baselines and another well-known method. We used the annotated data and the extracted features to train a random forest decision tree classifier using scikit-learn python library [3]. A majority vote was used to decide on the labels of the tweets. We validate the applicability of our classifier (CAT) by doing two different experimental setups. First, we compare CAT to three baselines. Then, we compare CAT to a state-of-the-art tweet credibility classifier - TweetCred (Gupta et al., 2014).

4.1 CAT versus Baselines

In this experiment, we use the 9,000 annotated tweets to train and test CAT. We trained our classifier using multiple machine-learning algorithms such as NaÃŕve Bayes, SVM and Random Forest Decision Tree, however, we only report the results of the highest attaining algorithm in terms of Weighted Average F-measure (WAF-measure), namely the Random Forest Decision Tree. The Weighted Average F-measure is the sum of all F-measures, each weighted according to the number of instances with that particular class label. The Weighted Average F-measure allows a fair comparison whilst taking into consideration the classifier performance within both credible and non-credible classes. Using 10-fold cross validation, CAT achieved a WAF-measure of 75.8%.

We compared the performance of CAT to three common baselines. The first baseline is the stratified baseline, where the classifier makes random predication in accordance to the distribution of credible and non-credible tweets in the training set. Hence, if the training set includes 80% credible and 20% non-credible tweets, the stratified baseline randomly predicts 80% of the test set to be credible and 20% to be non-credible. The second baseline is one that makes uniform predictions such that both credible and non-credible classes are equally likely. The third baseline is the majority class baseline. Such a classifier predicts all tweets to belong to a single class and this class is the majority class in the training set. Hence, if the training set is mostly composed of credible tweets then each instance in the test set will be labeled credible. Table 2 presents the Weighted average Precision, Recall, F-measure of our classifier CAT in comparison to the three baselines. CAT consistently surpassed the WAF-measure of the baseline approaches indicating that the user-based and content-based features we used are worthy indicators of credibility. When considering the highest WAF-measure among the baselines, CAT achieves a percentage improvement of 47% over the best baseline (difference / original number).

Classifier	Weighted Average Precision	Weighted Average Recall	Weighted Average F-measure
CAT	76.1%	76.3%	**75.8%**
Stratified	51.5%	51.3%	51.4%
Uniform	52.1%	50.5%	50.9%
Majority	35.6%	59.6%	44.6%

Table 2: CAT's against baseline classifiers

4.2 CAT versus TweetCred

We aim to compare CAT to a competitive approach existing in the literature, namely TweetCred (Gupta et al., 2014). We treated both TweetCred and CAT as black boxes, and obtained the credibility scores for each tweet in our data set using TweetCred's API and CAT's classifier . Consequently, we have two annotations for each tweet, the first annotation obtained from CAT and the second annotation obtained from TweetCred. Given these two labels we compare the performance of CAT to TweetCred. According to our knowledge, TweetCred is the best work available on credibility classification on Twitter.

TweetCred is a real-time web-based system for assessing credibility. It relies in its classification on features that can be extracted in real-time only; hence TweetCred may assess a new twitter feed in any language. Details of TweetCred were presented in the related work section. Since we could not receive TweetCred scores for some of the tweets, we removed those tweets from the experiment and re-evaluated CATs performance, in order to keep the comparison fair. The scores obtained from TweetCred API ranged from 1 to 7, where 1 indicates low credibility and 7 indicates high credibility. To fairly compare CAT to TweetCred we must project TweetCred's credibil-

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

ity scores to two values, namely credible or non-credible. To determine the cut-off threshold below which a tweet is non-credible using TweetCred, we used our annotations and TweetCred scores to train a decision tree. The cut-off threshold was determined to be 3. Hence, any tweet receiving a TweetCred score less than or equal to 3 is non-credible and is credible otherwise. Table 3 depicts the Weighted Average Precision, Recall, and the WAF-measure of both TweetCred and CAT. CAT outperforms TweetCred when classifying tweets by 16.7% in terms of the percentage increase in WAF-measure. Our intuition is that CAT outperformed TweetCred because TweetCred relies in its classification on real-time features only, such as, count of retweets and count of friends, while ignoring the tweet semantics and the author clues.

Classifier	Weighted Average Precision	Weighted Average Recall	Weighted Average F-measure
CAT	66.8%	67.1%	**67.1%**
TweetCred	58.6%	56.9%	**57.5%**

Table 3: CAT's against TweetCred Classifier

5 Feature Analysis

5.1 Content-based vs. User-based features

In this section, we present comparative analysis when training our classifier using content-based features versus user-based features. The main objective of this comparison is to know whether content based features only or user-based features only can be used as deciders for credibility. We trained our classifier using user-based features only and performed 10-fold cross validation. We repeated the same experiment but using content-based features only. As shown in table 4, a WAF-measure of 68.9% was achieved when using user-based features alone, which is 0.2% more than the WAF-measure achieved when content-based features were used alone. However, the best results are achieved when the features are combined. Consequently, we cannot solely rely on the tweet content alone or the author features alone to decide on tweet credibility; rather a combination of both cues is needed for a robust judgment.

5.2 Feature Reduction

In this section, we describe our feature reduction experiment that aimed at retaining worthy features and discarding features that might be misleading and harming the performance of our classifier. This process was composed of two steps. Step 1 involved picking a subset of features, and step 2 involved evaluating the efficiency of the selected subset. These two steps were repeated until the desired improvement was achieved.

Features	Weighted Average Precision	Weighted Average Recall	Weighted Average F-measure
user	69.7%	70.1%	68.9%
content	69.1%	68.5%	68.7%
CAT	76.1%	76.3%	**75.8%**

Table 4: CAT's evaluation using different feature sets

For Step 1 (picking a subset of the features), we used best-first search implementation available in WEKA - a well established data mining tool - to traverse the feature space (Hall et al., 2009). The feature space was represented as a graph and each node in the graph represented a possible combination of the available features. Hence, in total our feature space contained 2^{48} nodes. Edges connecting the graph nodes were determined by the content of each subset node. A node had an edge to another if the other node either added or removed a feature from the node's combination of features. Traversing the graph starting from an empty node (containing no features) and moving only along the edges that add a feature to the current combination is called forward traversal. On the other hand, starting from a full node (containing all 48 features) and moving along the edges that remove a feature from the current combination is called backward traversal. We performed feature reduction with both forward and backward traversal separately.

For Step 2 (evaluating the chosen subset), after deciding on the search direction and picking a starting node from the feature space graph, we evaluate the efficiency of the node's subset of features as follows. We build a classifier using the combination of features in the selected node and we perform 10-fold cross validation and keep track of the WAF-measure of the built classifier. Feature reduction is an optimization problem and we cannot predict how neighboring nodes will perform or whether a node will get us closer to our goal or not. Traversing the whole graph and evaluating every possible node in the feature space

will give us the best feature subset, but this is not feasible. Consequently, we must determine a stale state, that is, a state after which the algorithm terminates graph traversal. The stale state for this experiment was set to 100. Consequently, the traversal algorithm terminated once it had expanded 100 nodes that did not improve on the best WAF-measure seen so far.

We evaluated around 1000-4000 subsets and the selected feature subsets were each less than half the original size, yet each outperformed the original feature set when it comes to WAF-measure. The best representative features were found to be as follows:

- User Features: Follower count, listed count, has description, has url, retweet fraction, average hashtags per tweet, average urls per tweet, tweet spacing (in minutes), expertise, average tweet length (in words), follower/friends ratio
- Content Features: count of url, negative sentiment score, count of exclamation, has url, count of unique chars, count of hashtag, count of ellipse

One of the features that was found to be highly crucial in determining the credibility of a given tweet is its sentiment, specifically the negative sentiment. Also, five of the effective features are related to URLs. For example, one of the features that was found to be very useful is the presence of a URL in the author's Twitter profile linking to her website. We found that 74% of the tweets whose authors provided a URL were credible, in contrast to only 47% of tweets whose authors' profiles were missing a URL. We conclude that tweets whose authors' profiles contain URLs are more likely to be credible than those that do not. We also noticed that the presence of a URL in a tweet is a very important feature. We found that 80% of tweets that had a URL in them were credible, whereas only 40% of tweets without a URL were credible. The above highlights the importance of the presence of a URL in both the tweet and in its author's profile. In addition, we also noticed that not only the presence of a URL is important, but also the count of URLs. For example, the average URLs count per tweet is another crucial feature. It is computed by looking at the previous tweet posts for the tweet author and computing the count of URLs she uses on average. We found that 88% of the tweets that were generated by authors who had an average of one URL per tweet in their history were credible. Moreover, we found that only 39% of tweets that did not have a URL linking to an external source were credible. On the other hand, 79% and 100% of tweets with 1 and 2 URLs, respectively, were annotated as credible.

All of the above are clear indicators of the importance of the presence and the count of URLs whether in the author profile, his/her past tweets or in the tweet itself. Next, we highlight relevant features extracted from the user timeline. Timeline features are user features extracted from the author's tweet history. The original feature set included 8 such features. To the best of our knowledge none of these features have been previously used for credibility classification. We found 5 features from the user timeline to be highly effective for credibility classification, namely retweet fraction, average URLs per tweet, tweet spacing (in minutes), average tweet length (in words), average hashtags per tweet. For instance, we observed that 85% of the tweets whose authors had on average 1.25 hashtags in their history were credible. We also noticed that authors who had a high retweet fraction had a higher probability of generating non-credible tweets.

Finally, while it is interesting and useful to know what are the most relevant features for credibility, we were also interested in finding the least important ones. To find such features we performed the same steps used to find the most important features for credibility classification however instead of using the reduced feature sets, we used their inverses. The inverse set of a reduced feature set will include all features in the feature space that have not been selected by feature reduction. One feature that was missing from all the reduced sets was the day of the week. This means that the day at which the tweet is generated has no correlation to its credibility and this is intuitive. Another irrelevant feature is the count and the presence of a user mention. Also the count of character and character to word ratio of the tweet were deemed irrelevant to credibility classification.

6 Conclusion

In this paper, we presented a novel credibility model for tweets called CAT. Our model is based on a machine-learning approach, and makes use of an exhaustive list of features, some of which are

user based and some that are extracted from the text of the tweet itself. Our feature set includes many features extracted from the timeline of the tweet's author, which have never been looked at in the context of credibility. To test the validity of our model, we annotated a corpus of 9,000 Arabic tweets that are topic independent. The annotated corpus was used to train a binary classifier that consistently outperformed all baselines and a state-of-the-art approach in terms of Weighted Average F-measure. We also conducted a thorough analysis of the annotated corpus and carefully studied the effect of the various features on credibility prediction. For future work, we plan to incorporate Arabic specific features, for instance part of speech (POS) tags, and check their effect on classifying credibility. We also plan to try the exact method with the same set of features on tweets from other languages, and see if the proposed classifier continues to perform well on languages other than Arabic.

References

Ahmad A Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *ANLP Workshop 2015*, page 9.

Ahmad A Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B Shaban. in press 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Ayman Al Zaatari, Reem El Ballouli, Shady ELbassouni, Wassim El-Hajj, Hazem Hajj, Khaled Bashir Shaban, and Nizar Habash. 2016. Arabic corpora for credibility analysis. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4396–4401.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.

Amal Abdullah AlMansour. 2016. *CREDIBILITY ASSESSMENT FOR ARABIC MICRO-BLOGS USING NOISY LABELS*. Ph.D. thesis, King's College London.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. ANLP.

Gilbert Badaro, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2015. A light lexicon-based mobile application for sentiment mining of arabic tweets. In *ANLP Workshop 2015*, page 18.

Ramy Baly, Roula Hobeica, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, and Ahmad Al-Sallab. 2016. A meta-framework for modeling the human reading process in sentiment analysis. *ACM Transactions on Information Systems (TOIS)*, 35(1).

W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. 2006. Introducing the arabic wordnet project. Proceedings of the 3rd International WordNet Conference (GWC-06), pp. 295-299.

Kevin Canini, Bongwon Suh, and Peter Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. IEEE Third International Conference on Social Computing (Socialcom).

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. 20th International Conference on World Wide Web.

A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of LREC pp. 417-422.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: A real-time web-based system for assessing credibility of content on twitter. arXiv Preprint arXiv:1405.5490.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Roula Hobeica, Hazem Hajj, and Wassim El Hajj. 2011. Machine reading for notion-based sentiment mining. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 75–80. IEEE.

Byungkyu Kang, John O'Donovan, and Tobias HÃűllerer. 2012. Modeling topic specific credibility on twitter. ACM International Conference on Intelligent User Interfaces.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

M. Maamouri, D. Graff, B. Bouziri, S. Krouna, and S. Kulick. 2010. Ldc standard arabic morphological analyzer (sama) v. 3.1. LDC Catalog no.LDC2010L01.ISBN, pp. 1-58563, 2010.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to wordnet: An online lexical database. International Journal of Lexicography, vol. 3, pp. 235-244.

John ODonovan, Byungkyu Kang, Georg Meyer, Tobias Hollerer, and Sibel Adalii. 2012. Credibility in context: An analysis of feature distributions in twitter. Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom).

A. Pasha, M. Al-Badrashiny, A. E. Kholy, R. Eskander, M. Diab, N. Habash, M. Pooleery, O. Rambow, and R. Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland.

Mohammad Bin Sultan, Hend AlKhalifa, and Abdul-Malik AlSalman. 2010. Measuring the credibility of arabic text content in twitter. Digital Information Management (ICDIM).