A Survey on Credibility Detection Approaches in Twitter

Noha Y. Hassan [1], Wael H. Gomaa [2], Ghada A. Khoriba [3], Mohammed H. Haggag [4]
Assistant Lecturer, Computer Science Dept., Beni-Suef University, Egypt.[1]
Lecturer of Computer Science, Beni-Suef University, Egypt.[2]
Lecture of Computer Science, Helwan University, Egypt.[3]
Professor of Computer Science, Helwan University, Egypt. [4]

***Abstract* - Twitter is the most popular micro-blogging which allows users to exchange short messages, provides a platform for common people to share experiences and opinions. Nowadays, Twitter counts with an average 328 million monthly active users and is growing rapidly specially among young people who might be influenced by the information from anonymous sources. Detecting credible or trustworthy information on Twitter becomes a necessity, especially during high impact events. In this paper we discuss previous studies on assessing credibility of Twitter messages including human based models and automatic models. This survey classifies the different models for automatic credibility assessment into three approaches: classification-based approach, propagation-based approach and similarity-based approach. The survey presents a comparison of these models based on the used techniques and feature set. Furthermore, an overview of the existing systems which developed a practical implementation of the credibility problem is presented. The paper discusses the human based models including: the user surveys and statistical analysis models which can aid in the design of credibility assessment models by better understanding of features distribution and users' perceptions of information credibility.**

Keywords— Micro-blogging; Twitter; Credibility detection; Classification; Propagation; Feature extraction.

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]. Due to quick response time, they allow users to spread real time news update to many people [2]. Twitter is the most famous micro-blogging service that allows users to post and exchange short messages or "tweets". Tweets are shared with the author' followers and can be easily disseminated through "re-tweet". Recently, Twitter has been considered as the most micro-blogging platform used as news source [3, 4]. News on Twitter comes from different sources: some from authorized news organizations, while most from public users. Unlike traditional media sources, the absence of supervision and quality control makes Twitter a suitable environment for spreading rumors [5]. This issue becomes problematic as more people depend on social media to obtain information and news especially in high impact events [6, 7]. Another research by Gupta et al. [8] analyzed the spreading of rumors on Twitter during "Hurricane Sandy" and discovered that about 86% of the fake tweets were re-tweets. They found out that during the crisis people share news even if it is from an unknown source. Moreover, a recent study [9] stated that fake news published on social media during the American presidential elections in 2016, had a significant effect on voters. Many researches revealed that a lot of content on Twitter may be incredible [10-12]. Therefore, detecting credible or trustworthy information in Twitter becomes a necessity.

This paper is organized as follows: section two presents the information credibility problem on Twitter. Section three presents the different approaches for automatic credibility assessment. Section four presents the existing commercial credibility assessment applications and section five includes the human based approaches. Finally, section six includes the conclusion.

II. INFORMATION CREDIBILITY ON TWITTER

Credibility is defined as "the quality of being trusted and believed in"[1]. In fact, it is hard to determine the credible tweets manually. The number of followers of a user and the number of re-tweets of a tweet cannot indicate trust because malicious users can easily forge followers or re-tweets. Moreover, Twitter users often re-tweet without verifying the content [13]. Recently, several approaches have been proposed to handle this challenge. Figure 1 shows our classification for credibility assessment approaches which is divided into human based and automatic approaches. Human based approaches including user surveys (sample of people who are asked to fill out questionnaires to determine the most important credibility factors) and studies which apply statistical approaches are discussed in section five. Proposed approaches which use algorithms for automatic classification of tweets according to their credibility are discussed in the next section.

---

[1] http://www.oxforddictionaries.com

III. AUTOMATIC ASSESSMENT OF INFORMATION CREDIBILITY

There has been extensive research aiming at determining the credibility level of Twitter messages automatically using different methods. After surveying the literature, these methods are categorized into three categories: 1) classification based approaches which use machine learning mainly supervised learning methods [14-24, 42, 59], 2) propagation based approaches, which exploit the network structure of users and tweets using graph analysis [5, 25-30], 3) similarity based approaches which depend on measuring similarity with credible sources [31, 32].

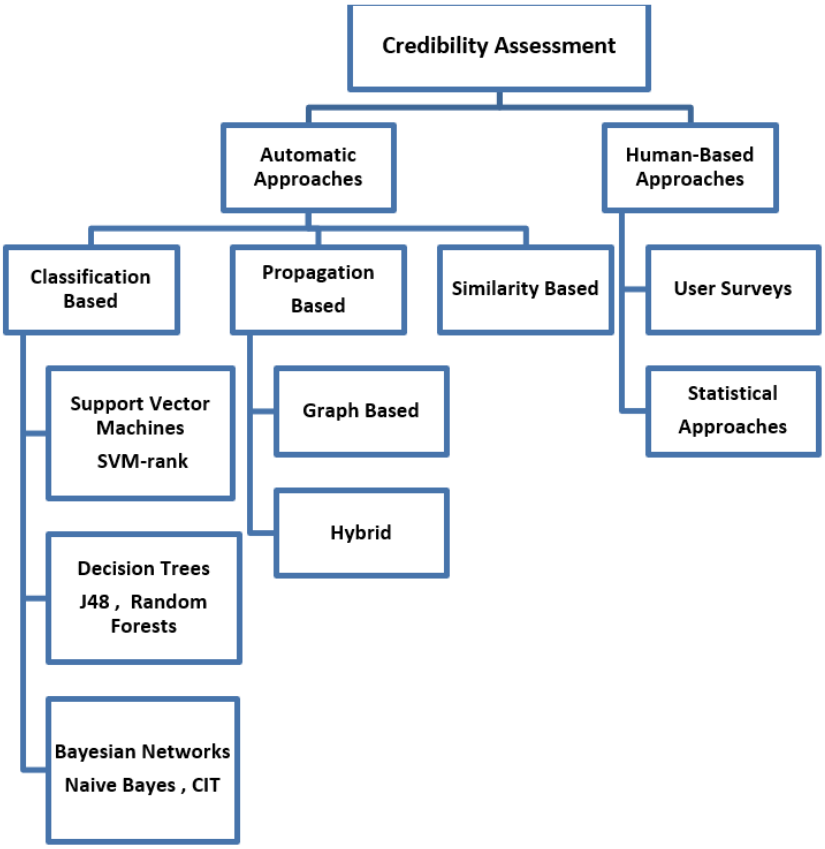

Figure 1: Classification of credibility assessment approaches

There are several useful cues or features introduced by the previous research to assess credibility. Most of these studies rely on content and source based features. Figure 2 shows the different types of features used by the previous studies. Content-based features focus on the content of the message itself and can be grouped into: textual, network, meta-data, linguistic and sentiment features. Textual features include features related to the text of the message such as the length of the message, number of unique characters or emoticons and if the message contains a hashtag (#) or URLs. Network features captures the network aspects of the message such as if the message is a re-tweet and the number of user mentions, while meta-data features includes features such as the posting date. Some studies focus on the linguistic features [31, 33, 34] such as unigram and bigram based lexical features. Moreover, sentiment features of the tweet whether positive, negative or neutral have been proved to be good indicator of credibility [14, 17, 21, 35, 36]. Source-based features consider characteristics of the user as the source of the tweet and can be grouped into two subcategories: social and meta-data features. Social features include features that capture the connectivity between the author and other Twitter users such as the number of followers and followees. Meta-data includes features such as registration age and if the user is verified, in addition some studies introduced new features from the timeline like the number of tweets the user has posted and the existence of profile pictures [21, 22]. Some of the previous studies computed aggregations from the content and source feature sets (Topic features), such as the fraction of tweets that contain URLs within a specific topic and the fraction of tweets from verified users. Other studies consider characteristics related to the propagation tree that can be built from the re-tweets of a message such as the depth of the re-tweet tree [14]. Finally, some studies introduced external resource features such as “Web of Trust” reputation (WOT) score for tweets containing URLs[16].

Most research in information credibility automatically was based on English content. According to experts, social media specifically Twitter played a significant role during the recent political changes in Arab countries [37]. Recently, there have been few efforts to assess and analyze the Twitter messages in Arabic language [18, 19, 23, 31] and to identify Arabic credibility prominent features [22]. Finally, there is some research has gone as far as to implement their proposed approaches in a practical

manner [16, 20, 21, 38-40]. TweetCred [16] and TweetBot [20] are examples of real-time web-based systems that aim at assessing the credibility of English Twitter messages. Moreover, a recent research [21] introduced CAT (Credibility Analysis of Arabic Content on Twitter) system to automatically predict the credibility of Arabic tweets. These systems are discussed in section four explaining the advantages and drawbacks of each system.

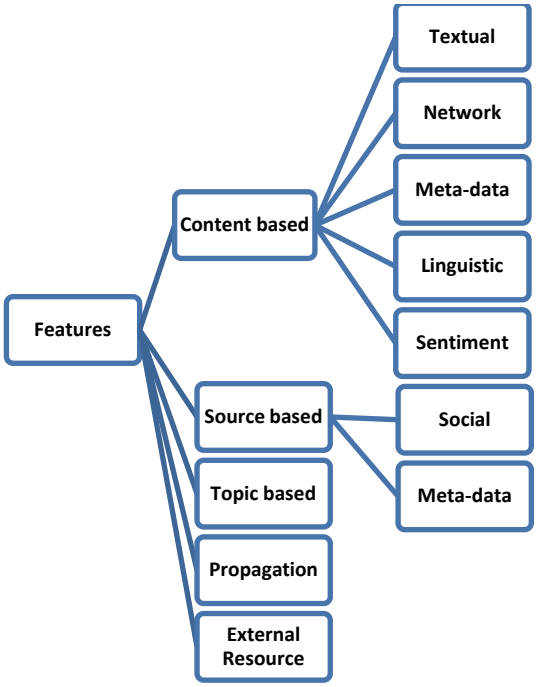

Figure 2: Existing credibility Twitter features categories

A. *Classification-Based Approaches*

Previous research in these approaches in general classifies tweets based on the extracted features using machine learning techniques especially supervised techniques [14-24]. As shown in figure 3, these techniques require building a ground truth that contains a collection of tweet messages with the features related to them. The messages in these datasets are then labeled by human annotators which is an important step affecting the prediction model [41]. This annotated dataset will then act as a source of training data (or ground truth) for machine learning techniques to build automatic classifiers that can accurately determine the credibility of a given tweet. Table 1 summarizes the research works and illustrating, for each approach, the used dataset size and language, the used algorithms and the type of extracted features. Decision trees, support vector machines (SVM) and Bayesian networks are the most popular supervised learning techniques used for classification.

Figure 3: General model for supervised classification-based techniques

1) Decision trees: Castillo et al. were the first to work on solving the Twitter credibility problem automatically [14, 42], by using classification-based approaches. In their first research [14] they focused on tweets related to "trending" topics and used automatic methods to measure their credibility using features extracted from them. They identified four types of features: content based, source based, topic based, and propagation-based features such as the depth of the re-tweet tree. Twitter Monitor [43] was used to collect 2500 trending topics with at most 10,000 tweets in each topic. The annotation of the dataset was performed by evaluators assistant and included two rounds: the first-round separates posts which contain information about news events (labeled as NEWS), from personal opinions (labeled as CHAT). Then for the NEWS tweets, another group of evaluators classify them into credible/not credible. For the human annotation task, they asked for 7 different assessments and labels for each topic require the agreement of at least 5 evaluators. As a next step, they trained several learning algorithms such as "SVM, decision trees, decision rules, and Bayesian networks", but best results were achieved by "J48 decision tree" [44]. Figure 4 shows the decision tree built by the research for the credibility classification where A="credible" and B="not credible".

Figure 4: Decision tree built by Castillo et al. [14] for credibility assessment problem

The research performed a 3-fold cross validation and reached 86% for credibility classification and 89% for news/chat classification. The research provided a feature analysis to illustrate the following contributions: 1) Tweets without URLs are often related to non-credible content, while negative sentiment indicates credibility. 2) More active users (users who have many tweets or have large number of followers/friends) tend to spread more credible information, 3) the tweet which has many re-tweets seems to be more credible. Castillo et al. extended their research in [42] by re-designing the learning scheme and tested the new model on data collected during and after the Earthquake in Chile 2010. The labeling process used a crowd-sourcing tool where labels are NEWS or CHAT or UNSURE. The 'UNSURE' label identifies tweets that were not labeled as NEWS or CHAT messages. By removing these tweets from the training set, the performance of this approach improved considerably. Finally, the research tested the accuracy of this model on Twitter topics in Spanish posted during the Earthquake.

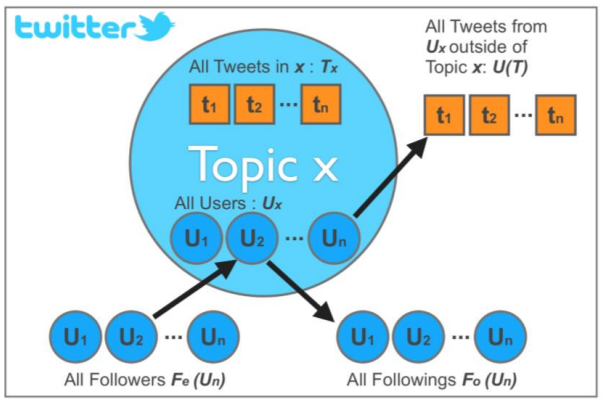

Figure 5: Social Model crawling algorithm [17]

Kang et al. [17] introduced three models for topic-based credibility assessment. The first model (social model) focuses on features related to the social network such as followers/following relationship and re-tweets. Figure 5 illustrates the social model

crawling algorithm. The second model (content model) is a language-based model defining 19 content features related to specific topics such as number of URLs and mentions. The last model (hybrid model) is a combination of the previous two models using different ways to predict credibility. By applying ten-fold cross validation and using J48 decision tree classifier, the social model achieved 88.17% accuracy rate while the content and hybrid models achieved 63% and 67% respectively. The approach represented in this research differs from Castillo et al. [14] which performed similar evaluation in that it focuses on individual tweets. The results of this research indicated that features from the underlying social network are better than content features in predicting credibility.

Another research that assesses the credibility of individual tweets using decision tree learning algorithm is the one by Lorek and Gupta [20]. The research focused on manual annotation, experts were asked to take some attributes into consideration such as visible features of the tweet (e.g., profile photo, profile name, and "account verified" mark). The research focused on the external link features as it was proved to be prominent features in assessing credibility [14, 18, 42]. Experts were asked to check if the target website reached just after clicking the link or it leads to an interactive ad instead and if the link's content matches the tweet content. The next step, they developed an algorithm of reconciling different manually assigned scores, according to reconciliation rules. All embedded URLs were extracted from tweet content and processed by the Reconcile platform. This adds the Reconcile features and Reconcile score to the dataset which gives a satisfactory enhancement to the machine learning process. For machine learning, they used random forest implementation [56] in three cases: based on twitter features, based on reconcile features and finally based on a combination of the two sets. The best results were given with the combined classifier as it achieved 89% of recognition precision. At the end of the experiments, the research concluded that it is still more important what the user posts and not who they really are. AlMansour et al. [22] used supervised machine learning methods and to identify credible information in Arabic context. A statistical approach based on features frequencies was used to identify credibility prominent features for Arabic content . The research introduced new features such as: the use of natural user names, the use of religious words, and whether the authors' locations are related to the topic. The dataset contained 199 source tweet messages, 6249 tweet messages including re-tweet. The experiments were held using different classification algorithms such as: J48, Random Forest tree, Naive Bayes, Logistic Regression, SVM and k-Nearest Neighbor to evaluate the performance of the proposed model. The accuracy results reached 77.4% using Random forest decision tree classifier using 10-fold cross validation. Similarly, CAT (Credibility Analysis of Arabic Content on Twitter) system proposed by ElBallouli et al. [21] to automatically predict the credibility of a given Arabic tweet using Random forest decision tree will be discussed later in section four.

2) Bayesian Networks: Xia et al. [24] presented a supervised method using Bayesian network to predict the credibility of Twitter messages in emergency situations. Five experts were asked to manually label 350 tweets related to English Riots into credible or not credible. Several features were extracted, from the annotated tweets, including the social behavior and diffusion features (related to time such as the time of the original tweet has been retweeted by other users) in addition to the content and topic-based features. Regarding the machine learning process, they used CIT "conditional independence test-based" learning algorithm (CIT) [68]. Experiments results showed an accuracy rate between 61% and 66% using different classifiers, while it reached 63.66% with the proposed algorithm.

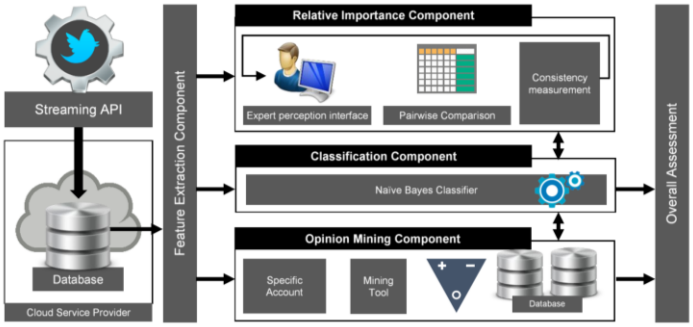

Figure 6: Credibility assessment model using relative importance [18]

Al-Rubaian et al. [18] proposed a multi-stage credibility assessment model for Arabic tweets which considered relative importance to study its effect on the credibility classification. Figure 6 illustrates their model. The research found out that some features are qualitative and require human assistance to determine its importance. They performed a pair-wise comparison in order to generate weights of the features in a numerical scale from 1 to 9 forming the priority vector. Naïve Bayes classifier [49] with the priority vector was used to build the classification model.

Another enhancement outlined in this research is the effect of sentiment analysis on credibility assessment. They used the SAMAR technique [50] to measure the sentiment of Arabic content. The overall credibility score was calculated depending on

three values: 1) the credibility score calculated from the classifier, 2) the value of user content verification calculated by applying the Naïve Bayes classifier on all user's content, and 3) the positive sentiment of all user's tweets which indicates the user's behavior [35]. The ground truth was created by collecting 1000 unique tweets from 700 unique accounts written in Arabic language. The proposed model achieved 86.24% Precision, 98.8% recall and 90.3% accuracy.

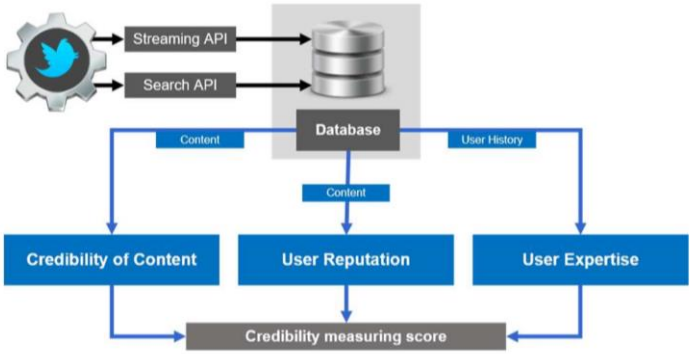

Figure 7: Hybrid reputation-based approach [19]

The relative importance concept was introduced again in [19] combined with other components to determine the credibility of Twitter messages. The hybrid approach proposed in this research comprises four integrated components: user expertise model, feature ranking algorithm, reputation-based model, and credibility assessment engine as illustrated in Figure 7. Reputation of a Twitter user can be assessed based on his popularity and sentimentality regarding a particular topic [52, 53]. Sentiment score is calculated based on the number of positive and negative tweets in the user's history, while the popularity score is based on features related to the user's reputation such as number of followers, favorites and re-tweets of the user [51]. The model was evaluated with data for 2,843 Twitter users and more than 11,000 tweets and achieved accuracy ranged from 93% to 95% when applying different classifiers, 96% with feature-rank Naïve Bayes. Finally, a recent research [23] applied the previous model on two large Twitter datasets (187,614 English tweets, 186,819 Arabic tweets) to study the effect of the relative importance algorithm and the results showed that the model with the relative importance approach achieved higher accuracy rate.

TABLE I : EVALUATION OF THE EXISTING SUPERVISED MACHINE LEARNING MODELS

Reference	Year	Language	Dataset	Extracted features	Classifier
[14] Castillo	2011	English	2,524 tweets	message, user, topic (aggregation), propagation features	J48 decision tree
[15] Gupta	2012	English	7,000 tweets	Message, user features	SVM rank
[17] Kang	2012	English	5,025 tweets	Message, user, aggregated features	J48 decision tree
[24] Xia	2012	English	350 tweets	Message, user, topic, diffusion(time) features	CIT Bayesian Network
[42] Castillo	2013	English, Spanish	165,312 tweets	message, user, topic (aggregation), propagation features	Random Forest, Bayes Net
[16] Gupta	2014	English	6,000 tweets	Message, message meta-data, user, network, linguistic and external resource features.	Coordinate Ascent
[18] AlRubaian	2015	Arabic	1,000 tweets	Message, user, aggregated features	Naïve Bayes
[20] Lorek	2015	English	1,206 tweets	User and reconcile features	Random forest
[19] AlRubaian	2016	Arabic	11,000 tweets	Message, user, aggregated features	feature-rank Naïve Bayes
[22] AlMansour	2016	Arabic	199 tweets	Message, user, aggregated features	Random Forest decision tree
[21] El Ballouli	2017	Arabic	9000 tweets	Content (sentiment, social, meta, textual features), user (network, meta, timeline) features	Random forest decision tree
[23] AlRubaian	2017	English, Arabic	187,614 English, 186,819 Arabic tweets	Message, topic, user features	Naive Bayes
[59] Sabbeh	2018	Arabic	800 tweets	User, content (similarity with verified content, comments polarity) features	Decision tree

3) SVM: Gupta et al. [15] used a combination of supervised machine learning and relevance feedback to measure credibility. Figure 8 illustrates the components of the model. They used "SVM ranking algorithm" [45] to build their classification model then they evaluated an enhancement to the ranking technique by using PRF "pseudo feedback relevance re-ranking" scheme [46]. PRF was used to re-rank the ranked documents by calculating text similarity between the most frequent unigrams from the top ranked tweets and the other tweets. Text similarity was calculated using metric BM25 [47], then they used NDCG (Normalized Discounted Cumulative Gain) [48] to evaluate the proposed approach. Similarly, as described in [14], the research performed a regression analysis to identify the most prominent features. Number of followers, number of unique characters and swear words were the most effective features. 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. Gupta et al. [16] extended the previous efforts and proposed a real-time system named TweetCred which will be discussed later in section four.

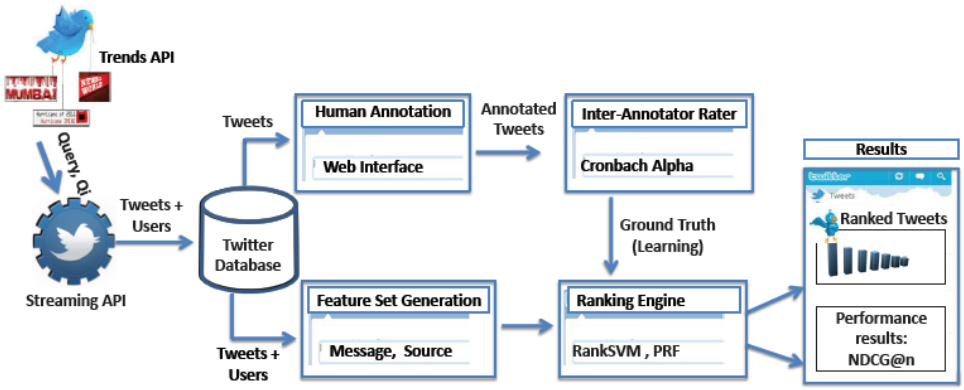

Figure 8: Twitter credibility ranking system developed by Gupta et al. [15]

B. *Propagation Based Approaches*

The approaches that focus on the propagation concept to detect the credibility rely on the network structure and social graph analysis. Social networks can be represented as a graph composed of nodes (Twitter users) and relationships connecting them (such as: follows, replies, mentions and tweet) called edges. These inter-entity relationships on Twitter can provide rich information and many researches incorporated graph analysis to measure information credibility. Ravikumar et al. [26] proposed a method to rank tweets according to their credibility scores and content-based popularity. The approach exploited the relationships between the tweets and modeled Twitter as a three-layer graph consisting of tweets (based on similarity), users (based on follower-followee relationship), and web pages using PageRank. The three layers are illustrated in Figure 9. Within the tweets layer, they computed the content (semantic) agreement between the tweets using "Soft-TFIDF" with "Jaro-Winkler" similarity [58] as it was proved that the agreement is more indicative of credibility than re-tweets [57]. The model derived trust scores of entities in the three layers, then propagated the scores to tweets considering the inter-layer relations to compute a single tweet score.

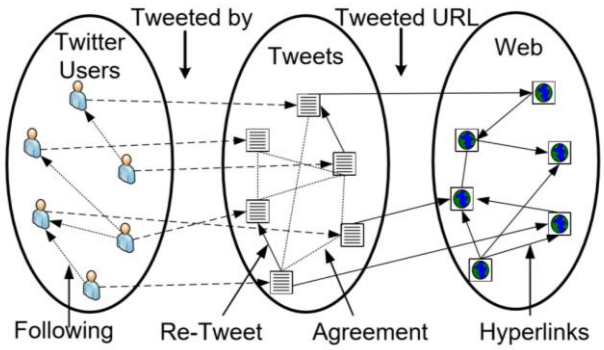

Figure 9: Three layers system of Twitter space [26]

Another research by Gupta [27] presented a hybrid approach for credibility analysis with event graph-based optimization and machine learning technique to compute the credibility of Twitter events. The research started from the model described by Castillo [14] and introduced new features for enhancement. The proposed approach included two additional modules: BasicCA (Basic Credibility Analysis) and EventOptCA (Event Graph Optimization). BasicCA first initializes the credibility of different tweets, users and events in the network using the classifier results. Then, it performs Page rank iterations to propagate authority

across the network as shown in Figure 10. EventOptCA builds another graph of events in each iteration and updates event credibility values by assigning similar scores to similar events. They performed the experiments on two different datasets and the results showed that the event graph optimization approach outperforms the classification based approach introduced by the authors. Moreover, it was clearly noticed that their classification based approach achieved lower accuracy compared to Castillo et al. [14] which uses the same classifier (J48 decision tree) and achieved 86% accuracy rate.

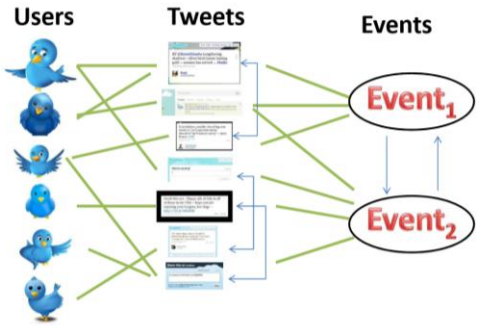

Figure 10: BasicCA module [27]

Jin et al. [25] proved that each event contains a combination of credible and not credible information. They defined a new layer named "sub-event layer" to capture deeper semantic information for an event where the credibility of an event is the expected credibility of all sub-events belong to it. Experiments on two real-world datasets showed that the proposed model can achieve improvements in accuracy by more than 6% and F-score by more than 16% compared with baseline methods. Zhao et al. [28] developed a new propagation model to measure the credibility of Twitter users and messages. The proposed model contains two modules: trust evaluation based on similarity and trust propagation. Figure 11 illustrates the components of the two modules. The first module was used to rank users and tweets against credible ones based on three similarity features (textual, spatial and temporal). The second module was used to update the user/tweet credibility scores through iterative propagation according to four propagation rules. Based on precision and F-measure values, this method outperforms the baseline supervised learning method. Mendoza et al. [5] explored the behavior of users under an emergency (Earthquake in Chile 2010) and analyzed how messages were propagated through the network. The analysis outlined the key differences between the propagation of tweets that correspond to rumors and tweets that spread news. The research discovered that rumors is more questionable than truthful news which outlined that it is possible to detect rumors by automatically identify highly questioned information.

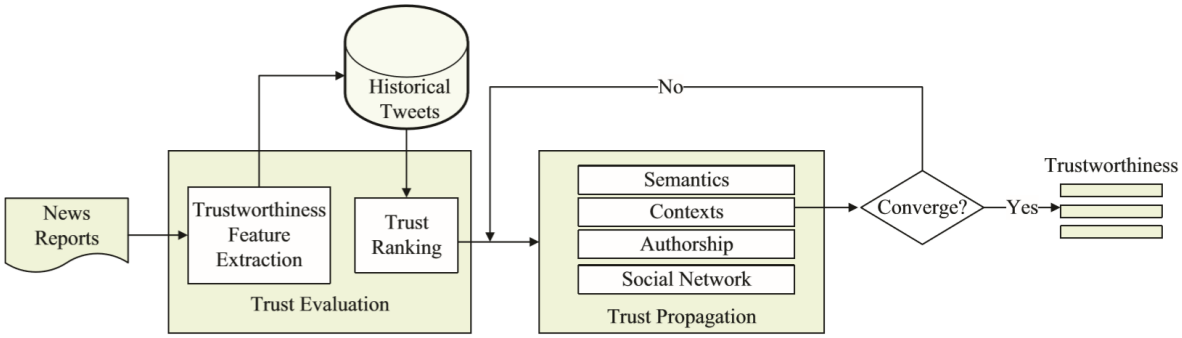

Figure 11: Topic-focused trust evaluation and trust propagation modules [28]

The approach described by Gündüz [55] proposed a hybrid approach by applying classification algorithms on the tweet set then the results were improved by means of graph-based techniques. In the graph-based phase, a connected graph from users and tweets was constructed according to authority (user and tweet), friendship (users) and similarity (tweets) relations. They introduced new credibility definition based on three dimensions: 1) being newsworthy, 2) is not a spam and 3) is free from offensive words. Each one of those dimensions was evaluated separately and the result is a combination of the three dimensions' results where a credible tweet is free from offensive words, free from spam and is newsworthy

C. Similarity Based Approaches

Credibility assessment has been studied from another point of view based on similarity. Al-Khalifa et al. [31, 32] developed a model to measure credibility of Arabic tweets and assign credibility level (high, low and moderate) to each tweet. The model uses two approaches: the first is based on the similarity between Twitter messages and authorized news sources like Aljazeera.net. The second model is based on the calculated similarity in addition to extra features related to the author and the content of the tweet.

The proposed model consists of four stages: text preprocessing, feature extraction, credibility calculation and credibility ranking. As shown in Figure 12(a), preprocessing includes normalization, stop words removal, POS Part of speech tagging and stemming. After text preprocessing steps, the research used term frequency inverse document frequency (TF-IDF) weight and cosine similarity measure to calculate the similarity with verified content value as shown in Figure 12(b). For the second approach, a set of features was used such as: presence of inappropriate words, link to authoritative existence and if the author is verified.

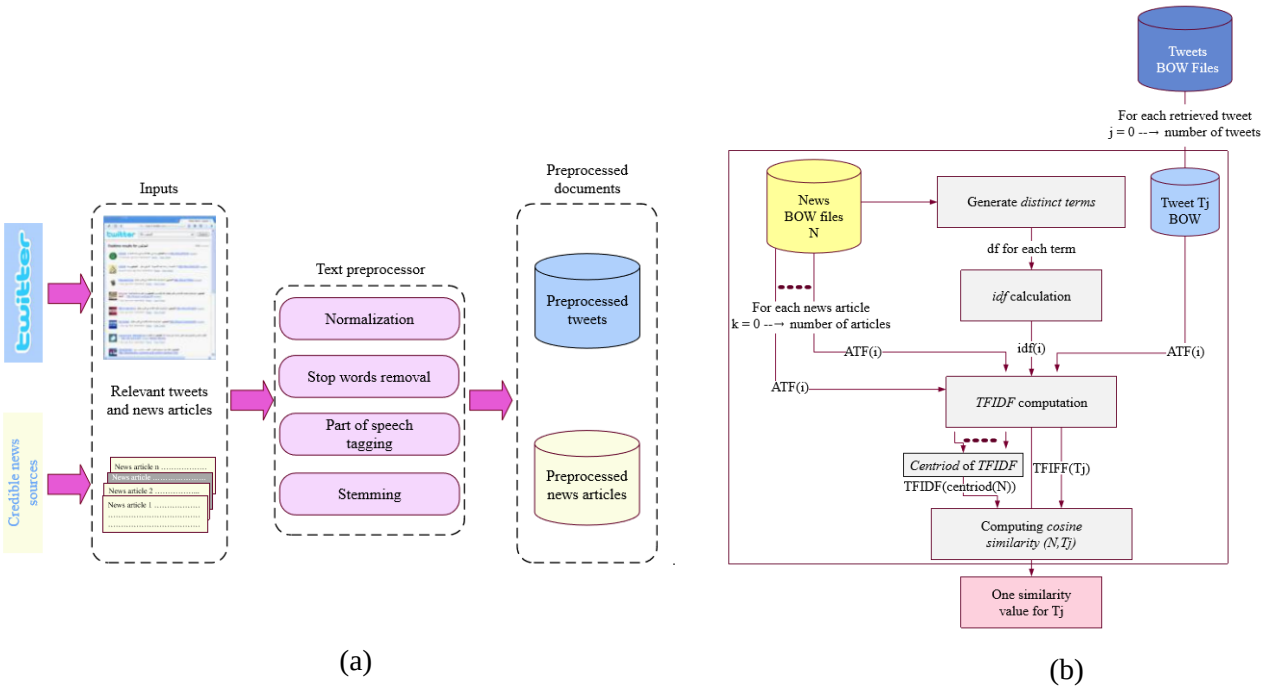

Figure 12: (a)Text preprocessing components , (b)Similarity computing steps [31]

The following formula was used to calculate credibility score:

"Credibility Score = 0.6*similarity + 0.2*inappropriate words + 0.1*authoritative link + 0.1*verified author" [31].

The experiments showed that the first approach (similarity only) has a higher precision and recall. The evaluation of the second approach indicated that the linking feature was the most prominent feature. The proposed system achieved acceptable results in assessing credibility of Arabic tweets but requires the existence of credible external sources. Another research that considered similarity with trusted news sources in determining the credibility of tweets is the work presented in [59]. The research utilized a hybrid set of features based on user and content features as well as content verifiability against verified external news sources. The model is illustrated in Figure 13 and contains four basic modules: feature extraction, content verification, users' comments polarity evaluation and credibility classification. Content verification module is based on the work in [30] which uses cosine similarity between tweets and external content. The research introduced a new feature related to the users' comments polarity as a significant indicator for credibility. The polarity of each comment is given based on the occurrence of negative and positive words, then the total polarity of the topic is calculated by the weighted sum of polarity of all comments. Feature matrix (includes extracted content and user features, verification score and comments polarity) was then fed into the credibility classification module. Three different classifiers (SVM, decision tree and Naïve Bayes) were trained using a dataset of 800 annotated news tweets. The experimental results indicated that decision tree classifier achieved best results in terms of accuracy and F-measure.

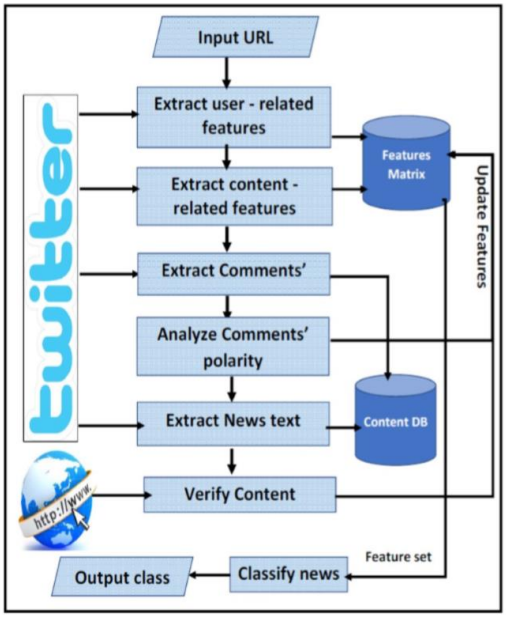

Figure 13: credibility assessment model used in [59]

IV. COMMERCIAL CREDIBILITY ASSESSMENT APPLICATION

There is some research trying to implement the proposed approaches in practical ways. Gupta et al. [16] proposed TweetCred, a real-time web-based system with browser interface to predict credibility. Figure 14 shows the components of TweetCred system. TweetCred takes Twitter posts as input and assigns the credibility level for each tweet from 1 to 7 levels. The model is a semi-supervised model and the training set contains tweets from six high impact topics of 2013. The research trained many classifiers such as: SVM-rank [45], Coordinate Ascent [60], AdaRank [61], and RankBoost [62] and the best results were achieved using AdaRank and Coordinate Ascent.

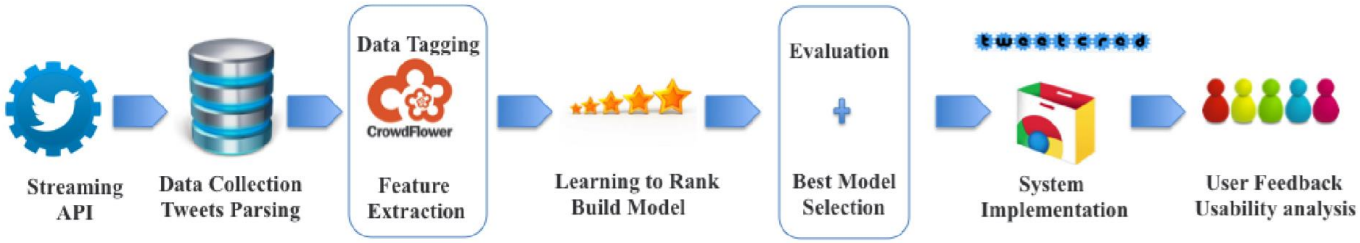

Figure 14: TweetCred system components [16]

The system enables users to have the credibility score within the user interface of Twitter as shown in Figure 15. The system was evaluated with 717 real Twitter users to test credibility and gave reviews for 936 tweets. The evaluation indicates user agreement with the computed credibility score in 43% of the tweets.

Figure 15: TweetCred user interface [16]

Differently from TweetCred, TwitterBot [20] aims at improving such systems due to the extra scoring by the Reconcile system. The classifier behind the TwitterBOT gives better results than TweetCred but the drawback of the system is the amount of time needed for links analysis. AlRubaian et al. presented CredFinder [38] a real-time system to assess credibility through user and content analysis based on their model that was discussed in [19]. CredFinder consists of two parts: 1) a Chrome extension which collects real time tweets from the user timeline page and 2) a web-based backend to analyze the tweets then calculate their credibility score. They used the U.S. presidential election event as a case study, but the system still needs to be tested by more users to prove its effectiveness.

Moreover, ElBallouli et al. [21] built a classification model to predict the credibility of Arabic tweets automatically depending on features extracted directly or computed from the author's profile. The model included 26 content-based features (grouped into sentiment, social, meta, and textual features), 22 user-based features (grouped into network, meta, and timeline features). The research introduced new features such as the presence of a profile picture, and then they performed face detection to useful features from the picture. To extract the sentiment features, the research used ArSenL [63] which is an Arabic sentiment lexicon and MADAMIRA [64], a morphological analysis tool for Arabic text. To train the model, they annotated a data set of 9, 000 tweets. Multiple machine-learning algorithms were trained but the best results were given by Random forest decision tree learning algorithm. The proposed model achieved a Weighted Average F-measure of 75.8% when applying 10-fold cross validation. CAT was compared to TweetCred proposed by Gupta et al. [16]. The experiments indicated that CAT outperformed TweetCred [16] with an improvement of 16.7% in terms of Weighted Average F-measure. Another interesting observation was that approximately 40% of the tweets generated per day were non-credible tweets when grouping tweets by the creation date. This emphasizes the necessity of Twitter credibility assessment models.

V. HUMAN-BASED APPROACHES

Many researches have studied the credibility problem using human-based approaches through user surveys or statistical analysis. In this section, we discussed the different ways to conduct surveys and analyze their results. We reviewed the statistical analysis which aimed at identifying influential credibility features and studying their distributions.

A. User Surveys

Morris et al. [65] presented a survey to investigate the users' perceptions of information credibility. They conducted two experiments to determine the effect of different tweet features like user name, user image and demographics (age, gender, or Twitter experience) on the perceptions of message and author credibility. They found that Twitter users cannot indicate credibility based on content only, they often depend on visual features like message topic, user name, user image and re-tweet. In addition, the results indicated that user demographics did not assist users in determining message or author credibility. Regarding the message topic type, participants were more concerned about credibility of political and emergency topics. Moreover, it was noticed that topically related user names were considered more credible.

Sikdar et al. [66] presented additional indicators of credibility based on re-tweet behavior. They collected two different data sets on the same topic but have different characteristics. They conducted two surveys in which different information about the same tweet are provided to participants to test how credibility judgments across different surveys are comparable. They concluded that user surveys and re-tweet behavior can be extremely noisy, and prediction based on re-tweet behavior may vary greatly from dataset to dataset.

A research by Yang et al. [67] studies the users' perceptions of credibility on two micro-blogging services, Twitter in USA and Weibo in China. The survey outlined the effect of different features (such as author's gender, profile image and name style)

on users' perceptions and how these features interact with culture to affect credibility assessments. The research found key differences between users in the two countries which imply the fact that users' credibility perceptions are culture dependent. The study indicated that Weibo's users in China trust in and rely on microblogs as an information source more than Twitter users and they depend on metadata integration when evaluating micro-blog credibility with high degree. Similarly, a research by AlMansour et al. [68] studied the effect of culture on assessing credibility. The authors stated the credibility perceptions in Arabic countries and how Arab users utilize Twitter.

B. Statistical Approaches

O'Donovan et al. [36] proposed a statistical analysis of features distribution which can aid in the design of credibility assessment models. They focused on how the features are distributed across three different contexts: credible / not-credible messages, information flow through long or short re-tweet chains and dyadic against non-dyadic messages. Dyadic messages represent pair-wise interaction between two users using the "@ mention" or "@reply" tags. The analysis considered a wide set of features grouped into three classes: content-based, social and behavioral features which focus on the dynamics of information flow such as the average number of friends in timeline. Their results indicated that the topics that cover emergency situations presents a considerable increase in the number of features included in the tweets. The most notable result in the analysis of re-tweet chains is the importance of the URL feature in long chains as it occurs in 50% of them, confirming that tweets including URL can be propagated more often than other tweets. In addition, longer tweets that have longer words and characters appear frequently in longer chains. The analysis of context features stated that dyadic messages include more words, more uppercase letters, more negative sentiments and less hashtags than standard tweets.

VI. CONCLUSION

Information credibility on Twitter can be identified through human based approaches or automatic approaches. In this survey, we have summarized state-of-the-art approaches addressing automatic assessment of information credibility on Twitter. Three approaches were discussed: classification based, propagation based and similarity-based approaches. We reviewed and compared previous research on automatic measuring credibility in different languages such as English, Spanish, Turkey and Arabic. Furthermore, some credibility assessment commercial applications were reviewed like TweetCred, CredFinder and CAT. Finally, human based approaches including user surveys and statistical approaches were reviewed and discussed.

References

[1] Flanagin, Andrew J., and Miriam J. Metzger. "Perceptions of Internet information credibility." Journalism & Mass Communication Quarterly 77.3 (2000): 515-540.

[2] Sharifi, Beaux, Mark-Anthony Hutton, and Jugal Kalita. "Summarizing microblogs automatically." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[3] 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.

[4] 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.

[5] Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. "Twitter Under Crisis: Can we trust what we RT?." Proceedings of the first workshop on social media analytics. ACM, 2010.

[6] 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.

[7] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[8] 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.

[9] Allcott, Hunt, and Matthew Gentzkow. Social media and fake news in the 2016 election. No. w23089. National Bureau of Economic Research, 2017.

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

[11] 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).

[12] Imran, Muhammad, et al. "Practical extraction of disaster-relevant information from social media." Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.

[13] Fogg, Brian J. "Prominence-interpretation theory: Explaining how people assess credibility online." CHI'03 extended abstracts on human factors in computing systems. ACM, 2003.

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

[15] 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.

[16] Gupta, Aditi, et al. "Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter." Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain. 2014.

[17] Kang, Byungkyu, John O'Donovan, and Tobias Höllerer. "Modeling topic specific credibility on twitter." Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM, 2012.

[18] AlRubaian, Majed, et al. "A multistage credibility analysis model for microblogs." Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on. IEEE, 2015.

[19] Alrubaian, Majed, et al. "A credibility analysis system for assessing information on twitter." IEEE Transactions on Dependable and Secure Computing (2016).

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

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

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

[23] Alrubaian, Majed, et al. "A Credibility Assessment Model for Online Social Network Content." From Social Data Mining and Analysis to Prediction and Community Detection. Springer International Publishing, 2017. 61-77.

[24] Xia, Xin, et al. "Information credibility on twitter in emergency situation." Intelligence and Security Informatics (2012): 45-59.

[25] Jin, Zhiwei, et al. "News credibility evaluation on microblog with a hierarchical propagation model." Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014.

[26] Ravikumar, Srijith, Raju Balakrishnan, and Subbarao Kambhampati. "Ranking tweets considering trust and relevance." Proceedings of the Ninth International Workshop on Information Integration on the Web. ACM, 2012.

[27] Gupta, Manish, Peixiang Zhao, and Jiawei Han. "Evaluating event credibility on twitter." Proceedings of the 2012 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2012.

[28] Zhao, Liang, et al. "A topic-focused trust model for Twitter." Computer Communications 76 (2016): 1-11.

[29] Seo, Eunsoo, Prasant Mohapatra, and Tarek Abdelzaher. "Identifying rumors and their sources in social networks." SPIE defense, security, and sensing(2012): 83891I-83891I.

[30] Yin, Guangyu, et al. "Autrust: A practical trust measurement for adjacent users in social networks." Cloud and Green Computing (CGC), 2012 Second International Conference on. IEEE, 2012.

[31] 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.

[32] Al-Eidan, Rasha M. BinSultan, Hend S. Al-Khalifa, and AbdulMalik S. Al-Salman. "Measuring the credibility of Arabic text content in Twitter." Digital Information Management (ICDIM), 2010 Fifth International Conference on. IEEE, 2010.

[33] Qazvinian, Vahed, et al. "Rumor has it: Identifying misinformation in microblogs." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[34] Bhattacharya, Sanmitra, et al. "Belief surveillance with twitter." Proceedings of the 4th Annual ACM Web Science Conference. ACM, 2012.

[35] Ikegami, Yukino, et al. "Topic and opinion classification based information credibility analysis on twitter." Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.

[36] O'Donovan, John, et al. "Credibility in context: An analysis of feature distributions in twitter." Privacy, Security, Risk and Trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom). IEEE, 2012.

[37] Howard, Philip N., et al. "Opening closed regimes: what was the role of social media during the Arab Spring?." (2011).

[38] AlRubaian, Majed, et al. "CredFinder: A real-time tweets credibility assessing system." Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016.

[39] Saez-Trumper, Diego. "Fake tweet buster: a webtool to identify users promoting fake news ontwitter." Proceedings of the 25th ACM conference on Hypertext and social media. ACM, 2014.

[40] McKelvey, Karissa Rae, and Filippo Menczer. "Truthy: Enabling the study of online social networks." Proceedings of the 2013 conference on Computer supported cooperative work companion. ACM, 2013.

[41] 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.

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

[43] Mathioudakis, Michael, and Nick Koudas. "Twittermonitor: trend detection over the twitter stream." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010.

[44] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[45] Joachims, Thorsten. "Optimizing search engines using clickthrough data." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[46] Buckley, Chris, Gerard Salton, and James Allan. "Automatic retrieval with locality information using SMART." Proceedings of the First Text REtrieval Conference TREC-1. 1993.

[47] Robertson, Stephen E., et al. "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track." Nist Special Publication SP 500 (1999): 253-264.

[48] Järvelin, Kalervo, and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques." ACM Transactions on Information Systems (TOIS) 20.4 (2002): 422-446.

[49] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM, 2001.

[50] Abdul-Mageed, M., S. Kuebler, and M. Diab. "SAMAR: A system for subjectivity and sentiment analysis of social media Arabic." Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ICC Jeju, Republic of Korea. 2012.

[51] Alrubaian, Majed, et al. "Reputation-based credibility analysis of Twitter social network users." Concurrency and Computation: Practice and Experience29.7 (2017).

[52] Morozov, Evgeny, and Mourjo Sen. "Analysing the Twitter social graph: whom can we trust." (2014).

[53] Al-Qurishi, Muhammad, et al. "A new model for classifying social media users according to their behaviors." Web Applications and Networking (WSWAN), 2015 2nd World Symposium on. IEEE, 2015.

[54] Bravo-Marquez, Felipe, Marcelo Mendoza, and Barbara Poblete. "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis." Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM, 2013.

[55] Karagöz, Ali Fatih Gündüz Pınar. "Credibility Analysis for Tweets Written in Turkish by a Hybrid Method." Feature Engineering in Hybrid Recommender Systems: 55.

[56] Breiman, L. "Random forests. Mach Learn45: 5–32." (2001).

[57] Balakrishnan, Raju, and Subbarao Kambhampati. "SourceRank: relevance and trust assessment for deep web sources based on inter-source agreement." Proceedings of the 20th international conference on World wide web. ACM, 2011.

[58] Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. "A comparison of string metrics for matching names and records." Kdd workshop on data cleaning and object consolidation. Vol. 3. 2003.

[59] Sabbeh, Sahar f., and Sumaia Y. Baatwah. "Arabic news credibility on twitter: an enhanced model using hybrid features." Journal of Theoretical & Applied Information Technology 96.8 (2018).

[60] Metzler, Donald, and W. Bruce Croft. "Linear feature-based models for information retrieval." Information Retrieval 10.3 (2007): 257-274.

[61] Xu, Jun, and Hang Li. "Adarank: a boosting algorithm for information retrieval." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[62] Freund, Yoav, et al. "An efficient boosting algorithm for combining preferences." Journal of machine learning research 4.Nov (2003): 933-969.

[63] Badaro, Gilbert, et al. "A large scale Arabic sentiment lexicon for Arabic opinion mining." ANLP 2014 165 (2014).

[64] Pasha, Arfath, et al. "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic." LREC. Vol. 14. 2014.

[65] Morris, Meredith Ringel, et al. "Tweeting is believing?: understanding microblog credibility perceptions." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012.

[66] Sikdar, Sujoy, et al. "Understanding information credibility on twitter." Social computing (socialcom), 2013 international conference on. IEEE, 2013.

[67] Yang, Jiang, et al. "Microblog credibility perceptions: comparing the USA and China." Proceedings of the 2013 conference on Computer supported cooperative work. ACM, 2013.

[68] AlMansour, Amal Abdullah. "Credibility Perception for Arab Users." Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2016.

[69] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

++**********

O'Donovan et al identified the most useful indicators of credible and noncredible tweets as URLs, mentions, retweets, and tweet lengths[19 lstm]. (credibility in context)