FCI-H
Informatics Bulletin

Helwan University

User-Generated Content (UGC) Credibility on Social Media Using Sentiment Classification

Esraa A. Afify [a], Ahmed Sharaf Eldin [a], Ayman E. Khedr [b], Fahad Kamal Alsheref [c]

[a] *Faculty of Computers & Information, Information Systems Department, Helwan University, Cairo, Egypt*
[b] *Faculty of Computers & Information Technology, Information Systems Department, Future University in Egypt, Cairo, Egypt*
[c] *Faculty of Computers & Information, Information Systems Department, Beni-Suef University, Cairo, Egypt*

KEYWORDS

Social Media,
Credibility on Social Media,
Sentiment Classification,
Machine Learning Techniques,
Performance Evaluation.

ABSTRACT

Web 2.0 technologies have seen a big evolution recently leading to the existence of a huge amount of unreliable and misleading content due to the openness and low publishing barrier nature of the content generated through social media platforms. As a fact, the User-Generated Content (UGC) on social media platforms suffers from a lack of professional gatekeepers to monitor this content. Consequently, most online users fall into the trap of being misled through fake information that spreads rapidly. They usually rely on this information without any verification and this prevents them from making accurate decisions concerning their social lives, politics, or business events. Because online users face difficulty in finding which piece of information is credible or not, the researchers found that assessing User-Generated Content (UGC) of social media is very important in resolving the issue of credibility. This paper adapted some of the existing literature and concluded that many previous approaches have investigated information credibility on Twitter and a limited number of Facebook for proposing a new approach for measuring posts credibility. The proposed model used to measure the credibility of Facebook posts through a formula combined from the page profile rank and the post-analysis score. The model was tested and achieved 87.45 % accuracy.

1. Introduction

Over the last ten years, there has been a noticeable change in the way online users pursue social lives. Sharing information about themselves, posting pictures and sending messages to friends about upcoming events, finding and interacting with people they never imagined talking with, and connecting with family members, classmates, friends and colleagues. Sharing knowledge, experiences, thoughts, and opinions with the world can be achieved by means of read-write Web and user-generated contents. With the widespread use of unmonitored web 2.0 content, many people may be negatively influenced, and consequently affect their ability to make sound social, political and economic decisions. This change felt by us is not only due to the exponential evolution of technology but also due to the existence of *"Social Media"*.

Nowadays, Social Media users can easily communicate and publish whatever they want. In this context, users can create and directly publish large quantities of any content: reviews or opinions, news, links, pictures, or videos extensively almost without any form of external reliable sources or trusted control. Due to the unlimited expression of user-generated opinions, there is a need to analyze them to implement better decision choices.

Paper organization. This paper is organized as follows: Section 2 presents the research problem and its relevance. Section 3 provides a background related to the research study. Section 4 explains the approaches provided for the credibility assessment by illustrating a survey of the existing works of literature concerning the research study problem. Section 5 illustrates a novel approach proposed for assessing credibility on Facebook content. Section 6 discusses the analysis results of the research study. Section 7 explores the research contribution. Finally, section 8 contains the conclusion conducted in the research study.

* *Corresponding authors.*
E-mail address: esraa_afify@hotmail.com, profase2000@yahoo.com, ayman.khedr@fue.edu.eg, drfahad@fcis.bsu.edu.eg

2. Research Problem and Relevance

People use Social Media platforms daily to obtain news and information first-hand. They also use them for communicating or sharing newsworthy information. They employ Social Media for practically every aspect of their lives. They utilize Social Media in disasters to report injuries, damage, or aids needed.

Due to open discussions on social media information is generated and shared at a high rate. Online users face numerous challenges such as content anonymity, the absence of standards for information quality, ease of manipulating and altering information, lack of clarity of context, and presence of many potential targets of credibility evaluation (i.e., the content, the source, and the medium.) This creates an urgent need to develop systems for serving users to automatically assess the credibility of information and information sources.

In the following subsections the researchers will discuss the fake news on social media, then credibility and its relationship with social media that relate to the research problem.

2.1 Fake News on Social Media

Fake news were founded for a long time, that intended to verify false news and mislead readers. Social media has demonstrated to be a strong source for fake news dissemination. A survey of some existing fake news detection on social media is still in the early stage of development, and there are still many challenging problems that need further investigations. Fig. 1. identifies the two aspects of fake news problem: *characterization* and *detection*.

2.1.1 Fake News Characterization

Malicious Accounts on Social Media for Propaganda [40]. Social media users could be malicious, and in many cases were not even real humans. The cheep of creating social media accounts motivated malicious user accounts, such as social bots, trolls, and cyborg users.

- *A social bot*- refers to the accounts that a computer algorithm controls to automatically produce content and interact with humans (or other bot users) on social media.
- *Tolls*- real human users whom the target is to disrupt online communities and excite people's inner negative emotions, such as anger and fear, that's rolling in spreading the fake news on social media.
- *Cyborg users*- propagate fake news that mix automated activities with human input. Commonly, cyborg accounts are registered by a human as a camouflage and set automated programs to perform activities in social media. The simple turn of functionalities between bot and human offers cyborg users' unique opportunities to spread fake news.

Echo Chamber Effect [40]. Social media supplies a new model of information innovation and consumption for users. Consumers are selectively exposed to certain kinds of news because of the way news feed appears on their homepage in social media, amplifying the psychological challenges of dispelling fake news. Usually, users on social media head for composing groups that have a like-minded people where they can gather their opinions, because of an echo chamber effect. For example, users on Facebook always follow like-minded people and thus receive news that promotes their favored existing narratives.

The echo chamber effect facilitates the way people consume and believe fake news due to the following psychological factors.

- *Social credibility*, which means people are more likely to take a source as credible if others perceive that, mostly when there is insufficient information available to prove the truthfulness of the source.
- *Frequency heuristic*, which means that users could naturally favor information heard frequently, even if it is fake news.

Researches had shown that increased exposure to an idea is enough to generate a positive opinion of it, and in echo chambers, users continue to share and consume the same information. As an outcome, the echo chamber effect created segmented, homogeneous communities with a very limited information ecosystem. Also, researches had shown that the homogeneous communities became the major driver of information propagation that supports strengthens polarization.

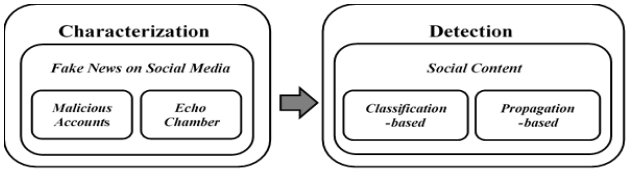

Figure 1: Fake News on Social Media: From Characterization to Detection [40].

2.1.2 Fake News Detection

Fake news detection structure includes two stages: *feature extraction*, and *model construction*. The feature extraction stage aims to clarify news content and related auxiliary information in a formal mathematical structure, and model construction stage otherwise constructs machine learning models that improve how to identify fake news and genuine news based on feature representations.

(i) Feature Extraction

In this stage, the fake news detection has two ways. The first one, on traditional news media, which depends on news content. While the second one, in social media depends on additional social context auxiliary information that used to help the detection of fake news. In the subsections, the researchers will briefly present how to extract and represent useful features from *news content* and *social content [40]*.

A. News Content Features

News content features depict the information regarding a part of the news. Below the news content attributes are shown [40]:

- **Source:** Author or publisher of the news article.
- **Headline:** Short title text that aims to attract the attention of readers and depicts the prime topic of the article.
- **Body Text:** Major text that makes the details of the news story; there is usually the main demand that is specifically highlighted and that forms the opinion of the publisher.
- **Image/Video:** Portion of the body content of a news article that expands visual cues to reach the story.

Based on these content attributes, the news content consists of *linguistic-based* and *visual-based*, described in more details below [40].

Linguistic-based: Fake news is purposely designed for monetary or governmental earning rather than to state impartial demands, includes opinionated and inflammatory language, crafted as "clickbait" (i.e., to seduce users to click on the link to read the entire news article) or to promote ambiguity. Therefore, it is rational to employ linguistic features that grab the different writing styles and rousing headlines to detect fake news.

Linguistic-based features are taken from the text-content idioms of document communities from various levels, such as characters, words, sentences, and documents. To gather the various sides of fake news and genuine news, researches have applied together *common linguistic features* and *domain-specific linguistic features*.

- *Common linguistic features*, which are applied to perform documents for various tasks in natural language processing. Typical common linguistic features are:
  - *Lexical features* include features for both "character-level" and "word-level", (e.g., "characters-per-word", "frequency-of-large-words", "unique-words" and "total-words";
  - *Syntactic features* include "sentence-level" features, such as "frequency of function words" and "phrases" (i.e., "n-grams" and "bag-of-words" approaches) or "punctuation" and "parts-of-speech" (POS) tagging.
- *Domain-specific linguistic features*, which are ranged for news domain, such as "quoted words", "external links", "number of graphs", and "the average length of graphs", etc. Furthermore, different features can be sited to capture the misleading cues in writing styles to differentiate fake news, such as false detection features.

Visual-based: A critical manipulator for fake news propaganda. That utilizes the sensitivity of people and therefore depends on exciting or even fake images to motivate anger or other sentimental response of consumers. Visual-based features are taken from visual elements (e.g. pictures and videos) to grasp the different characteristics of fake news. Unreal images are given based on several user-level and tweet-level hand-crafted features using the classification structure. Lately, several visual and statistical features have been extracted for news verification. Visual features cover clarity score, coherence score, similarity distribution histogram, diversity score, and the clustering score. Statistical features involve count, image ratio, multi-image ratio, hot image ratio, long image ratio, etc.

B. Social Content Features

Social context features are applied to the user-driven social engagements of news consumption on the social media platform. Social engagements clarify the news proliferation process over time, which gives useful auxiliary information to conclude the validity of news articles. Generally, there are three major aspects of the social media context: *users*, *generated posts*, and *networks [40]*.

User-based: As mentioned before, fake news pieces are likely to be created and spread by non-human accounts, such as social bots or cyborgs. Consequently, gathering users' profiles and characteristics by user-based features can supply helpful information for fake news detection. User-based features provide the characteristics of users who have interactions with the news on social media. These features enable categorization across various levels: *individual level* and *group level*.

- *Individual-level features* are applied to measure each of user reliability and credibility form several sides of user demographics, (e.g., the age of user registration, the number of user followers/followees, the number of tweets the user had created, etc.).
- *Group-level user features* take the characteristics of users' groups regarding the news. The claim is that the spreaders of fake news and genuine news can form various groups with unique characteristics that may be drawn by group level features. Usually applied group level features brings from assembling (e.g., averaging and weighting) individual level features, such as "percentage of verified users" and "average number of followers".

Post-based: users show their sentiment or opinions to fake news through social media posts, such as skeptical opinions, sensational reactions, etc. Therefore, it is rational to evolve post-based features to assist the chance of detecting fake news via responses from the people as expressed in posts. Post-based features concentrate on resembling useful information that concludes the validity of news from several sides of relevant social media posts. These features may be categorized as *post level*, *group level*, and *temporal level*.

- *Post level features* produce for each post feature values. Also, for each post, the linguistic-based features and several establishing approaches for news content can be performed. Especially, the unique features of posts that appear, people, social reactions, such as *stance*, *topic*, and *credibility.*
  - *Stance features* (or viewpoints) show the users' opinions across the news, such as supporting, denying, etc.
  - *Topic features* taken from using topic models, such as latent Dirichlet allocation (LDA).
  - *Credibility features* for posts estimate the reliability degree.
- *Group level features* intend to gather the feature values for all related posts for specific news articles by using "wisdom of crowds". For example, the average credibility scores are utilized to estimate the credibility of news.
- *Temporal level features* treat the temporal difference of post level feature values. Unsupervised embedding approaches, such as the recurrent neural network (RNN), are used to grasp the changes in posts over time. Based on the way of this time series for several metrics of related posts (e.g., number of posts), mathematical features may be computed, such as SpikeM parameters.

Network-based: Users compose various networks on social media in terms of interests, topics, and relations. As mentioned previously, fake news dissemination methods formed to make an echo chamber cycle, highlighting the value of extracting network-based features to perform these types of network patterns for fake news detection. Network-based features are founded by building certain networks among the users who publish linked social media posts, thus, various types of networks can be constructed.

- *The stance network*, which constructs with nodes showing all the tweets related to the news and the edge showing the weights of similarity of stances.
- *The co-occurrence network*, which construction formed on the user engagements by counting if those users compose posts related to the same news articles.
- *The friendship network*, which shows the structure of the following/followee of users who post relative tweets. An expansion of this friendship network is the diffusion network, that tracks the path of propagation of the news, where nodes clarify the users and edges illustrate the information diffusion paths among them.

(ii) Social Context Models Construction

Social context models consist of relevant user social engagements in the analysis, gathering this conducive information from a diversity of viewpoints. The current methods for social context modeling are classified into two groups: *Stance-based* and *Propagation-based [40]*.

Stance-based: (classification-based) approach used to conclude the validity of genuine news articles that relevant to users' opinions from related post contents. The stance of users' posts may be clarified either explicitly or implicitly. Explicit stances are direct terms of sentiment or opinion, such as the "thumbs up" and "thumbs down" responses came on Facebook. Implicit stances may be automatically taken from social media posts. Stance detection is an automatic functionality used to define if the user is in favor of, neutral toward, or against some target entity, event, or idea.

Propagation-based: approach used for predicting the news credibility of the fake news about the interrelations of relevant social media posts. The main claim is that the credibility of a news incident is highly concerned with the credibility of relevant social media posts. The propagation method consists of two credibility networks *homogeneous* and *heterogeneous*.

- *Homogeneous credibility networks* based on one type of entities, such as post or event.
- *Heterogeneous credibility networks* include various types of entities, such as posts, sub-events, and events.

Lately, the conflicting of opinion relationships is founded to construct a homogeneous credibility network via tweets and lead the operation of assessing their credibility.

2.1.3 Evaluation Metrics

To assess the performance of the algorithms for fake news detection issue, several appraisal metrics have been utilized. In the following subsection, a survey of the vastly applied metrics for fake news detection are adopted. Most current approaches treat the fake news issue as a classification problem that predicts whether a news article is fake or not:

- True Positive (**TP**): when predicted fake news pieces are annotated as fake news;
- True Negative (**TN**): when predicted true news pieces are annotated as true news;
- False Negative (**FN**): when predicted true news pieces are annotated as fake news;
- False Positive (**FP**): when predicted fake news pieces are annotated as true news.

By formulating this as a classification problem, the following metrics could be defined [40],

$$Precision = \frac{|TP|}{|TP| + |FP|} \quad (1)$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \quad (2)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (4)$$

These metrics are usually applied in the machine learning group and used to estimate the performance of a classifier from several perspectives. i.e. this involves measuring the accuracy of similarity between predicted fake news and genuine fake news. *Precision* measures all the fake news fraction detected that is commented as fake news, classifying the important problem of recognizing which news is fake. However, datasets of fake news are often skewed, a high *precision* may be easily accomplished by doing fewer positive predictions. Consequently, the *recall* applies to degree the sensitivity, or the fraction of commenting fake news articles that are predicted to be fake news. *F1* is utilized to merge with *precision* and *recall*, which can supply a total prediction performance for fake news detection. Note that for *Precision*; *Recall*, *F1*, and *Accuracy*, the higher the value, the better the performance.

2.2 Credibility on Social Media

PROBLEM DEFINITION: Previous studies have found that social media interaction is a type of unstructured data. Millions of interactions come from a wide range of sources. According to the Pew Research Center [41], the “top” Social Media platforms are *Facebook*, *YouTube*, *Twitter*, *Instagram*, *Snapchat*, *LinkedIn*, *Pinterest* and *WhatsApp.*

In the following subsections, the researchers will give a brief overview of the concept of *Credibility*, *Social Media*, and the relation between both relating to the idea of the research study in this paper.

2.2.1 Credibility

DEFINITION 1. Credibility is *"a multifaceted concept and has been defined as reliability, accuracy, fairness, and objectivity, as well as various combinations of these concepts"* [24]. Across definitions, credibility has often been correlated with “*trustworthiness*” and “*believability*”. It means a quality recognized by individuals, who are not permanently able to identify with their cognitive ability the genuine information from the fake.

Credibility has been discussed in the three perspectives of communication: *medium credibility*, *message/content credibility*, and *source credibility*. ***Medium credibility*** is the apparent level of credibility that users have of a certain medium, such as newspapers, TV, the Internet, or blogs; likewise, the medium's owners take responsibility for the content. ***Message credibility*** is the known trust of the linked message itself, such as informational quality, reliability, or prevalence. Past research on ***Source credibility*** has focused on the expertise or the trustworthiness of the source as the likelihood to provide credible information. But, on the issue concerning social media, the source may be unknown, so no one takes responsibility for the content. In many situations, a username found to be the only information known about the source (e.g., an insufficient or a fake profile on *Twitter* or *Facebook* that releases information about an incident).

Inspired by work of Hilligoss & Rieh [24] and Fogg [16], the three main components affecting UGC Information Credibility perception are introduced as follows: (1) Contexts such as, environment, topic and situation, (2) UGC available features (cues), and (3) evaluator traits and cognitive heuristics such as, topic knowledge and the selection of cues in making a credibility judgment.

2.2.2 Social Media

DEFINITION 2. Social Media is defined as "*a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of User-Generated Content*" [28]. In this paper, the researchers will focus on the two most popular user-generated content social networks in the world with the highest number of active users.

DEFINITION 3. *Twitter*- is *"a micro-blogging service that allows its users to post and exchange 140-character-long messages known as ‘tweets’ "*. It averages 100 million users per day sending 500 million tweets, amounting to 328 million monthly active users with 80% of these users socializing on mobile [45].

DEFINITION 4. *Facebook*- is *"a multi-purpose social networking platform that promotes and facilitates interaction between friends, family, and colleagues"*. It has over 2 billion monthly active users, 1.66 billion monthly mobile active users and 1.32 billion desktop daily active users. It is considered the largest leading source of news exceeding all other social media sites and targeting all age groups [15].

2.2.3 Credibility and Social Media

Credibility is an extremely wide term and can include various factors for diverse applications. In this research, the focus is on the credibility of information spread through social media networks. Recently, the credibility of information has become a major concern to online users. Therefore, various approaches have been proposed to automatically assess information credibility on social media platforms.

There are various factors that can be found on the social media platform itself, and that are helpful in assessing the information credibility. These factors include:

- The responses that specific topics generate, and the sentiment announced by users discussing the topic: e.g. if they use opinion expressions that clarify positive or negative sentiments about the topic;
- The scale of certainty of users spreading the information: e.g. if they question the information that is offered to them, or not;
- The cited external sources: e.g. if they cite a certain URL with the information they are spreading, and if that source is a known domain or not;
- The users' characteristics that spread the information: e.g. the followers' number that each user has on the platform.

There are four types of credibility features that social media depends on *message-based features*, *user-based features*, *topic-based features*, and *propagation-based features [11]*.

- **Message-based features** examine messages characteristics, these features can be independent or dependent. Independent features involve a message length, whether or not the text contains exclamation or question marks and the number of positive/negative sentiment words in a message. Dependent features involve: like, comment, share, a hashtag, and a re-tweet.
- **User-based features** examine the users' characteristics for the post messages, such as registration age of the users, followers' numbers, followees numbers, and user tweets number created in the past.
- **Topic-based features** computed from the message and user-based features; e.g., the tweets fraction include URLs, the tweets fraction with hashtags and the sentiment fraction of positive and negative in a set.
- **Propagation-based features** examine characteristics associated with the propagation network. These involve features such as the re-tweet depth, or the initial tweets number for a certain topic.

2.2.3.1 ***Credibility on Twitter***- Tweets can be released through sending e-mails, SMS text messages and instantly from smartphones. Therefore, it spreads a real-time propagation of information to large groups of online users. Tweet feeds includes many hints that refer to the credibility, distinguishing the trustworthy news automatically from non-credible ones is obviously a challenging job. Mostly, there is a lack of any genuine or instant verification technique for fresh news events. This makes it an ideal environment for the dissemination of newsworthy content directly from the news source and/or geographical location of events. Micro-blogging is the source of the extremely important and effective events, through emergency and important events. Therefore, it becomes significant to develop tools that prove the credibility of online generated content.

The main kinds of incredible events on Twitter, as follows:

a. *Clear incredible*, such as fake events regarding for celebrities or strategic locations, partly spiced up rumors or erroneous claims done by politicians.
b. *Seem incredible*, such as informally written tweets, tweets making conflicting claims, tweets lacking any supportive evidence like URLs, tweets without any credibility conveying words like news, breaking, latest, report, etc.

To perform an automatic credibility assessment on Twitter, some of the important hints concerning the attributes of the entities must be clarified, as follows [21]:

User Features: User credibility can be affected by the social reputation and profile completeness, measured using the following factors.

1) The number of friends, followers, and status updates.
2) User Twitter profile connected to user Facebook profile.
3) Verified user account by Twitter.
4) Age of user account on Twitter.
5) User profile description, URL, profile image, location.

Tweet Features: Tweet credibility can be affected by the following factors.

1) Professionally tweet writing (as no slang words, '?', '!', smileys, etc.).
2) External URLs included supportive evidence.
3) Words number for first, second, third person pronouns.
4) Frequent location concerning the event, the user who create many tweets related to the event, the most frequently cited URL related to the event, the most frequently used hashtag related to the event.
5) Complete details, as most entities related to the event.
6) Tweet sentiment matches with a sentiment of the event.

Event Features: Event credibility can be affected by the following factors.

1) The event number of tweets and retweets.
2) The event number of distinct URLs, domains, hashtags, user mentions, users, location.
3) The event number of hours to be popular.
4) The event tweets percentage concerning the day when the event reached its peak popularity.

2.2.3.2 Credibility on Facebook- As a fact, any post expands enormous engagement in the first five hours; this post could gain 75 % of its existence effect after around two and a half hours. Every month, Facebook users post 2.5 billion comments on Facebook pages. So, it is questionable which of the posts or comments on these pages are considered credible or not.

To perform features measuring for information credibility on Facebook, the Credibility computation can be classified into two categories, as follows [37]:

Web-page-independent features: This approach does not use any feature for measurement credibility but compares content with trusted news sources by using Natural language processing (NLP). It enables a computer to understand human languages such as comparing two documents similarly. However, in practices, this approach is unsuitable for social media data, which has not only text but also images and videos. Messages in social media are used for computing credibility by comparing the messages with trusted news sources. If the message is similar with trusted news sources, the credibility score of the message is high.

Web-page dependent features: Use features of each social media for computing credibility such as like, comment, and share.

3. Background of Research Study

From previous research studies, information provided to online users about the estimated credibility of generated content was very useful and valuable to researchers. In this context, the approach that is based on data-driven models uses sentiment classification and machine learning techniques that classify content and/or sources of information as credible or not credible. For that, the researchers found the importance to declare and discuss the concepts and techniques used for classification and the approaches related to the research study.

In this section, the researchers will first provide a brief background on the concept of Sentiment Analysis and Opinion Mining; then try to illustrate the most state-of-the-art approaches that discuss information Credibility on social media sites.

Nowadays, Sentiment Analysis is used as a popular study, because social networking sites contain online users who are free to show their ideas, feelings, and impressions about a specific topic. The massive amount of ready-made data pulls system developers for studying automatic mining and analysis. How users think about specific topics may be a classification task. Online users' feelings could be classified as positive, negative or neutral. A sentiment is often represented in simple or complex ways in a text. Online users may combine objective and subjective information about a certain topic. On top of that, data collected from the World Wide Web often includes many imperfections.

SENTIMENT ANALYSIS AND OPINION MINING: DEFINITION

Sentiment Analysis (*SA*) or *Opinion Mining* (*OM*), is a new field of research coined in Natural Language Processing (NLP), aiming at identifying the sentiment (positive or negative or neutral), detecting subjectivity (objective or subjective) in text and/or extracting and classifying opinions and sentiments [4].

Sentiment analysis and opinion mining is the field of study that analyzes people's sentiments, opinions, attitudes, evaluations, appraisals, and emotions towards services, products, individuals, organizations, issues, topics, events and their attributes [30] [36].

Sentiments can be described as emotions, judgments or ideas prompted or colored by emotion. Computationally, the concentration is on opinions rather than on sentiments, feelings or emotions, but the word 'sentiment' and 'opinion' are often utilized interchangeably. Understanding *Social Media* sentiments can help get a grasp of users' knowledge and capture their ideas without necessarily going through the entire data, which will save a huge amount of time in the analysis.

SENTIMENT CLASSIFICATION: LEVELS

The Sentiment Classification (SC) is a job of grouping a base unit in a document to the positive or negative class. There are three essential classification levels [32]:

- **Document-level:** classifies an opinion document as a positive or negative opinion or sentiment. It regards the entire document as a basic data unit (discussing one topic.)
- **Sentence-level:** classifies sentiment explicit in each sentence. On the off chance that the sentence is subjective, it groups it into positive or negative opinions.
- **Aspect-level:** classifies the sentiment concerning aspects of entities. Users can assign different opinions to different aspects of the same entity.

SENTIMENT ANALYSIS: PROCESS

Sentiment Analysis function is treated as a sentiment classification (SC) problem. The first stage in the SC problem is to extract and select text content features.

Sentiment Analysis usually contains three main stages [9]:

Pre-processing, feature selection, and sentiment classification.

A. PRE-PROCESSING

Text documents contain rich textual information, for example, words and phrases, punctuation, abbreviation, emoticons and so on. They also tend to have incorrect spelling, duplicate-characters (for example, "cooool"), particularly for social media generated content. Direct application of SA methods on such text usually leads to poor performance. Therefore, pre-processing is typically conducted to convert the text into textual features that could fit into the SA methods. Once the pre-processed text features are extracted, they are ready to be fitted in the next phase of SA – Feature Selection [10] [22].

Pre-processing is usually based on NLP techniques, such as:

- tokenization (splitting the sentences into words),
- de-noising (removing special characters, capturing symbols for emotions),
- normalization (removing duplicate characters, identifying root words etc.),
- stop-words removal (removing stop words and the words which are of no use to sentiment analysis),
- stemming (returning the word to its stem or root),
- lemmatization (converting inflected words to their root form) etc.

B. FEATURE SELECTION

The output of pre-processing is the extracted text features. Many text features are applied for SA, as following [2] [33]:

- **Term presence:** unigram (individual words), bigram (two consecutive words), or n-grams (n consecutive words.)
- **Term frequency:** To assign the word binary weighting (one, in case the word appears or zero, otherwise) or employing term frequency weights to show the relative importance of features.
- **Opinion words and phrases:** Words commonly used to express opinions such as the words *'good'* or '*like'*, '*bad'* or '*hate'*. Phrases commonly used to express intensification of opinions or negative words that change the opinion orientation such as, '*cost me an arm and a leg'*.
- **POS (Part-of-Speech) Tags:** To identify adjectives and adverbs which are usually used as sentiment indicators.
- **Negations:** Very common linguistic constructions that affect polarity, therefore, the appearance of words of negation may change the opinion orientation such as, '*not good'* is equivalent '*to bad'*.
- **Syntactic Dependency:** Word dependency is generated from parsing.

The goal of feature selection is to select important text features out of the immense number of extracted ones. Feature selection methods can be categorized into filter methods and wrapper methods. Filter methods rank the features according to certain metrics and select the top-ranked features. Wrapper methods, on the contrary, select the best subset of features by generation and evaluation of different subsets with a classifier. Therefore, the selected features tend to be classifier specific, namely they might perform well using the specific classifier that is used for the selection but not necessarily well with other classifiers.

FEATURE SELECTION METHODS:

They are divided into [32]:

- **Lexicon-based methods:** needs human annotation. They usually begin with a small set of 'seed' words. Then they bootstrap this set through synonym detection or online resources to obtain a larger lexicon.
- **Statistical methods:** fully automatic and more frequently used.

The feature selection techniques considered to be documented either as a group of words (Bag-of-Words (BOWs)) or as a string which saves the sequence of words in the document. BOW is mostly applied because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word to its stem or root i.e. flies →fly.)

C. *SENTIMENT CLASSIFICATION*

Sentiment Classification techniques consist of [31] [32]:

- **Machine Learning Approach (ML):** uses the most known ML algorithms and applies linguistic features.
- **Lexicon-based Approach:** relies on a sentiment lexicon, a familiar group, and precompiled sentiment expressions.
- **Hybrid Approach:** merges the two methods and runs a key part in most of the methods with sentiment lexicons.

The following section contains a brief explanation of two algorithms:

A. Machine Learning approach

Machine learning depends on the most known ML algorithms to resolve the SA as an orderly text classification problem that uses the syntactic and/or linguistic features [32].

Fig. 2. illustrates the famous SC techniques and the most popular algorithms used in these fields.

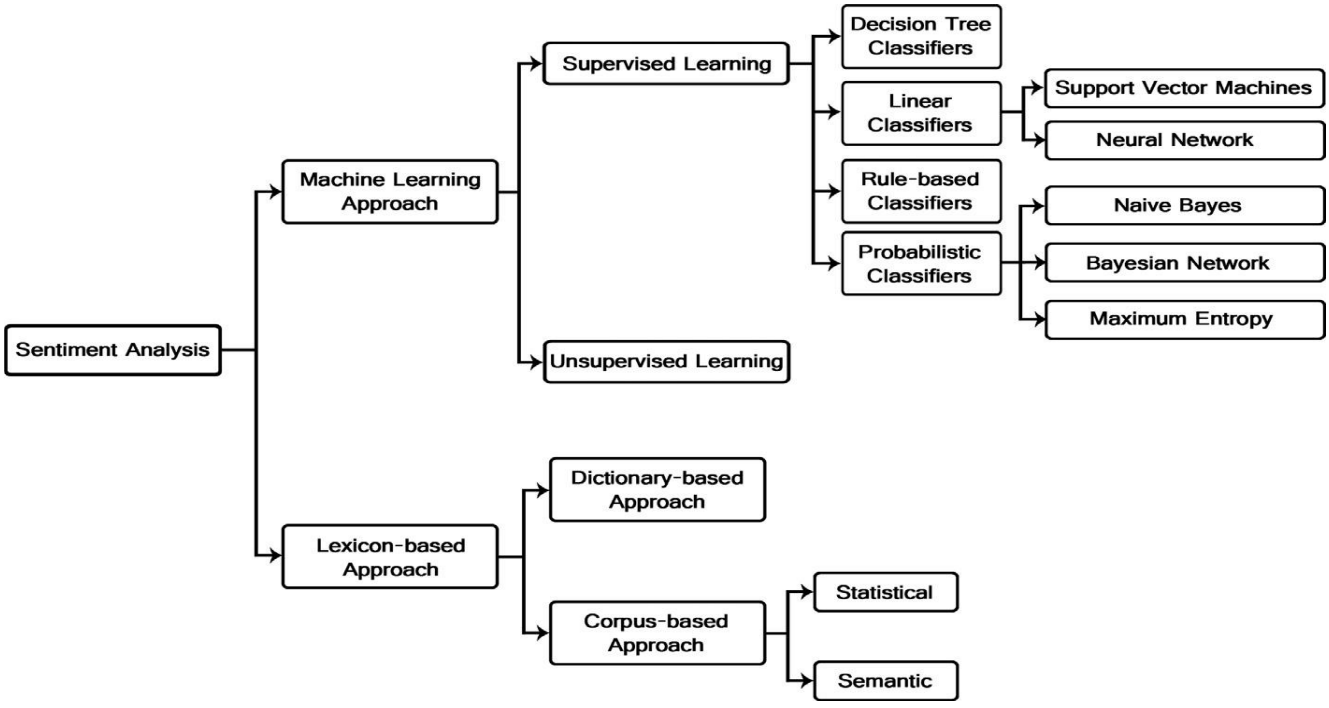

Figure 2: Sentiment classification techniques [32].

Text Classification Problem Definition:

Consider a set of training records $D = \{X_1, X_2, \ldots, X_n\}$ where each record is labeled according to a class. The classification model is related to the features in the underlying record to one of the class labels. If a class is unknown, the model is used to predict a class label for it.

There are two methods in this approach, as presented in the following subsections [32] [46]:

1) *Supervised learning*: relies on the existence of labeled training documents;

1.1. Decision tree Classifiers

The decision tree classifier applies a hierarchical decomposition that uses the training data space for performing the division on the data as a condition of the attribute value. The condition or predicate is the presence or absence of one or more words. The classification utilized as a division on the data space that done recursively until the lead nodes include specific minimum numbers of records.

There are several kinds of splits in decision trees, as follows below:

- *Single-Attribute split*, this split is performed using the presence or absence of words or phrases at a node in the tree.
- *Similarity-based multi-attribute split*, this split is performed using documents or frequent word clusters and the similarity of the documents to these word clusters.
- *Discriminate-based multi-attribute split*, this split is performed using discriminants such as the Fisher discriminate.

The decision tree implementations in text classification is used to be small differences on standard packages such as ID3 and C4.5.

1.2. *Linear Classifiers*

It is the normalized document word frequency as given $\bar{X} = \{x_1 \ldots\ldots x_n\}$, a vector of linear coefficients with the same dimensionality of the feature space as vector $\bar{A} = \{a_1 \ldots\ldots a_n\}$, and a scalar as b; $p = \overline{A} \cdot \overline{X} + b$ defined as the output of the linear predictor, which is the result of the linear classifier. The predictor p is a separating hyperplane between various classes [32].

Two of the most known linear classifiers are explained in the following subsections:

1.2.1. *Support Vector Machines (SVM)*

The essential principle of SVMs is to define linear separators in the search space that can best separate the various classes. SVMs are applied in many applications, among these applications are classifying reviews according to their quality.

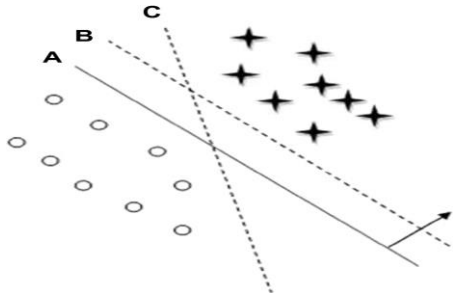

Figure 3: Using support vector machine on a classification problem [32].

In Fig. 3., there are two classes x, o and there are three hyperplanes, A, B, and C. Hyperplane A gives the best separation between the classes because the normal distance of any of the data points is the largest, so it clarifies the maximum margin of separation.

1.2.2. *Neural Network (NN)*

Neural Network contains many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector *overline*X_i which is the word frequencies in the *i*th document. There is a set of weights A which is associated with each neuron applied in order to compute a function of its inputs f(●). The linear function of the neural network is: $p_i = A \cdot \overline{X}_i$. In a binary classification problem, it is supposed that the class label of $\overline{X}_i$ is denoted by y_i and the sign of the predicted function p_i yields the class label.

1.3. *Rule-based Classifiers*

In rule-based classifiers, the data space is modeled with a set of rules. The left-hand side illustrates a condition on the feature set expressed in disjunctive normal form while the right-hand side is the class label. The conditions are in the term presence. Term absence is rarely applied because it is not informative in sparse data [32].

The most two popular criteria in rule-based, as follows below:

- *The support* refers to the absolute number of instances in the training data set that are relevant to the rule.
- *The Confidence* refers to the conditional probability that the right-hand side of the rule is satisfied if the left-hand side is satisfied.

1.4. *Probabilistic Classifiers*

Probabilistic classifiers employ mixture models for classification. The mixture model supposes that each class is a component of the mixture. Each mixture component is a generative model that gives the probability of sampling a term for that component [32].

1.4.1. *Naïve Bayes Classifier (NB)*

The Naïve Bayes classifier is the simplest and most popular classifier that used. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction to ignore the position of the word in the document. It applies Bayes Theorem to predict the probability if a given feature set belongs to a particular label.

$$P(label \backslash features) = \frac{P(label) * P(features \backslash label)}{P(features)} \quad (5)$$

P(label) is the prior probability of a label or the likelihood that a random feature set on the label. P(features\label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set occurs. Given the Naïve assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label \backslash features) = \frac{P(label) * P(f1 \backslash label) * \ldots. * P(fn \backslash label)}{P(features)} \quad (6)$$

1.4.2. *Bayesian Network (BN)*

The prime assumption of the BN classifier is the independence of the features. The other extreme assumption is to suppose that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes show random variables, and edges illustrate conditional dependencies. BN is seen as a complete model for the variables and their relationships.

1.4.3. *Maximum Entropy (ME)*

The Maxent Classifier (called a conditional exponential classifier), which transforms the sets of feature labeled into vectors by employing encoding. This encoded vector is after that used to calculate weights for each feature that can then be combined to define the most probable label for a set of feature. This classifier is parameterized by the X{*weights*} set, which is utilized to integrate the joint features that are created from a feature-set by an X{*encoding*}. Especially, pair with a vector C{(*featureset, label*)} for each encoding maps. The probability of each label can then calculated using the following equation:

$$P(fs \backslash label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l)) for l in labels)} \quad (7)$$

2) *Weakly, semi and unsupervised learning*: employs a set of inputs, such as clustering when it is hard to locate the labeled training of documents [32].

The major aim of text classification is to classify documents into a certain number of predefined groups. In order to achieve that, a large number of labeled training documents are utilized for supervised learning. In text classification, it is sometimes severe to make these labeled training documents, however, it is simple to gather the unlabeled documents. The unsupervised learning methods overcome these difficulties.

B. *The Lexicon-based Approach* relies on finding the opinion Lexicon then utilizing them for analyzing the text. Opinion words are used in many sentiment classification tasks. Positive opinion words are utilized to express several desired states, while negative opinion words are utilized to express some of the undesired states. As well, opinion phrases and idioms which together are called the *opinion lexicon [32]*.

There are three leading approaches used for compiling or collecting the opinion word list.

The manual approach is consuming a lot of time and it is not applied alone. It is usually combined with the two others automated approaches to avoid the mistakes in the final review which resulted from automated methods.

There are two automated methods in this approach, as presented in the following subsections:

1) *Dictionary-based approach:* search on the opinion seed words first then investigate on the dictionary of synonyms and antonyms. This method has the main disadvantage that is the inability to detect opinion words with domain and context specific orientations;
2) *Corpus-based approach*: used in solving the problem of searching for the opinion words with the context of certain orientations. Its methods rely on syntactic patterns or patterns that come together along with a seed list of opinion words to detect other opinion words in a large corpus. The solo use for the corpus-based approach is not as effective as the dictionary-based approach; because of the hardness of setting a huge corpus to cover all English words. Also, the main advantage of this approach is it can help to get a domain and context specific opinion words and their orientations using a domain corpus.

The corpus-based approach is applied using statistical approach or semantic approach as illustrated in the following subsections:

2.1. *The Statistical approach* applied to get the co-occurrence patterns or seed opinion words. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus.

2.2. *The Semantic approach* uses sentiment values directly and depends on several principles for calculating the similarity between words. This principle uses similar sentiment values to semantically close words.

4. Approaches to Credibility Assessment

The assessment of information credibility in the online user-generated content is often considerably more complex than in former media contexts due to "*the multiplicity of sources embedded in the numerous layers of online dissemination of content*" [43].

The researches on information credibility are extensive, thus the covering in these subsections are by no means complete. the researchers just provide an outline of the researches that is most closely related to the study.

4.1 Credibility on Twitter Content

In recent research studies, there is a variety of different approaches that have been proposed to assess the credibility of micro-blogging services, one of which is the time-sensitive nature of online content.

In the literature, these methods can be broadly classified into two categories: *classification-based approaches* and *propagation-based approaches* which both exploit the network structure of users and tweets.

4.1.1 Classification-Based Approaches

A research into Twitter credibility for Arabic content has been done by Al-Eidan *et al.*, (2010) [3] and Al-Khalifa *et al.*, (2011) [6], they proposed a system to evaluate the credibility of Twitter Arabic news content automatically by using an evidence-based method; using a *Weighting-based feature approach*; using bag-of-word comparing Twitter content in Twitter trusted news sources. The system used two approaches to evaluate the message credibility levels to (low, high, and questionable). The first approach is based on a computed similarity between thresholds between the content of both Twitter posts and verified news sources such as SPA, Aljazeera, and Google News. The second approach is based on a linear combination of the similarity value with verified content, in addition to a set of proposed extra features related to the content and the source. The formula used to calculate the credibility score: "Credibility Score = 0.6 Similarity + 0.2 Inappropriate Words + 0.1 Linking to authoritative source + 0.1 Author feature". They evaluated their classification result against three political experts' evaluation using a dataset of 29 tweets and four news articles on two topics: (Iran) and (Yeman & Houthi). Preliminary evaluations of the system showed that the first approach was more effective in rating the credibility of tweets compared to the second approach. They received average precision and recall as 0.52 and 0.56 respectively. However, using this approach, the system rated the tweets to only two credibility levels: (Low and High), while in the second approach, it was able to assign the tweets to all three levels of credibility: (Low, High, and Questionable). Linking source degree assigned by an expert was the main prominent feature in the second approach. It should be noted that the above method was only useful for tweets combined with credible external sources and didn't embrace most of the prominent features proposed by previous research such as hash-tags, re-tweets, and emoticons. Moreover, there was a need to evaluate the credibility formula and its performance after the addition of more features.

Castillo *et al.*, (2011) [11], presented a promising study for the information credibility of news propagated through Twitter. They focused on analyzing microblog postings related to "trending" topics using automatic methods for assessing the credibility of a given 'time-sensitive' set of tweets by applying J-48 Decision Tree algorithm. Specifically, the authors identified four types of features depending on the scope: message-based features, user-based features, topic-based features, and propagation-based features. *Message-based features* considered the content of the message (e.g., length, syntax, sentiment, etc.). *User-based features* considered users' behavior on posting messages (e.g., registration age, number of friends and followers). *Topic-based features* were acquired by aggregating the previous two feature sets (e.g., the fraction of tweets with "URLs", "hashtags" and "positive" or "negative" sentiments). Finally, *Propagation-based features* examined the propagation network that could be constructed from the retweets of a message (e.g., the depth of the re-tweet tree, or the number of initial tweets of a topic). This approach was established on a supervised learning method which helped in assessing its effectiveness. 2,500 trending topics via Twitter monitor were gathered to construct an appropriate dataset, focusing only on those cases including 10,000 tweets at most. The labeling of the dataset was executed in two phases. In the first phase, some evaluators (by means of the AMT) were applied to assess whether the collected tweets concerned news about specific events or rather if they were 'conversation' topics. Then, for the

subset of tweets specified as news, the second group of evaluators was applied to label them as credible or not credible. By executing a threefold validation strategy and by applying different classifiers (i.e., SVM, Decision Trees, Decision Rules, and Bayesian Networks), the authors obtained good results in identifying 'newsworthy' topics while using a J-48 decision tree; they reached an 86% accuracy in credibility classification with respect to a random predictor. A feature analysis was also provided to illustrate the contribution that, different types of features were given in terms of credibility assessment. The outcomes of this work were: (1) *message* and *user-based features*, these were not enough to effectively classify credible trending topics, (2) In general, credible news was propagated by authors who previously wrote a considerable number of messages, and (3) *propagation-based features* (having many re-posts) were particularly effective. Moreover, the authors outlined those features such as propagation level, URLs inclusion, and sentiment helped to effectively classify topics automatically as credible or not credible, that tweets not including URLs were in most cases related to non-credible news, while tweets including negative sentiment were related to credible content. However, their method was based on topic credibility rather than individual tweets.

A study focusing on individual tweets or users covered by Kang *et al.*, (2012) [27], proposed an approach for assessing credibility within specific microblog topics. The models evaluated to identify credibility ratings of 1023 tweets collected from topic-specific, Libya. The authors defined three computational models that were based on the features related to different models: (1) *Social (source) Model*: focused on credibility at the user level, harnessing various dynamics of information flow in the underlying social graph to compute a credibility rating. Its features were connected to the network-structure connecting users and tweets (e.g., 'follow' relationships, retweets, etc.), (2) *Content Model*: applied a content-based strategy to compute a finer-grained credibility score for individual tweets. Its features were extracted from the content of tweets related to specific topics (e.g., length of the tweet, number of URLs, number of mentions, etc.), and (3) *Hybrid Model*: used both source and content features to predict credible information and credible information sources, using both averaging and filtering hybrid strategies. The preliminary experiments were performed using Bayesian classifiers (among others) to learn a model based on the features of each prediction strategy. For the full experiment, a J-48 tree-based learning algorithm was used, firstly, because it performed well in preliminary tests, and secondly, to allow for comparison of similar results with Castillo *et al*., (2011) [11]. In contrast to Castillo *et* al., (2011) [11] who proposed an algorithm that acted on groups of 'newsworthy' tweets, the approach described in Kang *et al.*, (2012) [27], aimed at classifying each tweet individually, and at predicting user credibility. Not surprisingly, as already demonstrated in the previously described work, the features that consider the underlying network and the dynamics connected to information flows were better indicators of credibility in microblogs than linguistic features. The results obtained showed that the social features model outperformed both content-based and hybrid models, achieving a best predictive accuracy result 88.17%, compared with 62% and 69% for content-based and the next best-performing hybrid (weighted strategy) respectively.

Another work that assesses the credibility of individual tweets is the one by Gupta & Kumaraguru, (2012) [18]. In this work, the authors used the SVM rank algorithm (a variant of the SVM algorithm used to solve certain ranking problems by learning how to rank), to illustrate that ranking tweets based on Twitter features (i.e., content- and user-based) could help in assessing the credibility of tweets about an event. To this purpose, tweets were labeled by human annotators. To evaluate the effectiveness of the proposed approach, the standard normalized discounted cumulative gain (NDCG) metric was employed.

In another paper, Castillo *et al.*, (2013) [12], discussed more broadly the problem of information credibility on microblogging services. By presenting a case study about information propagation and news event credibility in Twitter, they first summarized their prior work, and then redesigned the learning scheme. A first supervised classifier was used to decide if an information cascade corresponds to a 'newsworthy' event. A second supervised classifier was employed to decide if this news event should be considered credible or not. Both classifiers were trained over labeled data, obtained using crowdsourcing tools. By introducing a label named 'unsure' to identify tweets that belong neither to news nor to chat messages, these 'unsure' tweets were removed from the training dataset, thus improving the performance of the approach.

Abbasi & Liu, (2013) [1], proposed an algorithm called "CredRank" to prove the user credibility in social media. To measure the credibility of the social media users' online behavior this algorithm used for analysis. The outcome had offered each independent user an equal vote and a chance to publicize his/her content. They used the following steps: (1) detect and cluster coordinated users (dependent users) together and (2) weight for each cluster based on the size of the cluster. Then they prepared the CredRank algorithm to perform these two steps. The authors suggested a method to detect coordinated behavior in social media and assigned a lower credibility weight to users who were engaged in the coordinated behavior. The proposed algorithm helped them to detect individuals who used many social media accounts and do so in a way to diffuse their content. In many cases, the CredRank algorithm could be applied, for example in forbidding the distribution of rumors, preventing coordinated activities, and disappointing fake product reviews.

Ikegami *et al.*, (2013) [25], tackled the problem of the credibility measurement on Twitter at Great Eastern Japan Earthquake in 2011 event. They propose a method for automatically assessing the credibility of information based on the topic and opinion classifications. They assessed the credibility of information by calculating the ratio of the same opinions to all opinions about a topic; count opinions in each post. For identifying which topic is mentioned in a tweet, the method uses topic models generated by Latent Dirichlet Allocation. For identifying whether an opinion of a tweet is positive or negative, the method performs sentiment analysis using a semantic orientation dictionary. If a user received several positive opinions with his/her post, the credibility of that post would be high. The result of performing the proposed approach is more than 0.6 in kappa statistics between their method and human score.

In subsequent work, Gupta *et al.*, (2014) [19], extended the previous work and proposed "TweetCred", a real-time web-based system to assess the credibility of content on microblog in the form of a Chrome extension. "TweetCred" takes a direct stream of tweets as input and computes the credibility for each tweet on a scale of 1 (low credibility) to 7 (high credibility). The authors expanded the work with a semi-supervised ranking model using SVM-Rank for assessing credibility. The works established on tweets were related to the six high impact crisis events of 2013. With respect to prior work, in this paper a more exhaustive and comprehensive set of features is using 45 features categorized as tweet meta-data features, tweet content features, user-based features, network features, linguistic features, and external resource features. The dataset was obtained through crowdsourcing, where human assessors labeled around 500 tweets per event, which were selected randomly. This system provides useful insights on how credibility evaluation models evolve over time.

As in the case of opinion spam detection, recent approaches are

focusing on the identification of spammers as well as fake news detection. Galán-Garca *et al*., (2014) [17], focused on the detection of troll profiles on Twitter. Their assumption was that since every trolling profile was followed by the real profile of a user behind the trolling one. It is possible to link a trolling account to the corresponding real profile of the user behind the fake account, analyzing different features present in the profile, data connections, and characteristics of tweets, using machine-learning algorithms. Using manually-selected genuine profiles and collecting at least 100 genuine tweets per profile, the authors compared the performance of different classification algorithms (i.e., Random Forests, J48, k-Nearest-Neighbor, Sequential Minimal Optimization, and NB) on features selected: the content of the tweet published by the user, the time of publication, the language and geolocation, and the Twitter client. Even if this approach achieved 'only' 68.47% of accuracy at best, this work constitutes one of the first significant approaches that tried to directly detect trolls in Social Media and Microblogs in particular, and it had been applied to a real case study for the identification of students responsible for cyberbullying in a school in the city of Bilbao (Spain).

AlMansour *et al.*, (2014) [7], presented a systematic review of the current developments in assessing information credibility automatically in UGC platforms, focusing on micro-blogging service. They designed eight criteria to compare several credibility models to examine how they integrate the three credibility components (*contexts*, *features*, and *evaluator*). They presented a novel theoretical credibility model which integrates *context* factors and *evaluator* traits to assess information credibility; first, in terms of *context*, they considered the Arabic community and assessed the impact of cultural differences on credibility perceptions. Then second, in terms of the *evaluator*, characteristics relevant to information credibility perception in assessment. Therefore, they proposed a theoretical credibility assessment model that takes evaluators' trait differences into account.

Also, Gupta & Kaushal, (2015) [20], proposed an integrated approach for spammer detection, which combines three learning algorithms, (i.e., NB, Clustering, and Decision Trees), with the aim of improving spammer detection accuracy. The considered features are followers/followees, URLs, spam words, replies, and hashtags. The algorithms' accuracy in the detection of non-spammers is about 99.1%, but the accuracy of detecting spammers reaches 68.4%, which was the same result obtained by Galán-Garca *et al.*, (2014) [17].

In another work, AlMansour & Iliopoulos, (2015) [8], presented a study that automated the credibility assessment of Arabic Twitter messages using machine learning approaches. They employed the idea of crowd-sourcing where users could explicitly express their opinions about the credibility of a set of tweets. They distinguished three main groups of features: authority and topical expertise (of the source), data quality (of the content), and popularity (of the content and the source). *Implicit* and *explicit* methods have been used to check the prominent features consumed to assess the tweet messages credibility. *For the implicit method*, a histogram was used to detect the percentage of occurrences of these features in different credibility classes. *For the explicit method*, a user survey was used to rate the importance of features for assessing messages' credibility. The study was conducted using an online survey that took place from Oct/13/2014 to Dec/10/2014 with a sample of 52 raters. They independently evaluated and submitted 4173 credibility evaluation values for a sample of 199 tweet messages from 9 news topic categories: hard news topics such as crises, politics, health, and soft news topics such as entertainment and sports. The corpus of evaluated tweet messages was gathered using NodeXL Twitter Search API tool. They found that features related to the source authority and expertise, and data quality factors are common in both methods and would be used to identify high-credibility messages. In addition, for data quality factor, the linguistic features based on analyzing textual content and writing style were more capable to classify credibility levels. Their findings suggested the importance of using linguistics features and text analysis tools to automate credibility classification solutions and to identify correct information.

El Azab *et al.*, (2016) [13], presented a classification method for recognizing the fake user accounts on Twitter. The authors proposed an approach based on determining the minimum set of attributes that could detect the fake user accounts with the highest accuracy, and then the determined factors were applied using various classification techniques such as Random Forest, Decision Tree, Naïve Bayes, Neural Network, and Support Vector Machine. The attributes have been collected from different researches, they have been clarified by extensive analysis as a first phase, and thereafter the features have been measured. Different experiments have been processed to reach the less set of attributes with realizing the best accuracy results. From 22 attributes, the proposed approach had reached only 7 effective attributes for fake accounts detection. They had compared the study with different recent researches in the same area; this comparison proved the accuracy of the proposed study. Although, they claim that this study could be constantly utilized on Twitter social media to automatically detect the fake accounts; furthermore, the study could be used on different social network sites such as Facebook with minor changes according to the unique nature of each social network.

El-Ballouli *et al.*, (2017) [14], presented a novel credibility model for Arabic content on Twitter called CAT. This model was established on a Machine Learning approach such as Naïve Bayes, SVM, and Random Forest Decision Tree. They used features extracted directly or indirectly from the users' profile and timeline. The used feature-set consisted of 48 features categorized into content-based and user-based features. The content-based features have consisted of 26 features. These features were classified into four subcategories, as sentiment, social, meta, and textual features. However, the user-based features have consisted of 22 features. These features were classified into three subcategories, as network, meta, and timeline. To train the proposed model, they created a corpus of 9,000 Arabic tweets that were a topic-independent. Then, they validated the proposed model by performing two tests. (1) comparing CAT to three baselines, (2) comparing CAT to a state-of-the-art tweet credibility classifier-TweetCred (Gupta et al., 2014). The outcomes from comparison proved that TweetCred was the best work available on credibility classification on Twitter. The scores obtained from TweetCred API ranged from 1 to 7, where 1 indicates low credibility and 7 indicates high credibility. To fairly compared CAT to TweetCred, they projected TweetCred's credibility scores to two values, namely credible or non-credible. After implementing the CAT, the model trained a binary classifier that achieved 21% in Weighted Average F-measure (WAF-measure) which outperformed all baselines and a state-of-the-art approach. The WAF-measure was the sum of all F-measures, each weighted due to the instances number with that class label. The WAF-measure allowed 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%.

4.1.2 Propagation-Based Approaches

The approaches that focus on the concept of propagation for assessing the credibility of tweets/news, usually consider the propagation of rumors (false claims) in microblogs, by exploiting the network structure constituted by retweets and the social graph constituted by followers and followees. Furthermore, trust propagation on the social graph can be assessed.

Mendoza *et al.*, (2010) [34], explored the behavior of Twitter users in emergencies (the 2010 earthquake in Chile); a preliminary study on the dissemination of false rumors and confirmed news was reported. On a crawled dataset containing tweets and other user-related information, the approach analyzed characteristics of the social network of the community surrounding the topic and how trending topics propagate. The considered characteristics were the relation between the number of followers/followees, the number of tweets each user posts, and the retweet activity during the first hours of the emergency. From this analysis, it emerged that the network topology characteristics remain unchanged with respect to normal circumstances and that the vocabulary used in critical situations exhibits a low variance. Concerning credibility and the spread of rumors through the network, the authors had manually selected some confirmed truths, i.e., reliable news items confirmed by reliable sources, and some false rumors, i.e., baseless rumors that emerged during the crisis. The outcome of this credibility study is that rumors tend to be questioned much more commonly than confirmed news by the Twitter community. Therefore, the authors suggested that microblogs could implement methods to warn people by automatically reporting highly questioned information back to them.

With respect to Mendoza *et al.*, (2010) [34], the approach proposed by Seo *et al.*, (2012) [38], studied how to identify sources of rumors when there was a limited view on the rumor provenance, and how to determine whether a piece of information was a rumor or not. The method assumed that rumors were initiated from only a small number of sources, whereas truthful information could be observed and originated from many unrelated individuals. This approach relied on the use of network monitors, i.e., individuals had heard a piece of information (from their social neighborhood), but who did not want to disclose either who told it to them, nor when they had learned it. If a monitor receives a rumor, it was called a positive monitor, or a negative monitor otherwise. To find a rumor source, the authors base their approach on the intuition that the source must be close to positive monitors and far from negative ones. For this reason, the authors introduced four metrics: the number of reachable positive monitors, the sum of distances to reachable positive monitors, the number of reachable negative monitors, and the sum of distances to reachable negative monitors. By computing these metrics for each node, it was possible to sort all the nodes in the network. In the resulting sorted list, the first node was the top suspect source of the rumor. In addition, to identify if a piece of information was a rumor, the authors proposed two strategies based on the set of monitors that received the information. A first strategy tried to assign as many positive monitors as possible to each source, producing in this way large information propagation trees. To solve this problem, a second strategy estimated the disparity between actual propagation trees and the ones constructed by the above strategy. To evaluate the proposed approach, a case study involving a real social network crawled from Twitter was reported. Using the experimental data, they evaluated how accurately logistic regression could classify rumor and non-rumors. The proposed approach has shown good potential to help users in identifying rumors and their sources.

Gupta *et al.*, (2012) [21], proposed a credibility analysis approach for assessing the credibility of news events on Twitter; this approach enhanced with event graph-based optimization to solve the problem. The authors start the study by executing the PageRank-like credibility propagation on a multi-typed network depends on events, tweets, and users. Then they enhanced the approach of the basic trust analysis on every iteration via updating event credibility scores employing regularization on a new graph of events. The outcomes concluded from employing the events of two tweet feed datasets, that for each dataset with millions of tweets presented the event graph optimization approach outperforms the basic credibility analysis approach. Furthermore, the used methods were significantly more accurate (∼86%) than the decision tree classifier approach (∼72%). The authors presented the study after starting from the approach described by Castillo *et al.* (2011) [11], the authors introduced some new features to improve it. They then claim that this classification-based approach is neither entity-nor network-aware because events are described by features originally related to tweets and users. Gupta *et al.* (2012) [21], to compute the credibility of Twitter events, they proposed an approach constituted by two modules: (1) a Basic Credibility Analyzer, namely BasicCA, and (2) an Event Graph Optimization Credibility Analyzer, namely EventOptCA. BasicCA acts on a graph constituted by users, tweets, and events. At each iteration, each node shares its credibility value (learned from a classifier) with its neighbors only. Since the assumption of BasicCA is that credible entities are strictly connected, every iteration helps in mutually enhancing the credibility of genuine entities and reducing the credibility of non-genuine ones, via propagation. EventOptCA enhances BasicCA by supposing that similar events have similar credibility scores. Thus, it performs event credibility updates on a graph of events whose edges are weighted with event similarity values. The experiments were conducted using 457 news events extracted from two tweet feed datasets: the first dataset from Castillo *et al.* (2011) [11], and another crawled dataset, whose events were manually labeled as social gossip or news, using 250 of the news events (of which 167 labeled as credible) in their study. On an average, the proposed approach outperforms the classifier-based approach discussed by the authors.

Jin *et al.*, (2014) [26], proposed a hierarchical credibility propagation approach on Microblog. This approach evaluated news credibility with three-layers, which consisted of a message layer, a subevent layer, and an event layer, with links built with semantic and social relations among these entities. As, the authors believed that it was not always true that credible users provide credible tweets with a high probability; furthermore, they were convinced that considering an event as only constituted by one kind of information (i.e., genuine or fake) is debatable. The assumption was that the hierarchical structure of the message to subevent, and subevent to an event, could model their relations and the process of credibility propagation; furthermore, with a subevent layer, deeper semantic information could be revealed for an event. Credibility propagation was modeled as a graph optimization problem. To validate the effectiveness of the proposed model, two datasets on Sina Weibo (the leading microblogging platform in China), were collected: one with random fake news in a year and truthful news at the same time; another with both fake and truthful news related to the same topic. Experiments on both datasets showed the effectiveness of the proposed model in terms of accuracy by more than 6% and F-score by more than 16% over a baseline method.

Zhao *et al.*, (2016) [49], proposed a novel method to estimate the user/data trustworthiness in Twitter for a specific topic domain. The authors considered the relationships between users/ tweets and multiple characteristics (i.e., textual, spatial, and temporal features) connected to them. First, the approach evaluated the trustworthiness of each tweet and each user. To do that, they evaluated the similarity between features, under the assumption that a candidate tweet (and the user who wrote it) was considered trustworthy if its features did not conflict with the features of trustworthy news. Then, by means of four propagation rules defined on the social graph, the trustworthiness of tweets and users was refined and propagated. The evaluation of the similarity-based trust evaluation method was based on two datasets: a manually labeled set, and the dataset provided

by Castillo *et al.*, (2011) [11]; both datasets were also employed to identify emerging events. Based on the resulting precision and f-score values, this method outperforms classification-based supervised learning approaches.

4.2 Credibility on Facebook Content

Limited numbers of recent researches had attempted to do studies on credibility content on Facebook. In the below subsections, the researchers tried to survey the existing studies related to the research idea. But due to the few researches that been found, the researchers additionally tried to survey studies in sentiment classification concerning Facebook content.

Yaakop *et al.*, (2013) [48], presented a study that concerning the issue of credibility in Facebook Advertising. The study tested the factors that impact consumers' perceptions and attitudes towards advertising on Facebook. The outcomes proposed that there were three online factors that impact in consumers' attitudes significantly towards advertising on Facebook. The factors were perceived interactivity, advertising avoidance, and privacy. Also, questionnaires were done on 350 respondents who studying the program of Bachelor of Management Marketing in the Faculty of Management & Economics in University Malaysia Terengganu (UMT).

Li & Suh, (2015) [29], presented a study examined the factors that influence individuals' perceived information credibility on social media platforms. The authors have identified five factors from two dimensions of credibility (medium and message credibility), based on the persuasion theory—the Elaboration Likelihood Model (ELM) were key ingredients in the online information assessment. Then developed a research model that predicted individuals' perceived information credibility on social media platforms. They tested and validated the proposed model with empirical data from 135 users of the Facebook page; among the 135 respondents, 72 (53.3%) were male and 63 (46.66%) were female. They were generally young, less than 30 years old (85.18%). The results were given that interactivity, the credibility of medium dimension (medium dependency) and the credibility of the message dimension (argument strength) were the main determinants of the information credibility. To test the proposed research model, they adopted a cross-sectional survey method for data collection and evaluated them hypotheses by applying the partial least squares (PLS) method. The unit of analysis for this study was an individual user of social media platforms.

Saikaew *et al.*, (2015) [37], proposed features then developed a system for measuring credibility on Facebook information. First, the authors proposed a FB credibility evaluator for evaluating the credibility of each post by manual human's labeling; then gathered the training data for constructing a model using SVM. Secondly, they developed a chrome extension of FB credibility for Facebook users to evaluate the credibility of each post. Based on the usage analysis of their FB credibility chrome extension, about 81% of users responded agree with suggested credibility automatically computed by the proposed system. Relating to the work of Gupta *et al.*, (2014) [19], There was a difference between the two works; TweetCred chrome extension retrieved data from Twitter API. By getting post ID of Twitter and sent it back to the server. It is called the Twitter API for retrieving completed data from Twitter API, but Facebook did not allow total data access. They retrieved Facebook feature data by using JavaScript code that parses the Facebook page.

Algarni *et al.*, (2016) [5], presented a study for measuring source credibility of Social Engineering attackers on Facebook. The authors proposed four dimensions of source credibility. They classified them as sincerity, competence, attraction, and worthiness. The study founded on the 13 Facebook-based source characteristics that impact users to judge the attacker as per one of the credibility dimensions. They additionally designed 20 different Facebook profiles using Fractional Factorial Design. Then, computed reliability coefficients of the scales using Cronbach's alpha. The accuracy results showed high levels of reliability for the source credibility dimension as 0.97, 0.96, 0.95, and 0.98 for perceived sincerity, perceived competence, perceived attraction, and perceived worthiness, respectively.

Wani *et al.*, (2016) [47], presented a novel approach for the prediction of fake profiles on Facebook using supervised machine learning algorithms like SVM, DT, ANN and NB to classify the user profiles into fake and genuine. Before doing the analysis, the proposed model had used sophisticated noise removal and data normalization techniques on datasets. A method was applied to recognize the non-significant attributes in datasets and to made attribute reduction accordingly by performing natural inspired algorithms like Artificial Bee Colony (ABC), Ant Colony Optimization (ACO). The authors applied the 5-fold cross validation where the data was split into 5 equal subsets, each of which was held out in turn as the testing data while the algorithm was trained on the remaining 4 subsets. The end results were noted according to the performance of the model which was gathered through False Positive and False Negative Analysis. The outcomes indicated that the AdaBoost classifier performance rise with the growth in several profiles in the training dataset. Based on the analysis there is no model used for fake or genuine Facebook profiles detection. However, a combination of two or more machine learning algorithms could be applied for detecting fake or genuine profiles on Facebook.

4.3 Sentiment Analysis On Facebook Content

Troussas *et al.*, (2013) [44], proposed a system used for classifying Facebook status updates. They had employed sentence-level classification to classify opinions whether it positive, negative or neutral emotions. They applied three classifiers, namely Naïve Bayes, Rocchio, and Perceptron, to match their performance in predicting whether a Facebook status update was positive or negative. They gathered about 7000 status updates from 90 users. The status updates were then manually labeled as positive or negative. To assist the language learning, sentiment analysis had applied. They performed stimulating on the educational process and empirical outcomes on the Naive Bayes Classifier. Moreover, they offered leading features that fulfill a significant earn over a unigram baseline.

Hamouda *et al.*, (2013) [23], developed a corpus for sentiment analysis and opinion mining purposes using different machine learning algorithms for Arabic Facebook news pages. They applied several sets of features with various classifiers; SVM, NB, and DT to detect the features that offer the best performance. The classifiers classified the comments into three categories: supportive comments 'y', attacking comments 'n', and neutral comments 'u'. Adding negation words and similarity features for all words in posts and comments features given the best performance. The evaluation results, Naive Bayes gives 59.9%, while with the Decision Tree, the precision and recall improved with 10%. Finally, Support Vector Machine classifier gave the best results with 73.4% of accuracy for precision and recall.

Soliman *et al.*, (2014) [42], proposed a sentiment analysis approach to classify Arabic slang comments on Facebook, based on Support Vector Machine (SVM) to classify comments as satisfy or dissatisfy comments. In addition, they developed a Slang Sentimental Words and Idioms Lexicon (SSWIL), containing opinion words and idioms used by the Arab youth generations. The new lexicon was collected manually from microblogs websites. They applied three types of classification: (1) classifying comments without applying SSWIL, (2) classifying comments after the creation of SSWIL, and (3) classifying comments using SSWIL only. All three were based on SVM classifier. They evaluated 1355 random

comments, by applying the proposed mining in all comments the performance of the proposed classifier, several Facebook news' comments were used, where 86.86% accuracy rate was obtained with precision 88.63% and recall 78%.

Ortigosa *et al.,* (2014) [35], proposed an approach for sentiment analysis on Facebook. The authors applied it in two steps by giving the users' messages: (1) detected information about the users' sentiment polarity (positive, neutral or negative). (2) discover critical emotional changes and form the users' ordinary sentiment polarity. They have developed SentBuk, a Facebook application that retrieved the messages, comments, and likes on the user's profiles, then classified the messages according to their polarity, and built/updated the user sentiment profile. SentBuk pursued a hybrid approach that classifies through merging lexical-based and machine-learning techniques (as NB, SVM & DT). They used Naïve-Bayes and SVM as the commonly used algorithms for sentiment analysis, and decision-trees as it is easier to explain the rationale behind a given classification. They explained that the newest version of SentBuk concentrated on ''status messages'', addition to comments and likes associated to these messages. And SentBuk acquired the outcomes via an interactive interface, that also helped the representation of emotional change detection, friend's emotion finding, user classification relating to their messages, and statistics, among others. They have formed a lexicon-based approach, (1) to join the famous techniques with them enhancements to get better results from the analysis of Facebook messages. (2) to join the lexicon-based approach with several machine-learning techniques from creating different hybrid classifiers, to get better performance of Facebook message sentiment analysis. The accuracy gained from joining the lexicon-based techniques (for pre-processing) and Support Vector Machines (for classifying) was the highest one (83.27%). This was the combination supported by Sentbuk currently.

Setty *et al.,* (2014) [39], proposed a system for classification of Facebook news feeds and automatic detection of sentiments. They adopted the approach of Gmail that enriched with automatic classification of emails into (primary, social and promotions) to automatically classify Facebook news feeds into (life posts and entertainment posts). Further, they performed sentiment analysis of life events posts into (happy, neutral and sad posts) to provide a better structure to the Facebook. The dataset was used for training the learning classifier. The training set (new posts) was classified as life events posts and entertainment posts according to the learned classifier. Moreover, life events posts were labeled as happy, neutral and sad posts according to the sentiment score value specified using SentiWordNet dictionary and POS tagger. Classification accuracy of posts was performed through several classifiers using WEKA and the learning model approach. Two-fold cross-validation was applied to assess the accuracy using Weka. They used various classification algorithms for comparison as Binary Logistic Regression, Bayes Net, J48, Naive Bayes, and SVM. Approximately 2000 posts were considered for analysis. According to the results showed SVM followed by Bayes Net had shown better accuracy than other classification algorithms for a used dataset.

DISCUSSION

Several approaches presented in this research study have been defined in the last years to tackle the problem of assessing the credibility on Twitter. As previously illustrated, most of the proposed approaches are based on supervised or semi-supervised techniques that make use of multiple kinds of characteristics that can be related to credibility. With respect to credibility on Facebook, a limited number of approaches have been proposed to assess the generated-content credibility. Table 1. summarizes the main approaches that have been proposed so far for assessing the credibility on Twitter, while table 2., summarizes the approaches that have been proposed for assessing the credibility on Facebook. Additionally, the researcher in table 3. summarizes the approaches that have been proposed for sentiment analysis on Facebook.

5. A Novel Approach Proposed for Assessing Credibility on Facebook

In this section a new proposed approach for assessing credibility on Facebook is presented as illustrated in Fig. 4.

(1) The proposed model tries to evaluate the content credibility and to detect the trusted users on Facebook. Firstly, by detecting the Facebook spammers and remove the posts created and published by them as a pre-processing step. Then by evaluating the other Facebook users in terms of proposing credibility measures to use them in the evaluation of the text content for the user profile. After filtering the spammers and the fake users; proposing to perform text content detection and clustering credibility for text content on Facebook.

(2) The evaluator for the text content consists of Machine Learning Models used to train a classifier, and then run it in the Engine. After that, training and building models based on support vector machines are adopted to extract deeper meaning and increasing accuracy for analytics.

(3) The classified users' profiles ranked through the rank builder according to the number of likes, posts, shares, comments, shares, fake and genuine profiles.

(4) When a new post created by Page, a complete analysis will be performed to the user post or comment. The researchers will try to cover these dimensional views in the proposed research study: "Text Analysis, URLs Inclusion, Demographics, Engagement, and Sentiment Analysis."

(5) The post credibility score will be measured through a formula combined from the Page profile rank and the post-analysis score.

6. Result Analysis and Discussion

The performance of the proposed model has been evaluated by applying it on 35 Facebook pages related to Press agencies and each page has 20 average post numbers. In order to test the performance of the proposed model, an extensive experiment has been done.

The experiment steps are as follows:

1) Spammer detection is responsible for detecting and removing the fake pages from the model.
2) Building the Page rank according to the following criteria:
   - Likes
   - Shares
   - Posts
   - Comments
   - Fake or genuine
3) Each new created page's post is analyzed through the following steps:
   - Text analysis
   - Engagements
   - Sentiment analysis
   - Links
   - Demographics

4) The final credibility score is calculated through the following formula: $Credibility\ Score = Page\ profile\ rank + Post\ score$ (8)

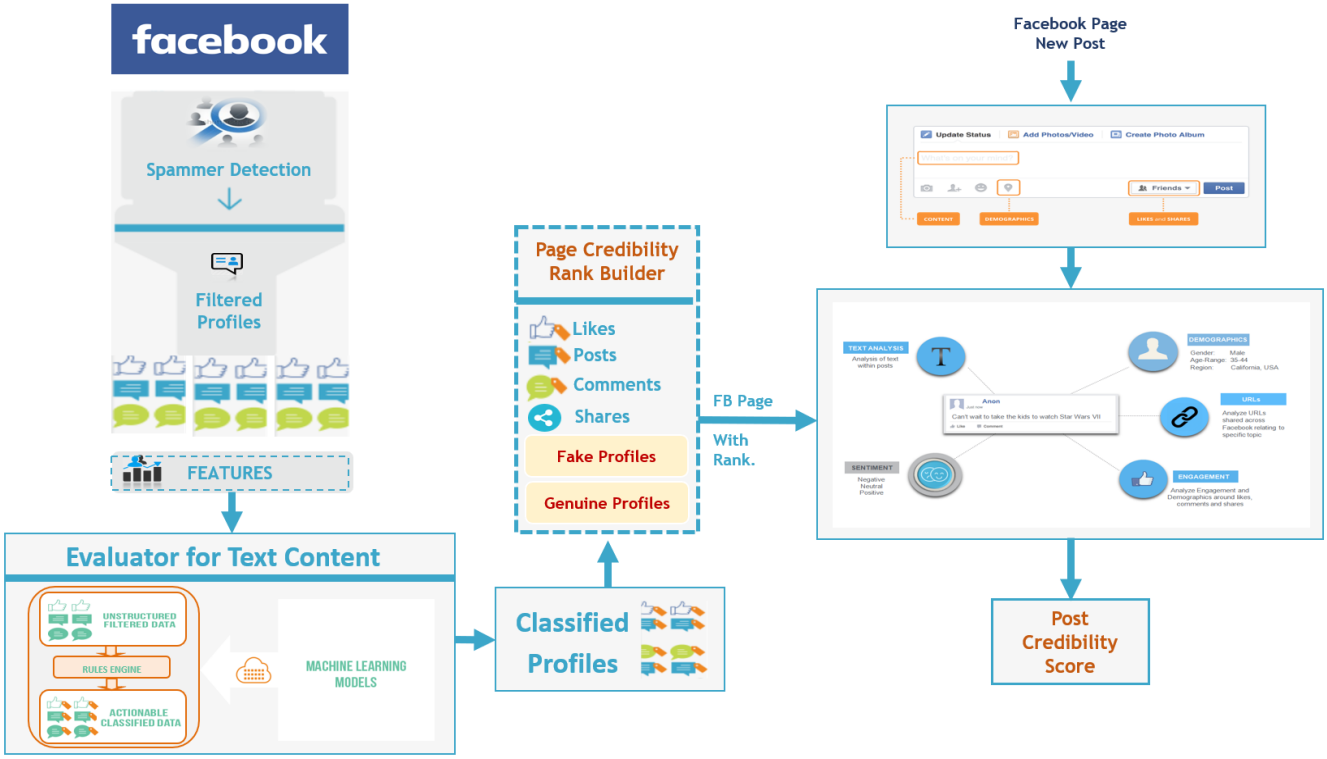

Figure 4: A Proposed Model for Assessing Credibility on Facebook.

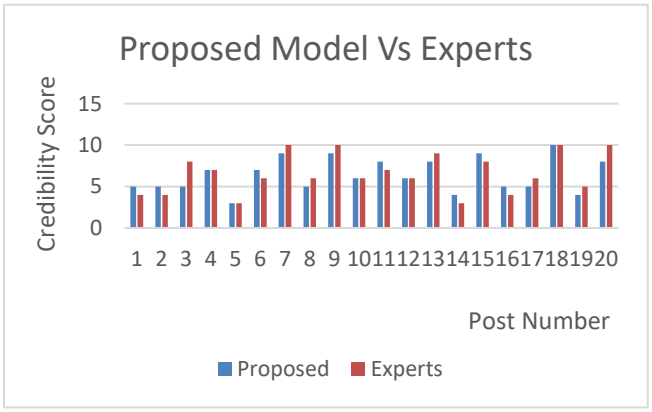

Figure 5: Proposed Model Vs Experts.

The output of the system is compared with experts' judgments for verifying the effectiveness of the proposed model.

Table 4. shows a sample of the evaluation process for verifying the accuracy of the proposed system, and Fig. 5 visualizes the convergence between the credibility score created by the proposed model and the credibility score created by experts.

The total number of test posts is 648 posts for 35 Facebook pages and as shown the proposed model achieved 87.45% accuracy that proved the effectiveness of the proposed model.

7. Research Contribution

The contribution of this research can be summarized as follows:

- Proposing a new model used to measure the credibility of Facebook posts through a formula combined from the page profile rank and the post-analysis score.
- The model was tested and achieved 87.45 % accuracy.

8. Conclusion

In this paper, the researchers presented some methods that were adopted and investigated for assessing credibility on social media-generated content using Sentiment Classification. Many previous studies have been investigated credibility content in Twitter, but only a small number of recent researches have been attempted to study credibility content on Facebook. They started by discussing the research problems and its relevance, such as Fake news detection and Credibility on Social Media. The study explored the relating background needed for the study such as sentiment classification and machine learning techniques to classify the content and sources. After that, they illustrated the approaches used for assessing the credibility content on Twitter and Facebook. In the end, they discussed a novel approach proposed to assist the credibility on Facebook posts.

Table 1: Summarization of Some Approaches for Assessing Credibility on Twitter.

REF.	AUTHOR	YEAR	ALGORITHM USED	LANGUAGE	FOCUS	APPROACH	OUTCOME
[3] [6]	Al-Eidan *et al.* & Al-Khalifa *et al.*	2010 2011	Semantic (compute similarity)	AR	Content and source credibility	Classification-based	Low, Avg & High, precision & recall 0.52 & 0.56
[11]	Castillo *et al.*	2011	J48 DT	EN	Trending topic credibility	Classification-based	80%Credible or Not
[27]	Kang *et al.*	2012	BN, J48 DT	EN	Tweet and source credibility	Classification-based	Source 88.17%, Content 62%, & Hybrid 69%
[18]	Gupta & Kumaraguru	2012	SVM	EN	Event credibility	Classification-based	0.37 NDCG metric
[12]	Castillo *et al.*	2013	RF, LR, Meta-L	EN	Trending topic credibility	Classification-based	Credible or Not
[1]	Abbasi & Liu	2013	Semantic	EN	User credibility	Classification-based	CredRank algorithm
[25]	Ikegami *et al.*	2013	Semantic	EN	Trending topic credibility	Classification-based	Positive/Negative
[19]	Gupta *et al.*	2014	SVM	EN	Tweet credibility	Classification-based	TweetCred & extension tool
[17]	Galán-Garca *et al.*	2014	RF, J48, k-NN, NB SMO	EN	Troll detection	Classification-based	68.47% accuracy at best
[7]	AlMansour *et al.*	2014	DT, SVM, BN, Statistical, Semantic	AR	Information credibility	Classification-based	Novel theoretical credibility model
[20]	Gupta & Kaushal	2015	NB, Clustering, DT	EN	Spammer detection	Classification-based	Non-spammers 99.1%, Spammers 68.4% accuracy
[8]	AlMansour & Iliopoulos	2015	Statistical	AR	Content and source credibility	Classification-based	Source/Content credibility
[13]	El Azab *et al.*	2016	RF, DT, NB, NN, SVM	EN	Fake user accounts	Classification-based	Only 7 effective attributes for fake accounts detection
[14]	El-Ballouli *et al.*	2017	Semantic	AR	User and content credibility	Classification-based	Credible or Not
[34]	Mendoza *et al.*	2010	Aggregate analysis	EN	Rumor propagation	Propagation-based	Truth/False rumors
[38]	Seo *et al.*	2012	Logistic regression	EN	Rumor and source credibility	Propagation-based	Rumor/Non-rumor
[21]	Gupta *et al.*	2012	Semantic	EN	Event credibility	Propagation-based	Accuracy (~86%)
[26]	Jin *et al.*	2014	Semantic	EN	Tweet and source credibility	Propagation-based	6% accuracy
[49]	Zhao *et al.*	2016	Semantic	EN	Topic-focused credibility	Propagation-based	Trustworthy tweets

Table 2: Summarization of Some Approaches for Assessing Credibility on Facebook.

REF.	AUTHOR	YEAR	ALGORITHM USED	LANGUAGE	FOCUS	APPROACH	OUTCOME
[48]	Yaakop *et al.*	2013	Regression analysis	EN	Advertising on Facebook	Classification-based	3 dimensions
[29]	Li & Suh	2015	Partial Least Squares (PLS)	EN	Information credibility	Classification-based	Medium/Message credibility
[37]	Saikaew *et al.*	2015	SVM	EN	Information credibility	Classification-based	81 % users agree
[5]	Algarni *et al.*	2016	Semantic	EN	Source credibility	Classification-based	4 dimensions
[47]	Wani *et al.*	2016	SVM, DT, NN, NB	EN	Fake profiles detection	Classification-based	Fake or Genuine

Table 3:Summarization of Some Approaches for Sentiment Analysis on Facebook.

REF.	AUTHOR	YEAR	ALGORITHM USED	LANGUAGE	FOCUS	APPROACH	OUTCOME
[44]	Troussas *et al.*	2013	NB	EN	Facebook status update	Classification-based	Positive or Negative NB- Precision 0.77
[23]	Hamouda *et al.*	2013	NB, DT, SVM	AR	Facebook news pages (Post and Comments)	Classification-based	Best accuracy SVM 73.4%
[42]	Soliman *et al.*	2014	SVM	AR	Slang Comments on Facebook	Classification-based	86.86% accuracy
[35]	Ortigosa *et al.*	2014	Lexical-based & (DT, NB, SVM)	EN	Facebook status message (users' sentiment polarity)	Classification-based	Lexical with SVM 83.27% accuracy
[39]	Setty *et al.*	2014	BN, J48, NB, SVM	EN	Facebook news feeds detection	Classification-based	SVM 97-99% Acc.

Table 4: Sample Output of the Proposed Approach for Assessing Credibility on Facebook.

Post Number	Facebook Page Id	Page Rank 1-10	Post Rank 1-10	Credibility Score Proposed Model 1-10	Credibility Score Experts 1-10
1	10155675667923700	8	4	5	4
2	10155675667923700	6	5	5	4
3	10155675667923700	3	5	5	8
4	10213333312077900	8	9	7	7
5	10213333312077900	7	7	3	3
6	10213333312077900	5	8	7	6
7	10155767830832300	5	9	9	10
8	10155675667923700	9	3	5	6
9	10213333312077900	10	7	9	10
10	10213333312077900	10	10	6	6
11	10213333312077900	6	8	8	7
12	10213333312077900	9	5	6	6
13	10213333312077900	9	4	8	9
14	10213333312077900	6	8	4	3
15	10155767830832300	6	3	9	8
16	10155675667923700	5	9	5	4
17	10213333312077900	9	8	5	6
18	10213333312077900	8	8	10	10
19	10213333312077900	3	9	4	5
20	10213333312077900	7	5	8	10

REFERENCES

[1] Abbasi, M. A., & Liu, H., (2013). "Measuring User Credibility in Social Media.", *In Social Computing, Behavioral-Cultural Modeling and Prediction*, 441-448, Springer.

[2] Aggarwal, C. C., & Zhai, C., (2012). "Mining Text Data", *Springer Science & Business Media*.

[3] Al-Eidan, R. M. B., Al-Khalif, H. S., & Al-Salman, A. S., (2010). "Measuring The Credibility of Arabic Text Content in Twitter.", *In Digital Information Management (ICDIM), 2010 Fifth International Conference on*, 285-291, IEEE.

[4] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T., (2015). "Approaches, Tools and Applications for Sentiment Analysis Implementation", *International Journal of Computer Applications*, 125(3), 26-33.

[5] Algarni, A., Xu, Y., & Chan, T., (2016). "Measuring Source Credibility of Social Engineering Attackers on Facebook.", *In System Sciences (HICSS), 2016 49th Hawaii International Conference on*, 3686-3695, IEEE.

[6] Al-Khalifa, H. S., & Al-Eidan, R. M., (2011). "An Experimental System for Measuring The Credibility of News Content in Twitter.", *International Journal of Web Information Systems*, 7(2), 130-151.

[7] AlMansour, A. A., Brankovic, L., & Iliopoulos, C. S., (2014). "Evaluation of Credibility Assessment for Microblogging: Models and Future Directions.", *In Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, 32, ACM.

[8] AlMansour, A. A., & Iliopoulos, C. S., (2015). "Using Arabic Microblogs Features in Determining Credibility.", *In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1212-1219, IEEE.

[9] Assiri, A., Emam, A., & Aldossari, H., (2015). "Arabic Sentiment Analysis: A Survey", *International Journal of Advanced Computer Science & Applications (IJACSA)*, 6(12), 75-85.

[10] Bao, Y., Quan, C., Wang, L., & Ren, F., (2014). "The Role of Pre-processing in Twitter Sentiment Analysis", *In International Conference on Intelligent Computing*, pp. (615-624), Springer, Cham.

[11] Castillo, C., Mendoza, M., & Poblete, B., (2011). "Information Credibility on Twitter.", *In Proceedings of the 20th international conference on World wide web*, 675-684, ACM.

[12] Castillo, C., Mendoza, M., & Poblete, B., (2013). "Predicting Information Credibility in Time-Sensitive Social Media.", *Internet Research*, 23(5), 560-588.

[13] El Azab, A., Idrees, A. M., Mahmoud, M. A., & Hefny, H., (2016). "Fake Account Detection in Twitter Based on Minimum Weighted Feature set.", *International Journal of Computer Electrical Automation Control and Information Engineering*, 10(1), 13-18.

[14] El Ballouli, R., El-Hajj, W., Ghandour, A., Elbassuoni, S., Hajj, H., & Shaban, K., (2017). "CAT: Credibility Analysis of Arabic Content on Twitter.", *In Proceedings of the Third Arabic Natural Language Processing Workshop*, 62-71.

[15] "Facebook - Statistics & Facts",2018. *Available from:* https://www.statista.com/topics/751/facebook/ [*Accessed December,2018*].

[16] Fogg, B. J., (2003). "Prominence-Interpretation Theory: Explaining How People Assess Credibility Online.", *In CHI'03 extended abstracts on human factors in computing systems*, 722-723, ACM.

[17] Galán-García, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I., & Bringas, P. G. (2014). "Supervised Machine Learning for The Detection of Troll Profiles in Twitter Social Network: Application to A Real Case of Cyberbullying.", *In: International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, 419–428, Springer.

[18] Gupta, A., & Kumaraguru, P., (2012). "Credibility Ranking of Tweets During High Impact Events.", *In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*. New York, NY, 2, ACM.

[19] Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P., (2014). "Tweetcred: A real-time Web-based System for Assessing Credibility of Content on Twitter.", *In* Proceedings *of 6th International Conference on Social Informatics (SocInfo), Barcelona, Spain*.

[20] Gupta, A., & Kaushal, R., (2015). "Improving Spam Detection in Online Social Networks.", *In Cognitive Computing and Information Processing (CCIP), 2015 International Conference on*, 1-6, IEEE.

[21] Gupta, M., Zhao, P., & Han, J., (2012). "Evaluating Event Credibility on Twitter.", *In Proceedings of the 2012 SIAM International Conference on Data Mining*. *Society for Industrial and Applied Mathematics*, 153-164.

[22] Haddi, E., Liu, X., & Shi, Y., (2013). "The Role of Text Pre-processing in Sentiment Analysis", *Procedia Computer Science*, 17, 26-32.

[23] Hamouda, A. A., & El-Taher, F., (2013). "Sentiment Analyzer for Arabic Comments System.", *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(3), 99-103.

[24] Hilligoss, B., & Rieh, S. Y., (2008). "Developing a Unifying Framework of Credibility Assessment: Construct, Heuristics, and Interaction in Context.", *Information Processing & Management*, 44(4), 1467-1484.

[25] Ikegami, Y., Kawai, K., Namihira, Y., & Tsuruta, S., (2013). "Topic and Opinion Classification Based Information Credibility Analysis on Twitter.", *In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 4676-4681, IEEE.

[26] Jin, Z., Cao, J., Jiang, Y. G., & Zhang, Y., (2014). "News Credibility Evaluation on Microblog with A Hierarchical Propagation Model.", *In Data Mining (ICDM), 2014 IEEE International Conference*, 230-239, IEEE.

[27] Kang, B., O'Donovan, J., & Höllerer, T., (2012). "Modeling Topic Specific Credibility on Twitter.", *In Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, 179-188, ACM.

[28] Kaplan, A. M., & Hanlein, M., (2010). "Users of the World, Unite! The Challenges and Opportunities of Social Media.", *Business Horizons*, 53(1), 59-68.

[29] Li, R., & Suh, A., (2015). "Factors Influencing Information Credibility on Social Media Platforms: Evidence from Facebook Pages.", *Procedia computer science*, 72, 314-328.

[30] Liu, B., (2012). "Sentiment Analysis and Opinion Mining", *Synthesis Lectures on Human Language Technologies*, , 5(1), 1-167.

[31] Maynard, D., & Funk, A., (2011). "Automatic Detection of Political Opinions in Tweets", *In Proceedings of the 8th international conference on The Semantic Web, (ESWC '11)*, pp. (88-99), Springer-Verlag.

[32] Medhat, W., Hassan, A., & Korashy, H., (2014). "Sentiment Analysis Algorithms and Applications: A Survey", *Ain Shams Engineering Journal*, 5(4), 1093-1113.

[33] Mejova, Y., & Srinivasan, P., (2011). "Exploring Feature Definition and Selection for Sentiment Classifiers", *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[34] Mendoza, M., Poblete, B., & Castillo, C., (2010). "Twitter Under Crisis: Can we trust what we RT?.", *In Proceedings of the first workshop on social media analytics*, 71-79, ACM.

[35] Ortigosa, A., Martín, J. M., & Carro, R. M., (2014). "Sentiment Analysis in Facebook and Its Application to E-Learning.", Computers in Human Behavior, 31, 527-541.

[36] Pang, B., & Lee, L., (2008). "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, 2(1–2), 1-135.

[37] Saikaew, K. R., & Noyunsan, C., (2015). "Features for Measuring Credibility on Facebook Information.", *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(1), 174-177.

[38] Seo, E., Mohapatra, P., & Abdelzaher, T., (2012). "Identifying Rumors and Their Sources in Social Networks.", *In Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III* , 8389, 838901-1, International Society for Optics and Photonics.

[39] Setty, S., Jadi, R., Shaikh, S., Mattikalli, C., & Mudenagudi, U., (2014). "Classification of Facebook News Feeds and Sentiment Analysis.", *In Advances in Computing, Communications and Informatics (ICACCI)*,18-23, IEEE.

[40] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H., (2017). "Fake News Detection on Social Media: A Data Mining Perspective." , *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36, ACM.

[41] Smith, A., & Anderson, M., (2018). " Social Media Use in 2018.", [Annual Report], *Pew Research Center*.

Available from: http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/, [Accessed June,2018].

[42] Soliman, T. H., Elmasry, M. A., Hedar, A., & Doss, M. M., (2014). "Sentiment Analysis of Arabic Slang Comments on Facebook.", *International Journal of Computers & Technology*, 12(5), 3470-3478.

[43] Sundar, S. S., (2008). "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility.", *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, 73-100.

[44] Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J., (2013). "Sentiment Analysis of Facebook Statuses using Naive Bayes Classifier for Language Learning.", *In Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International* Conference,1-6, IEEE.

[45] "Twitter - Statistics & Facts", 2018. *Available from:* https://www.statista.com/topics/737/twitter/ [*Accessed December,2018*].

[46] Vaghela, V. B., & Jadav, B. M., (2016). " Analysis of Various Sentiment Classification Techniques." , *International Journal of Computer Applications*, 140(3), 975–8887.

[47] Wani, S. Y., Kirmani, M. M., & Ansarulla, S. I., (2016). "Prediction of Fake Profiles on Facebook using Supervised Machine Learning Techniques-A Theoretical Model.", *International Journal of Computer Science and Information Technologies (IJCSIT)*, 7(4), 1735-1738.

[48] Yaakop, A., Anuar, M. M., & Omar, K., (2013). "Like It or Not: Issue of Credibility in Facebook Advertising.", *Asian Social Science*, 9(3), 154-163.

[49] Zhao, L., Hua, T., Lu, C. T., & Chen, R., (2016). "A topic-focused trust model for Twitter.", *Computer Communications*, 76, 1-11.