Fake News Detection Using Machine Learning Approach: A Survey

Ms. Suchitra Deokate

P.G. Student, Department of Computer Engineering

VPKBIET, Baramati, Maharashtra, India.

deokate.suchitra@gmail.com

---***---

Abstract - *The role of social media in our day to day life has increased rapidly in recent years. It is now used not only for social interaction, but also as an important platform for exchanging information and news. Twitter, Facebook a microblogging service, connects millions of users around the world and allows for the real-time propagation of information and news. Fake news has become a major problem in these social networks Fake news has vast impact in our modern society. Existing approaches to news verification depend on features extracted mainly from the text content of news tweets. Detecting Fake news is an important step. This approach that combines analysis of the users reputation on a given topic within the social network, as well as a measure of the users sentiment to identify relevant data and credible score of information it also find fake news or real news used machine learning techniques to detect Fake news, using Support Vector Machine (SVM) algorithm used for classify the with fake news or real news.*

***Key Words*: Social networking site, Twitter, Reputation, Credibility, Fake news, Machine learning.**

1. INTRODUCTION

The role of social media in our day to day life has increased rapidly in recent years. Online social media is a popular platform where lots of people can communicate with each other in real time. These are the dynamic data sources where the users can create their own profiles and communicate with each other independent of physical location. It provides communication platform with large scale and large extent. Furthermore these tools are outside the boundaries of the physical world in studying human relationship and behaviors. As these social Medias are becoming more popular, cybercriminals have exploited these as a new platform for communicating different types of cybercrimes. Twitter, Facebook connects millions of users around the world and allows the real-time circulation of information and news. [3]These factors have resulted in playing a critical role in world events, particularly emergency events, where it has been useful in emergency response and recovery.

Nowadays, different cybercrimes are happening such as phishing, spamming and spread of malware and fake news is considered as a major problem along with the recent development of social media .It is a technique by which users get harass from other individual user of the group of user. Online social media such as Facebook, twitter have become integral component of a user's life. Because of this, these websites have become the most common platform for spared the fake news. Fake News is an inaccurate, sometimes sensationalistic report that is created to gain attention, misinformation, deceive or harm a reputation. Unlike misinformation, which is wrong because a reporter has tangled facts, fake news is created with the intent to manipulate to the user. Fake news can spread quickly when it provides misinformation that is aligned with the audiences point of view because such content is not likely to be interrogated or discounted. Twitter has, however, not only been used for the spread of valid news, and fake news. This fake news can come in the form of spam, AstroTurf is a technique used in political movements to fake support numbers.by making a message appear to have popular origins when in authenticity it originated from one person or organization. [6]The increase in the volume of fake news has level controlled to our current times being labeled the age of misinformation and therefore stresses the importance of assessing the credibility of tweets. Hence, aimed to developed useful information in tweets to detect fake news. Selecting the twitter dataset with streaming API and search API tweets is a complex task that requires considerable efforts in building the machine learning model. Therefor to develop a fake news detection method by identifying which is the fake news or real news that can be used in machine learning techniques. In particular use many features of twitter such as Structural feature, Content features and User features these features use for the user reputation and Credibility of content. User and content features use for calculate the user reputation and Credibility of content. Using these features to train our fake news detection model and improve its performance. Apply machine learning techniques to detect Fake news by using the support vector machine (SVM), algorithm used for classify the tweet with fake or real.

2. REVIEW OF LITERATURE

Ikegami et al. [12] performed a topic- and - based credibility analysis of Twitter tweets, using the Great Eastern Japan earthquake as a case study. The researchers assessed credibility by computing the ratios of similar opinions to all opinions on a particular topic. The topics were identified using latent Dirichlet allocation (LDA). Sentiment analysis was performed using a semantic orientation dictionary to assess whether a tweets opinion was negative or positive. Shlok gilda. (2017)[5] Demonstrates that term frequency is potentially predictive of fake news it is important first step toward using machine Classification for identification. Conroy, Rubin, and Chen [4] outline several approaches that seem promising toward the aim of correctly classifying misleading articles. They note that simple content-related ngrams and shallow part-of-speech (POS) tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep SyntaxManalysis using Probabilistic Context Free Grammars (PCFG) have been shown to be particularly valuable in combination with n-gram methods. [5] [4]. Mykhailo Granik, Volodymyr Mesyura (2017) proposed the Fake News Detection Using Naive Bayes Classifier used for fake news detection method based on one of the artificial intelligence algorithms naïve Bayes classifier. Naïve Bayes classifiers are a general statistical technique of email filtering. Nave Bayes typically use bag of words features to identify spam e-mail and method commonly used in text classification. And this particular method works for this specific problem given a physically labeled news dataset and to support the idea of using artificial intelligence for fake news detection.

Majed Alrubaian (2016) proposed A Credibility Analysis System for Assessing Information on Twitter it is new credibility assessment system .conducted a survey to understand users' perceptions regarding credibility of content on Twitter. They found that the prominent features based on which users judge credibility are features visible at a glance, for example, the username and picture of a user and comprises four integrated components, namely, a reputation based model, a feature ranking algorithm, a credibility assessment classifiers engine, and a user expertise model. All of these components operate in an algorithmic form to analyze and assess the credibility of the tweets on Twitter. The reputation-based technique helps to filter ignored information before starting the calculation process. The classifier engine component distinguishes between credible and non-credible content. [3] [6]. Supanya Aphiwongsophon and Prabhas Chongstitvatana they are select dataset from Twitter are summarized with twenty two attributes. From this information, all the machine learning methods: Naïve Bayes, Neural Network, Support vector machine, are very good at detecting Fake news There are classified to two classes with believable and unbelievable. The results from the classification are precision, recall, F-Measure, and accuracy [4]. Jacob Ross, Krishnaprasad Thirunarayan,they are create a robust and general feature set for learning to rank tweets based on credibility and newsworthiness. Use a set of features that are suggestive of tweets credibility regardless of the time period and topic of that tweet. These features are derived by combining popular and actual features from previous works, as well as new features. Features can be broadly categorized as either user based features or tweet based features. They create sentiment based features that aim to capture when a tweets sentiment is irregular given the context of its topic. Fake News Detection on Social Media According to the sources that features are extracted from, fake news detection methods generally focus on using news contents and social contexts [13]. News content based approaches extract features from linguistic and visual information. Linguistic features aim to capture specific writing styles and sensational headlines that commonly occur in fake news content, such as lexical features and syntactic features [14].

3. SYSTEM ARCHITECTURE

3.1 Features Selection

Initial motivation for feature selection is that the social data often contain many different features that are difficult to deal with this feature, and most of the features are terminated except for specific tasks. To deal with a problem, apply feature extraction. Feature selection is frequently preferred over extraction; [8] they select the three main features first is Structural Features Structural features capture Twitter-specific properties of the tweet stream, including tweet volume and activity distributions. Second is User features capture properties of tweet authors, such as interactions, account ages, friend/follower counts, and Twitter verified status and third features Content features measure textual aspects of tweets, like polarity, subjectivity, no of comments and agreement.

3.1.1 Structural Features

Twitter, Facebook is common online social network services that provide the platform of communication. It enables users to read and send message of length 140 character. Structural features include the text features and sentiment features. Text features: Text features are specific to each Twitter conversation thread and are calculated across the entire thread. These features include the number of tweets, average of tweet length; thread lifetime is number of minutes between first and last tweet, and the depth of the communication tree and frequency and ratio of tweets that contain media like images or video audio, mentions (@), re-tweet, and web links. Tweet metadata is the number of seconds since the tweet; Source of tweet (mobile / web/ etc); Tweet comprises geo-coordinates. Number of characters, Number of words, Number of URLs, Number of hash-tags, Number of unique characters, Occurrence of typical symbol, Occurrence of happy smiley, Occurrence of sad smiley, Tweet contains via; Occurrence of colon symbol.

The credibility of the information based on the ratio of positive to negative tweet. The credibility of the information was then based on the ratio of positive to negative tweet. For every user $u_i \in U$, we calculate a sentiment score (denoted by Δu) based on analysis of his previous tweets, using the

following equation:

$$\Delta\,(u_i) \;=\; \frac{\sum T^+(u_i)}{\sum T^+(u_i) + \sum T\text{-}\,(u_i)}$$

where, T is a user's positive tweets and T- is a user's negative tweets.

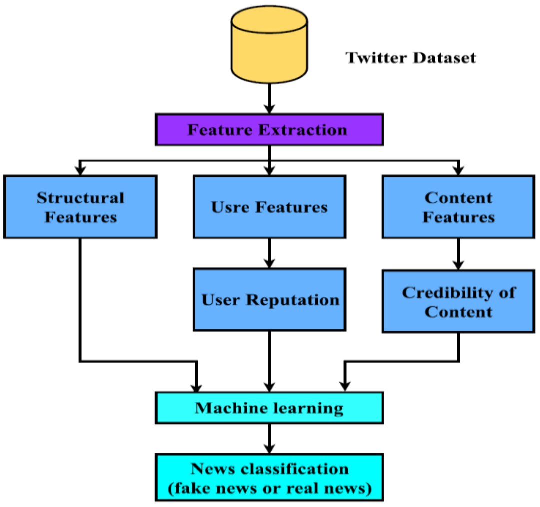

Fig -1: System Architecture

3.1.2 User Features

User features focuses on activities and thread characteristics, following Features attributes of the users taking part in the conversations, their connectedness, and the density of interaction between these users. Each user had a personal record of information in his profile. Some of these features are latent and some of them explicitly revealed in user profiles. For example, age, gender, education, political orientation, and even any user preferences are considered as latent attributes. The number of followers, number of friends and the number of re-tweeted tweets as well as the replies of users tweets, number of likes and unlike on specific topic, and authored status counts, occurrence of verified authors, and whether the author of the first tweet in the thread is verified.

3.1.3 Content Features

Content features are based on tweets textual aspects and include polarity is the average positive or negative feelings expressed a tweet, subjectivity is the score of whether a tweet is objective , and disagreement, as measured by the amount of tweets communicating difference in the conversation. [9] Also borrowing from Castillo et al., we include the frequency and proportions of tweets that contain question marks, exclamation points, first/second/third person pronouns, and smiling emoticons Frequency and proportions of tweets that contain question marks, exclamation points, first/second/third person pronouns, and smiling emoji. .

Textual Features: A text describing the news event. They provide details of the event and may contain certain opinions or thoughts towards the story. General textual Features are derived from the syntax of a text; three categories of general textual features are commonly used: lexical features, syntactic features, and topic features. Lexical features are features extracted at the word-level of a rumor, which could be statistics of words, lexical rumor patterns or sentimental lexicons. Syntactic features represent rumors at the sentence level. The basic syntactic features are simple statistics of a rumor message, such as the number of keywords, the sentiment score or polarity of the sentence and part-of speech tagging. Topic features are extracted from the level at the message set, which aim to understand messages and their underlying relations within a corpus.

3.2 User Reputation and Credibility Assessment

User reputation systems are commonly used in Ecommerce website and social networking sites, such as Twitter, Facebook etc. Most of the user reputation systems use the rule-based method or the voting systems to calculate user reputations. In present on-line social network has become the most popular communication tool of people. They can publish, transfer and rate different contents in a social network. The boundary between content provider and customers becomes more and more partial, and each user has parts, content provider and content customer. [1] Once a user pastes a message, other users can read, comment, transfer, and add to favorite and degree it. These interactions between social network users are called as social activity of users. Social network are often to communicate with different user. For security reasons, it is necessary to build a reputation system to evaluate his/her behaviors. In present most of the user reputation systems use the rule-based method or the voting systems to calculate user reputations but in this paper using machine learning techniques to calculate the user reputation. [4]The creditability of information was defined by many words such as trustworthiness, acceptability, reliability, accuracy, fairness, objectivity, and other with the same concepts and definitions.

A critical part of the system is the assessment of the credibility of tweets and the reputations of the users who posted them rescore to represent the level of trustworthiness of the posted content. Use the term reputation score to represent the level of dependability of the user who posted the content. Users reputations are based on popularity measures. Describe a popular user as one who is recognized by other users on a similar network. The measures include the Follower-Rank and the Twitter Follower Followed ratio. In addition, consider replies and rewets a measures of a user's popularity. In this viewpoint, sentiment defines the degree of antagonism of any user Tweet that affects social relationships, the emotional states of other users, and their orientation with respect to the given topic. Propose a new reputation-based source credibility assessment method that introduces several new features into the existing models. Main approach users sentiments to identify and evaluate topically relevant and credible sources of information sentiment defines the degree of resentment of any user Tweet that affects social relationships, the emotional states of other users, and their orientation with respect to the given topic and also calculated the number of positive and negative words in a message, based on a predefined list of sentiment words. And also use the user popularity, the users social popularity score can be quantitatively evaluated using a simple algorithm that defines a user's popularity score based on certain features that are related to the users reputation Based on some topics, retweets are considered to be one of the best indicators of user popularity from the calculable perspective. [7]This suggests that a tweet that has been re-tweeted many times is considered to be attractive to the user. But, the most critical indicators of the popularity of the person who posts the tweet (the tweeter) are qualitative, such as the relationship between the user and the tweeter. In particular, measure the reputation or credibility of a Twitter user based on how popular he/she is, and how sentimental he/she is regarding a particular topic. In developing this approach, we have identified new features that can be used to find the most credible Twitter users [3].

4. FAKE NEWS DATA SETS

The following are widespread data-sets that have been used for fake news detection:

4.1 BuzzFeedNews

Come from the FakeNewsNet dataset, recently published by [4]; we used both the PolitiFact and BuzzFeed news sets they provide: the former contains a ground truth of 240 news (half labeled as fake, half labeled as real by the well-recognized fact-checking website BuzzFeedNews is a gathering of title and links to an actual story or a post that is considered fake news. This data-set is useful for testing Linguistic methods, nevertheless, multimedia content is not part of this data-set, therefore certain analysis are not possible on text-only data-set.[15]

4.2 PHEME

PHEME journalist-labeled dataset. PHEME is a curated data set of conversation threads about rumors in Twitter replete with journalist annotations for truth,This data-set includes rumor tweets, collected and annotated within the journalism use case of the project [19]. It contains Twitter conversations which are initiated by a rumor tweet. Also, it is linguistic based data-set. It contains about 330 conversations (297 in English and 33 Germany).[17]

4.3 CREDBANK

CREDBANK crowdsourced dataset. CREDBANK is a large-scale set of Twitter conversations about events and corresponding crowdsourced accuracy assessments for each event .The only data-set has limited social media data and allows users to perform investigation on Twitter data. This data-set signs off on all the classes except the visual is images video data. It errors out on having multimedia data, but still makes it a very compelling choice for researchers who are also focused on fake news detection on social media.[18]

4.3 LIAR

LIAR is a bench-marking background made available by University of California, Santa Barbara researchers. This data-set is also linguistic-based dataset and only contains text only data and has similar limitations like BuzzFeedNews data-set.[16]

5. CONCLUSIONS

This work demonstrates an automated system for detecting fake news in popular Twitter threads. Identifying misinformation is authoritative in online social media platforms, because information is circulated easily across the social media by unsupported sources. Automatically detect fake news using machine learning algorithm. Using the reputation-based technique to each users profile and calculating sentiment score established based on the users history. Their tweets also solve the problem of assessing information credibility on Twit-ter. The issue of information credibility has comes under scrutiny.

ACKNOWLEDGEMENT

This paper would not have been written without the valuable advices and encouragement of Mrs. S. S. Nandgaonkar, guide of ME Dissertation work. Authors special thanks go to all the professors of computer engineering department of VPKBIET, Baramati, for their support and for giving an opportunity to fake news detection using machine learning approach

REFERENCES

[1] Cody Buntain, Jennifer Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads". 2017 IEEE International Conference on Smart Cloud

[2] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).

[3] Mohammad Mehedi Hassan, Member, IEEE and Atif Alamri, "A Credibility Analysis System for Assessing Information onTwitter Member", IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING .

[4] Supanya Aphiwongsophon, Prabhas Chongstitvatana, "Detecting Fake News with Machine Learning Method", IEEE (2017).

[5] Shlok Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection", 2017 IEEE 15th Student Conference onResearch and Development (SCOReD).

[6] Juan Cao, Junbo Guo, Xirong Li, Zhiwei Jin, Han Guo, and Jintao Li, "Automatic Rumor Detection on Microblogs: A Survey", arXiv:1807.03505v1 [cs.SI] 10 Jul 2018.

[7] Kai Shu,Suhang Wang, Huan Liu, "Understanding User Profiles on Social Media for Fake News Detection",2018 IEEE Conference on Multimedia Information Processing and Retrieval

[8] Tanushree Mitra and Eric Gilbert, "CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations" ,School of Interactive Computing GVU Center Georgia Institute of Technology tmitra3, gilbert@cc.gatech.edu,Proceedings of the Ninth International AAAI Conference on Web and Social Media.

[9] Jacob Ross, Krishnaprasad Thirunarayan, "Features for Ranking Tweets Based on Credibility Newsworthiness "Kno.e.sis: Ohio Center of Excellence in Knowledge-enabled Computing Department of Computer Science and Engineering Wright State University Dayton, Ohio 45435 ross.138, t.k.prasad@wright.edu,2016 .

[10] Granik, M., Mesyura, V. 2017. "Fake news detection using naïve Bayes classifier". 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 900-903.

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

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective, KDD exploration newsletter", 2017.

[13] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, " A stylometric inquiry into hyperpartisan and fake news", arXiv preprint arXiv:1702.05638, 2017.

[14] https://en.wikipedia.org/wiki/Fake news website. Accessed Feb.6,2017.[]

[15] "Buzzfeednews: 2017-12-fake-news-top-50," https://github. com/BuzzFeedNews/2017-12-fake-news-top-50.

[16] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv: 1705.00648*, 2017.

[17] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PloS one*, vol. 11, no. 3, p. e0150989, 2016.

[18] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations." In *ICWSM*, 2015, pp. 258–267.

[19] Yassin M. Y. Hasan and Lina J. Karam, "Morphological Text Extraction from Images", IEEE Transactions On Image Processing, vol. 9, No. 11, 2000