IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 11, November 2019

Natural language Processing Based Fake News Detection using Text Content Analysis with LSTM

Geetanjali Jain[1], Ankur Mudgal[2]

Research Scholar, CSE Department, Shri Ram Institute of Science & Technology, Jabalpur, India[1]

Assistant Professor, CSE Department, Shri Ram Institute of Science & Technology, Jabalpur, India[2]

Abstract: False or unverified information spreads just like accurate information on the web, thus possibly going viral and influencing the public opinion and its decisions. Fake news and rumours represent the most popular forms of false and unverified information, respectively, and should be detected as soon as possible for avoiding their dramatic effects. The interest in effective detection techniques has been therefore growing very fast in the last years.
In this paper we propose the different approaches to automatic detection of *fake news* and *rumours*. In particular, we focus on five main aspects. First, we report and discuss the various definitions of fake news and rumours that have been considered in the literature. Second, we highlight how the collection of relevant data for performing fake news and rumours detection is problematic and we present the various approaches, which have been adopted to gather these data, as well as the publicly available datasets. Third, we describe the features that have been considered in fake news and rumour detection approaches. Fourth, we provide a comprehensive analysis on the various techniques used to perform rumour and fake news detection. Finally, we identify and discuss future directions.

Keywords: Fake news, Rumours, Natural language processing, Data mining, Text mining, Classification, Machine learning, Deep learning

I. INTRODUCTION

A lot of people use social media platforms not only to keep in touch with friends and family, but also to gather information and news from around the world. Thus, social media play a fundamental role in the news fruition. The case study for Britain reported in [1] shows an increase in the usage of social media, and more importantly their relevance to news consumption. According to Zubiaga et al. [2], social media have become a critical publishing tool for journalists [3, 4] and the main consumption method for citizens looking for the latest news [5]. Journalists may use social media to report on public opinions about breaking news stories, and even to discover potential new stories, whereas citizens may follow the development of breaking news and events through official channels (i.e. news outlets official accounts on social media platforms) or through posts of their own network (e.g. friends, family, public figures). Indeed, social networks have proved to be extremely useful especially during crisis situations, because of their inherent ability to spread breaking news much faster than traditional media [6]..Nowadays, social media facilitate the spread of the unverified and false information among a larger number of users, thus deeply influencing the global perception and the understanding of events [2]. Probably, one of the most striking examples of how fake news can influence opinions has been the U.S. presidential campaign in 2016. Authors in [7] thoroughly studied the subject, reporting interesting findings: during the campaign voters were exposed to higher number of pro-Trump than pro-Clinton articles. However, it is unclear how fake news can have been effective in influencing the final vote. An analysis performed through surveys has however shown that Republican voters were in general more inclined to believe in both real and fake news articles [7]. Thus, in this particular case, this analysis suggests that the influence of fake news on the final vote was relatively low. Nonetheless, we can argue that fake news and, more broadly, disinformation are becoming a huge problem on the web, and might have an important social cost in the future.For this reason, both social network platforms and the research community are very active in identifying potential fake claims and assessing their veracity. In this paper a machine learning based classifier for fake news detection has been explained.

II. RELATED WORK

According to the Global digital report 2019 [8] out of the world's total population of 7.676 billion, there are 4.388 billion internet users and 3.484 billion social media users. Almost half of the world's total population depends upon the internet for their knowledge. However, how much or up to what extent the circulated facts are verified is still a big question. How much we can rely on the information content that we are browsing every day. False information is

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019

created and initiated by a small number of people. People, relations, content and time are four critical dimensions of networked data analysed multi dimensionally by proposing an iOLAP framework based on polyadic factorization approach [9]. This framework handles all types of networked data such as microblogs, social bookmarking, user comments, and discussion platforms with an arbitrary number of dimensions. Origination, propagation, detection and Intervention are the four main facets of information pollution, which are diagrammatically represented in Figure 2.1.

Figure 2.1: Life Cycle of False Information System.

Origination deals with the creation of fake content by a single person, account or multipleaccounts. Propagation analyses the reason behind the fast and large-scale spread of fraudulentcontents online. The analysis is done by [10],[11] sheds new light on fake news writing style, linguistic features and fraudulent contentpropagation trends; concludes that falsehood disseminates significantly faster, deeper, fartherand more broadly than the truth in all the categories. False news was 70% more likely to beretweet by more number of unique users, as fake stories are more novel, surprising and eye-catching;attracts human attention hence encourages information sharing. Identification of themisinformation and disinformation from the massive volume of social media data usingdifferent Artificial Intelligence technologies comes under detection. Finally, interventionmethods concentrate to restrict the outspread of false information by spreading the truth.Fake product review is an emerging field of forgery in online social networks, specificallyin the field of e-commerce, as more and more people share their shopping experiences online through reviews [12]. The customer reviews directly related to thereputation of a product in the E-commerce era. People consider ratings, feedback reviews,and comments by previous buyers to make an opinion on whether to purchase a particularitem or not. The algorithms suggested in [13][14] for detecting fake movie reviews are based on sentimentanalysis, temporal, statistical features and text classification. Ahmed et al. [14] use six supervised machine learning classifiers SVM, LSVM, KNN, DT, SGD, LR to detect fake reviews of hotels and fake news articles on the web using textclassification. Their experiments achieve a significant accuracy of 90% and 92%respectively. Different content-based, features based, behavior-based and graph-based approaches can be used to detect opinion spams present in differentformats of fake reviews, fake comments, social network posting and fake messages. Inaddition to the mainstream news media; there is also a concept of alternative mediathat aims to just present the facts and let readers use their critical thinking to explorereality by means of discussions.

The most recent work on the text in the field of fake news detection is given as follows: [15] assess the problem related to information credibility on Twitter. They have proposed an automated classification system, including four major components:

1) The reputation based technique,
2) A credibility classifier engine,
3) A user experience component, and
4) A feature rank algorithm.

Novelty and pseudo feedback (PF) based features have been introduced by [16] to detect rumours on early basis, along with features based on the presence of several URLs, hash-tags and user-names, POS tags, punctuation characters as well as eight different categoriesof sentiment and punctuation emotions. Many authors have worked on veracity classification task. [17] Introduced three sets of features related to linguistic, user-oriented, and temporalpropagation. The Twitter dataset has been used for evaluation. Study reveals that the bestperforming features were those in the temporal category. Sarcasm is also one of the crucial issuesover social media. M. Bouazi*et al.* [18] assessed the problem related to sarcasm on twitter usingpattern-based approach and introduced four sets of features that cover the different type of sarcasmand classified tweets as sarcastic and non-sarcastic.Social media is an open community where anyone can create their content, without any checkon its veracity. Also, data present on social media is highly

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019

heterogeneous. Though, manycredible sources are there whose integrity cannot be questioned and the content produced bythem is verified and double checked. Inspired by these ideas, we exploit this property in our work

Support Vector Machines (SVMs) are one of the most widely used machine learning method for classification in a number of research areas. SVMs are discriminative classifiers formally defined by a separating hyperplane. According to the experiments in [19], SVMs have outperformed a number of supervised machine learning approaches for deception detection in text, obtaining an F-measure F 1 of 0.84. However, as pointed out by the authors themselves, there exists a significant variation in performance depending on the dataset selected for training [18].

Content-based features (e.g. linguistic and visual features) were exploited in most SVM-based approaches to fake news and deception detection. In particular, Afroz et al. [20]has obtained highly competitive scores for the task of deception detection on a number of datasets by exploiting only lexical, syntactic, and content-specific features. Rubin et al. [21] has trained an SVM for satirical fake news detection with a number of content-based features, obtaining an F 1 of 0.87. As regards rumour related tasks (i.e. detection, verification etc.), SVM-based approaches have made a more prominent use of context-based features as well as content-based ones.

SVMs were also exploited for the task of clickbaitdetection. Another widely studied family of algorithms, proposed particularly for rumour analysis tasks, is decision tree [22]. Decision tree performs a recursive split on feature values in order to determine the class. Decision trees are generated from data with algorithms such as J48 (C4.5) [20]. Despite their relative simplicity with respect to other machine learning schemes, they exhibit competitive performance on the task at hand.

III. PROPOSED SYSTEM

Main proposed work is to develop a machine learning method to identify fake/unreliablenews based on content acquired. Figure below represents overall architecture of the system.

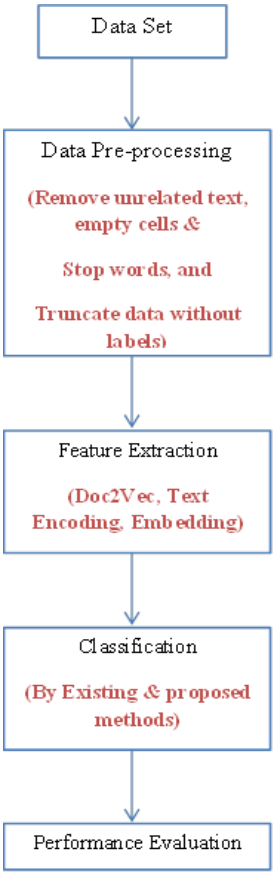


Figure 3.1: Architecture of Proposed System

Data Pre-processing

Text data requires special preprocessing to implement machine learning or deep learning algorithms on them. There are various techniques widely used to convert text data into a form that is ready for modeling. We start with removing stop words from the text data available. Stops Words (most common words in a language which do not provide much context) can be processed and filtered from the text as they are more common and hold less useful information. Stop

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019

words acts more like a connecting part of the sentences, for example, conjunctions like “and”, “or” and “but”, prepositions like “of”, “in”, “from”, “to”, etc. and the articles “a”, “an”, and “the”. Such stop words which are of less importance may take up valuable processing time, and hence removing stop words as a part of data preprocessing is a key first step in natural language processing. Preparing the text from the body and headline of the news article for modeling is quite challenging. To perform text analytics, we need to convert raw text into numerical features.

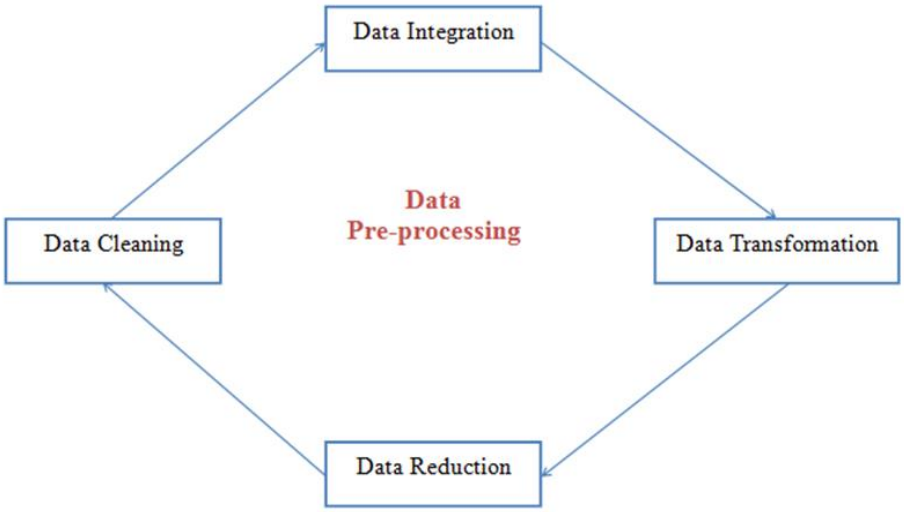


Figure 3.2: Data cleaning steps.

The embeddings used for the majority of our modelling are generated using the Doc2Vec model. The goal is to produce a vector representation of each article. Before applying Doc2Vec, we perform some basic pre-processing of the data. This includes removing stop words, deleting special characters and punctuation, and converting all text to lowercase. This produces a comma-separated list of words, which can be input into the Doc2Vec algorithm to produce a 300-length embedding vector for each article.

Doc2Vec is a model developed in 2014 based on the existing Word2Vec model, which generates vector representations for words. Word2Vec represents documents by combining the vectors of the individual words, but in doing so it loses all word order information. Doc2Vec expands on Word2Vec by adding a”document vector” to the output representation, which contains some information about the document as a whole, and allows the model to learn some information about word order. Preservation of word order information makes Doc2Vec useful for our application, as we are aiming to detect subtle differences between text documents.

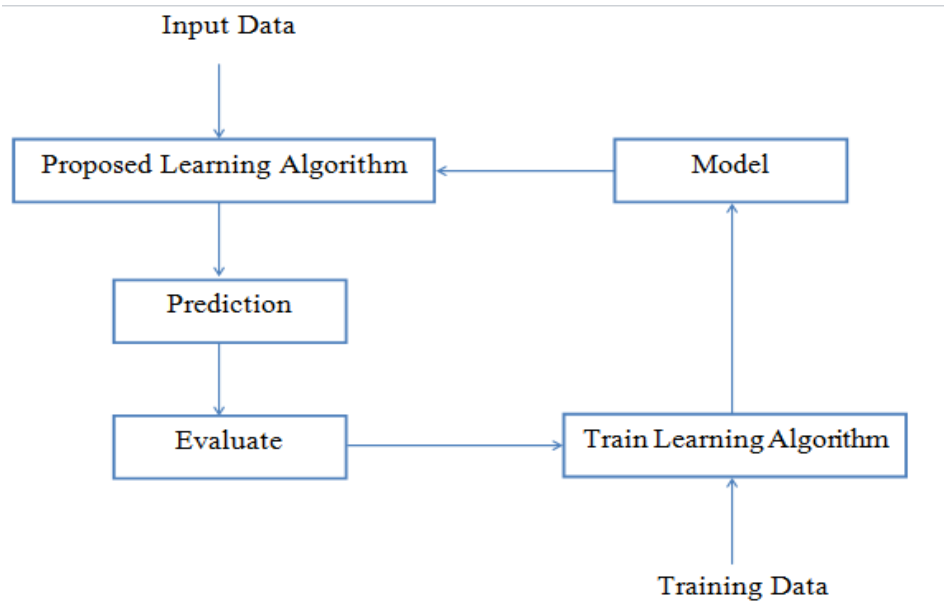
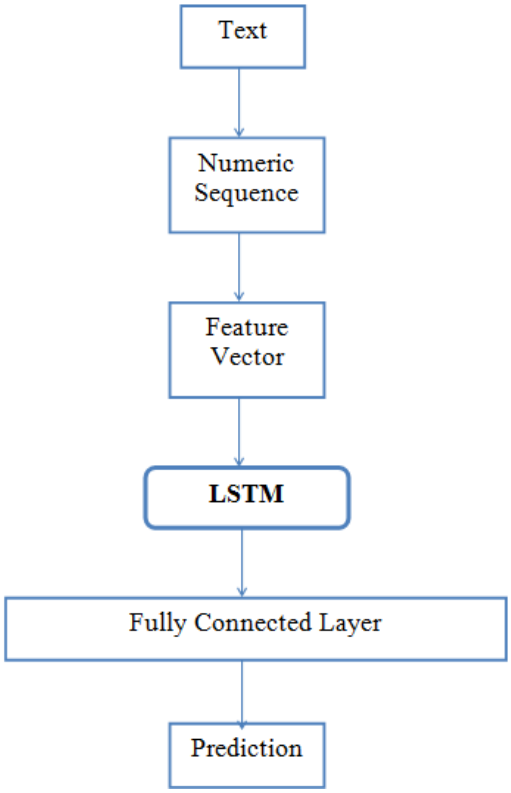


Figure 3.3: Proposed Learning Algorithm.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019



The Long-Short Term Memory (LSTM) unit was proposed by Hochreiter and Schmidhuber [8]. It is good at classifying serialized objects because it will selectively memorize the previous input and use that, together with the current input, to make prediction. The news content (text) in our problem is inherently serialized. The order of the words carries the important information of the sentence. So the LSTM model suits for our problem.

Since the order of the words is important for the LSTM unit, we cannot use the Doc2Vec for preprocessing because it will transfer the entire document into one vector and lose the order information. To prevent that, we use the word embedding instead. We first clean the text data by removing all characters which are not letters or numbers. Then we count the frequency of each word appeared in our training dataset to find 5000 most common words and give each one an unique integer ID. For example, the most common word will have ID 0, and the second most common one will have 1, etc. After that we replace each common word with its assigned ID and delete all uncommon words. Notice that the 5000 most common words cover the most of the text, as shown in Figure 1, so we only lose little information but transfer the string to a list of integers. Since the LSTM unit requires a fixed input vector length, we truncate the list longer than 500 numbers because more than half of the news is longer than 500 words. Then for those lists shorter than 500 words, we pad 0's at the beginning of the list. We also delete the data with only a few words since they don't carry enough information for training. By doing this, we transfer the original text string to a fixed length integer vector while preserving the words order information. Finally we use word embedding to transfer each word ID to a 32-dimension vector. The word embedding will train each word vector based on word similarity. If two words frequently appear together in the text, they are thought to be more similar and the distance of their corresponding vectors is small.

The pre-processing transfers each news in raw text into a fixed size matrix. Then we feed the processed training data into the LSTM unit to train the model. The LSTM is still a neural network. But different from the fully connected neural network, it has cycle in the neuron connections. So the previous state (or memory) of the LSTM unit Ct will play a role in new prediction ht.

$$ht = ot .\tanh(ct)$$
$$ct = ft.ct\text{-}1 + it .ct$$

During the training of RNN, as the information goes in loop again and again which results in very large updates to neural network model weights. This is due to the accumulation of error gradients during an update and hence, results in an unstable network. At an extreme, the values of weights can become so large as to overflow and result in NaN values. The explosion occurs through exponential growth by repeatedly multiplying gradients through the network layers that have values larger than 1 or vanishing occurs if the values are less than 1.

The above drawback of RNN pushed the scientists to develop and invent a new variant of the RNN model, called Long Short Term Memory. LSTM can solve this problem, because it uses gates to control the memorizing process.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 11, November 2019

IV. RESULT AND EVALUATION

The datasets used for this project were drawn from Kaggle [3]. The training dataset has about 16600 rows of data from various articles on the internet. We had to do quite a bit of pre-processing of the data, as is evident from our source code, in order to train our models.
A full training dataset has the following attributes:
1. id: unique id for a news article
2. title: the title of a news article
3. author: author of the news article
4. text: the text of the article; incomplete in some cases
5. label: a label that marks the article as potentially unreliable1: unreliable 0: reliable

Table 1 showing accuracy on first run.

Classifier	Accuracy (in %)
Naive Bayes	71.61
Neural Network with TensorFlow	81.94
SVM	88.51
Proposed	**94.53**

Most of the data-sets previously used to construct the classification model only comprised of the malware binaries. In this thesis we are also considering the ASM file generated via the IDA disassembler tool, giving us more ground to find patterns. Feature extraction in the existing models is based on manual selection of unique features which were identified by ”looking” at the binary files. Most of the models have not employed a systematic machine learning approach to extract the frequently occurring features. The few who have, have intuitively applied the n-gram model, with n = 2, 4, 10 etc. The selection of n did not have any machine learning basis. We use frequent item-set approach which helped us construct a solid framework to use various classification algorithms.

Our modelling uncovered that all our features could be segregated as either sparse or denseFeatures:
- A large number of sparse features degraded the performance of most of the models since a lot of non-relevant features were being used for model construction.
- The innate randomness of curating dense nodes from subset of features gives Proposed Algorithm the robustness against over-fitting. Since the model is created using through dense features it gave the best performance even before feature selection.
- Apart from proposed classifier, all the other classifiers did not perform well before feature selection.

We observed that LSTM gave us the best results. We had to use a different set of embeddings for preprocessing the data to be fed to our LSTM model. It uses ordered set of Word2Vec representations. The LSTM achieves the highest F1 score in comparison to all the other models, followed by the Neural Network model using Keras. One of the reasons that LSTM performs so well is because the text is inherently a serialized object. All the other models use the Doc2Vec to get their feature vectors and hence, they rely on the Doc2Vec to extract the order information and perform a classification on it. On the other hand, the LSTM model preserves the order using a different pre-processing method and makes prediction based on both the words and their order. This is how it outperforms others.

V. CONCLUSION

A lot of work has been done in the past years to make online content more reliable and trustful and some of the key areas remain unaddressed. Quick and real-time detection of the source is useful to control the spread of false information and reduce the adverse impact on society. Real-time collected datasets, automatic detection of rumors and finding its original source is a challenging issue.Information pollution, fake news, rumours, misinformation, disinformation has become a by-product of the digital communication ecosystem, which proves to be very dangerous. Approximately 40% of the studied research concentrated on detection of false content using machine learning and deep learning implicit as well as explicit feature engineering and pattern analysis techniques. Finally, open issues and challenges are also highlighted to further explore potential research opportunities. This work may be helpful to the new researchers to understand the different components of digital online communication from a social and technical perspective. Multilingual cross platform fake news spreading, complex and dynamic network structure, huge volumes of unlabelled real-time data and early detection of rumors are some challenging issues that are still unaddressed and need further research. Improving the reliability and future of online information ecosystem is a joint responsibility of the social community, digital policymakers, administration, technical and research scholars.

REFERENCES

[1]. N. Newman, W.H. Dutton, G. Blank, Social media in the changing ecology of news: the fourth and fifth estates in Britain, Int. J. Internet Sci. 7 (1) (2012) 6–22.
[2]. Zubiaga, A. Aker , K. Bontcheva , M. Liakata , R. Procter , Detection and resolution of rumours in social media: a survey, ACM Comput. Surv. 51 (2) (2018) 32:1–32:36.
[3]. N. Diakopoulos, M. De Choudhary, M. Naaman, Finding and assessing social media information sources in the context of journalism, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 2451–2460.
[4]. P. Tolmie , R. Procter , D.W. Randall , M. Rouncefield , C. Burger , G. Wong Sak Hoi , A. Zubiaga , M. Liakata , Supporting the use of user generated content in journalistic practice, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, 2017, pp. 3632–3644 .
[5]. Hermida, Twittering the news: the emergence of ambient journalism, Journal. Pract. 4 (3) (2010) 297–308.
[6]. S. Vieweg, Micro blogged contributions to the emergency arena: discovery, interpretation and implications, in: Proceedings of the ACM Conference on Computer Supported Cooperative Work, ACM, 2010, pp. 241–250.
[7]. H. Allcott , M. Gentzkow , Social media and fake news in the 2016 election, Technical Report, National Bureau of Economic Research, 2017
[8]. Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). Reuter Institute DigitalNews Report 2018. https://doi.org/10.2139/ssrn.2619576.
[9]. Chi, Y., Zhu, S., Hino, K., Gong, Y., & Zhang, Y. (2009). iOLAP: A framework for analyzing the internet,social networks, and other networked data. IEEE Transactions on Multimedia, 11(3), 372–382.https://doi.org/10.1109/TMM.2009.2012912.
[10]. Vosoughi, S., Deb, R., & Aral, S. (2018). The Spread of True and False News Online. Science, 359(6380), 1146–1151.
[11]. Horne, B. D., &Adali, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, RepetitiveContent in Text Body, More Similar to Satire than Real News. Eleventh International AAAI Conferenceon Web and Social Media, 759–766.
[12]. Martens, D., &Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. EmpiricalSoftware Engineering, 1–40. https://doi.org/10.1007/s10664-019-09706-9.
[13]. Elmurngi, E., &Gherbi, A. (2017b). Detecting fake reviews through sentiment analysis using machine learningtechniques. Sixth International Conference on Data Analytics, 65–72.
[14]. Elmurngi, E., &Gherbi, A. (2017a). An Empirical Study on Detecting Fake Reviews Using Machine LearningTechniques. IEEE Seventh International Conference on Innovative Computing Technology (INTECH), 107–114.
[15]. Ahmed, H., Traore, I., &Saad, S. (2017). Detecting opinion spams and fake news using text classification.Security and Privacy, 1(1), e9. https://doi.org/10.1002/spy2.9.
[16]. Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y. R., & Collins, C. (2014). #FluxFlow: Visual analysis of anomalousinformation spreading on social media. IEEE Transactions on Visualization and Computer Graphics,20(12), 1773–1782. https://doi.org/10.1109/TVCG.2014.2346922.
[17]. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., &Hovy, E. (2016). Hierarchical Attention Networks forDocument Classification. 2016 Conference of the North American Chapter of the Association forComputational Linguistics: Human Language Technologies, 1480–1489.
[18]. Shelke, S., & Attar, V. (2019), Source detection of rumor in social network – A review. Online Social Networksand Media, 9, 30–42. https://doi.org/10.1016/J.OSNEM.2018.12.001.
[19]. M. Bouazizi and T. O. Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," IEEE Access,vol. 4, pp. 5477-5488, 2016.
[20]. K. Shu , A. Sliva , S. Wang , J. Tang , H. Liu , Fake news detection on social media: a data mining perspective, ACM SIGKDD Explore Newslett. 19 (1) (2017)22–36.
[21]. S. Afroz , M. Brennan , R. Greenstadt , Detecting hoaxes, frauds, and deception in writing style online, in: Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, DC, USA, 2012, pp. 461–475 .
[22]. V.L. Rubin, Y. Chen, N.J. Conroy, Deception detection for news: three types of fakes 52 (1) (2015) 1–4.
[23]. J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI, address, 2016, pp. 3818,3824