Fake News Detection using Data Mining Techniques

S. Sunil Kumar Aithal[1], Krishna Prasad Rao[2], R. P. Puneeth[3]

[1,2,3]Assistant Professor, Department of Computer Science Engineering, NMAM Institute of Technology, Karnataka, Nitte, India

E-mail: [1]sunilaithal@nitte.edu.in, [2]krishnaprasad@nitte.edu.in, [3]puneeth.rp@nitte.edu.in

Abstract

Nowadays, internet has been well known as an information source where the information might be real or fake. Fake news over the web exist since several years. The main challenge is to detect the truthfulness of the news. The motive behind writing and publishing the fake news is to mislead the people. It causes damage to an agency, entity or person. This paper aims to detect fake news using semantic search.

Keywords: Real news, Fake news, Twitter API Keys, Semantic Search.

1. Introduction

Fake news exist from the time when the news began to circulate widely. Fake news are found in fake news websites, traditional news and in social media and such news are presented to look like real news. These news cause negative impact on an individual and also on the society. These news, apart from misleading people, it also affects their decisions. It changes the perception of the people towards the news. The truthfulness of the entire news ecosystem is damaged due to the fake news. Many researchers have already worked on this topic but complete solution is not obtained yet. Nowadays, spreading of fake news increases. Therefore, a method to detect the truthfulness of the news is most needed. For this, a concept called Semantic search is used. The concept of language semantics is the base for Semantic search. Semantic search is different than the traditional search algorithms. It relies on the factors such as substance, intent, context and meaning of the phrase searched. It also includes synonyms, location, word variations, current trends and several other natural language components in the search process.

DOI: https://doi.org/10.36548/jtcsst.2021.4.002
Received: 14.11.2021, received in revised form: 03.12.2021, accepted: 21.12.2021, published: 31.12.2021

2. Related Works

2.1 A survey on detecting media rich fake news

Nowadays, it is a trend among huge number of the humans to get trending news from social sites and other online sources. Most of the time, people are unaware about the authenticity of news from online sources. Lack of awareness of web functionality among the people, additionally facilitates the spread of the fake news and stories. Social media sites emerge as an important tool for the spread of the fake or false stories. In order to mislead the readers, fake news are intentionally created. The motive behind creating the fake or false news is predominantly for attracting the consumers usually by gaining the benefits of financial or economic or political reasons [1,13].

2.2 Text classification based detection of fake or false news and opinion spams

Fake reviews and fake news are called as opinion spams. Both of them contains the writings and spreading fake information or beliefs. Nowadays, it is easy for anyone to write false news or reviews on the internet. Today’s critical challenge is that there is a need for a structured way to differentiate between the fake review and the real one. Hence, a new model called ‘n-gram model’ is proposed for detecting the fake news and reviews automatically. This model uses six machine learning classification techniques - SVM, stochastic gradient descent, Linear SVM, LR, K-nearest neighbor and decision tree [2,14].

2.3 A data mining perspective on fake or false news detection on social sites

In this article, the problems of false or fake news are explored in two steps namely characterization step and tracking or detection step. In the first step (a characterization phase), the fundamental principles and concepts of fake or false news in social media sites as well as traditional media are introduced. In the second step (detection phase), the various pre-existing fake or false news tracking or detection techniques are reviewed from the perspective of data mining, comprising model construction and feature extraction [3,15].

2.4 An overview of semantic search with meanings systems

Using semantic web, acquisition of information, processing, storage and retrieval are performed to optimize the performance of conventional information search and retrieval methods. Traditional search techniques are developed on the basis of modelling the number of words and its computation and performing link analysis for further improvement. In this article, authors establishes the objectives, methodologies and associated functionalities and prepares six categories of research on semantic search [16]. They are Semantic Analytics, Mining-based Searching, Knowledge-based and Entity-based Search, Document- oriented Search, Multi-media Data or Information Search and Relationship-oriented Search [4].

2.5 Detection of fake or false news using naive bayes classifier approach

The naive Bayes classifier is used for detecting the fake news in this approach. For testing purpose, Facebook's news feed data set is identified and implemented as software systems. Using this classification approach an accuracy of nearly 74% is achieved on the test data set which is considered as a decent result. Further the article describes on how to improve the accuracy in several ways [17]. This article concludes using artificial intelligence techniques, and observing the obtained results, fake news detection problem is solved [5].

2.6 Survey of semantic based search research

Semantic based search research is performed is in work. Search is performed using semantic based approaches and also based on semantic contents annotation. This article finds and interprets various common research aspects in semantic search, and also extracts common methods used in them. The five aspects presented are: augmenting semantic techniques with traditional keyword search, concept of location, complex query constraints, solving problems and discovering the path connections [6, 18].

2.7 Semantic and Text Based

In this proposal, a Semantic based search which is "search the text with its associated meaning" has been used. Semantic search finds applications for different perspectives of real-world problem and utilizes for research involving different communities. It is two dimensional based survey which considers the type of collected data and the type of search. Type of data can be a text, knowledge basis or both and the search can be keyword-based or structured or natural language based. The article focuses on the basic principles, concrete systems, and benchmarks. Advanced topics like inference, indexing, ranking, ontology match and merge operations are taken into consideration [19]. NLP techniques such as keyword based search, sentence parsing, named entity recognition and POS tagging are discussed. Keyword searching fails to give accurate results as they are unable to interpret the exact meaning for the searched keywords. In this article, keyword based and semantic based techniques are compared. The advanced web technology correspond to the semantic web. Semantic web interprets and better recognizes the information both for humans and the machines. The main and valuable component of semantic web is an Ontology. Nowadays, the existing web are being enhanced from only machine readable form to highly machine-understandable version. Ontology plays a significant role in the intuition of the Semantic Web. It provides the huge semantic vocabulary, which can be used for annotating the websites in order to achieve more machine understandable information. It provides an explicit information and knowledge about the system or an application using which runtime interactions and information sharing for other systems can be achieved [7].

2.8 Semantic search Engines - a comparative study

The recent World Wide Web also known as Web2.0 is a huge repository of interlinked documents that are transferred from one computer to another and are presented to people. Search engine plays an important role in discovering any information needed by the people from www. In spite of having lots of research and development in the search engine techniques, they are still a syntax based search i.e. the search results they display is based on the keyword matching without understanding the semantics of the query and thus resulting in the production of list of Webpage links having more of unrelated links as an output [20]. The future of World Wide Web also called as Web3.0 (Semantic Web), is being developed. The main intention of Web3.0 is to lessen the problems faced in Web2.0 by representing data in a structured way and to discover such data using Semantic Search Engines (SSE) that are being developed across various domains. This article provides a survey on architectures of some of the common SSEs and also presents a comparative study on the basis of techniques used for crawling, indexing, reasoning, etc. [8]

2.9 Semantic based Search Engines and Approaches utilized for Semantic based Searching – A study

In this busy era, everyone wants accurate answers to their queries which can only be provided by Semantic Search Engines (SSE). Semantic based search engines are more accurate than the traditional searching engines. With a huge increase of voluminous data, the number of answers provided by the conventional search based engines have been improved and increased to satisfy all the users. This makes the user difficult in selecting the desired answer out of the large number of answers. Because of this, day by day the need for developing semantic based search engine increases. The most suitable answer for the user queries will be provided by the Semantic search engines. The semantics of the collected words that are typed along with the queries are not known by the traditional search engines. They just perform searching based on the keywords. Therefore, the semantic based searching engines are far more superior to the old traditional search based engines yielding us well defined and meaningful results [9].

2.10 Comparative study of semantic and keyword based search engine

As the Keyword based Search engines could not interpret the actual meaning or semantics of the words that are used in the search, they are unable to give relevant results. This article provides comparison between the keyword and semantic search engines. Semantic Web can be considered to be the future of the current web [21]. It depicts an information in a structured and better understandable way for humans as well as for machines. Ontology is a very important component of the Semantic web. Using Ontology, the current Web shifts from being machine readable scenario to machine understandable scenario. To annotate semantic and to identify comprehensible foundation for the Semantic Web resources, Ontology is the best technique [10].

2.11 Semantic Based Search

Establishing the distributed web, which provides machine with easily understandable data web services as well as semantic web, plays an important role. In this article, an application is built utilizing these technologies which is called as called Semantic based Search and focuses on improving the traditional web based searching. An outline of TAP framework, the domain application platform upon which the Semantic based Search is developed. The implemented Semantic based Search system is based on the type of the search query that supports the augmentation of the search results [11].

2.12 Modified framework for semantic based search engine

This system based on the data mining saves the data which is in machine readable form and facilitates the intelligent matching of the data with other related data. This engine is used in various large companies, but in order to perceive the maximum usefulness of the semantic web, it needs to be made more effective. Semantic web technology is highly dependent on a number of various components such as a good UI, an optimized query language and its processor, an output optimizer, ranking of the results and the appropriate selection of data structure for data storage. Thus, for prioritizing these components and optimizing the experience of the search over the web, various strategies are established [12].

In this section, taking news from twitter by applying data mining technique is discussed. Data Mining is the process of identifying the information which are useful taken from large data repositories. The main objective of data mining approach is to retrieve information from huge datasets and transforms them into useful information. In order to take news from twitter four API keys (access token, consumer key, access secret and consumer secret) are required. By creating twitter account and developer account, the API keys are obtained. Twitter data is accessed using API keys. Advantages of taking news from Twitter are that the news are reliable and people get updated with the current and trending news very fast. Twitter provides a variety of news and it also provides free API Keys. The data are stored in a file for further analysis. Once the data are collected, the next step is to pre-process it which is preliminary step for all Natural Language Processing techniques. Necessary steps that are required for pre-processing activity is based on the type of the application and targeted requirements. In this model, pre-processing of gathered tweets are performed. The standard Twitter message size is 140 characters which includes hash tags, user-mention, and URLs which has to be filtered using regular expression.

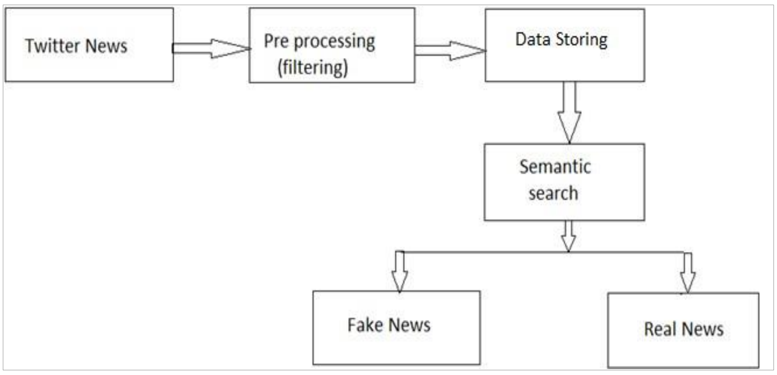

Figure 1. Extracting news from the Twitter

3. Proposed Work

Semantic search can be used to identify similarity between the user input and the data present inside the file which gets updated dynamically. It is a popular data searching technique. In this approach, the search query not only focuses on finding the trending keywords, but also it determines the purpose and background meaning or semantics of the words a person has used for searching. Semantic based search yields more significant search results. It evaluates and understands the search text and finds the most appropriate results in a particular website, data repository or database. The intention of the semantic based search is to go beyond the conventional dictionary word meaning or a phrase to recognize the searcher’s query target within a specific context. By using the concept of synonyms, matching, and natural language processing technique, semantic based search provides more interesting search results by modifying structured and unstructured data into an inherent and responsive data.

4. Results and Discussion

The Architecture diagram illustrates the overall structure and components associated with the system. The data are collected from social media like twitter. After the data collection, it starts with pre-processing unnecessary characters, and then URL are removed from the data. Hybrid CNN models can be used on the collected dataset. This helps to minimize the total size of the original data by eliminating the unrelated and unnecessary information that is already present in the gathered data. Using the concept called semantic search, the news entered by the user is real or fake can be identified by understanding the meaning of searcher's intent.

In this system, news from twitter are taken using Twitter APIs keys. When the application executes every time, current news from twitter gets loaded, and this news is considered as real news. Then, the pre-processing of news using regular expression by removing all URL, hashtag and special character that are present in the news is performed and are stored in a file. Any amount of news can be stored based on the memory size of the system. At first, a new user must register by filling required details before searching for the news. This user information will get stored in the database, and after successful login, user can search for the news. When user searches for the news, news present in the file one by one are considered. Each sentence is then tokenized into words using word tokenize. Tokenization of words is also known as word segmentation. It is the process of segmenting the textual sentence which is written using natural language into its equivalent individual words. Once the sentence is tokenized, stop words are looked for in it. During the pre-processing and post-processing of the text, the stop words present are filtered out. Examples of the most appearing stop words are "and", "a", "the", "an" occurring in natural language, however there is no complete set of defined stop words. Depending on the type of application, the set of possible stop words may change. Stop words may add plenty of noise. Therefore, stop words are removed. There may be words with the similar meaning. For grammatical reasons, the derivationally related forms and inflectional forms of a word in the sentence are reduced to a general common base form. This process is known as Lemmatization. It actually refers to proper use of a morphological analysis and vocabulary of a word. The same procedure is followed for the user's search sentence and compared with the news which has to be verifyied. The similarity between sentences are determined by considering the synonyms of each word. If the meaning of the news entered by the user matches more than 80% with the information stored in file, it is considered as real news or otherwise fake.

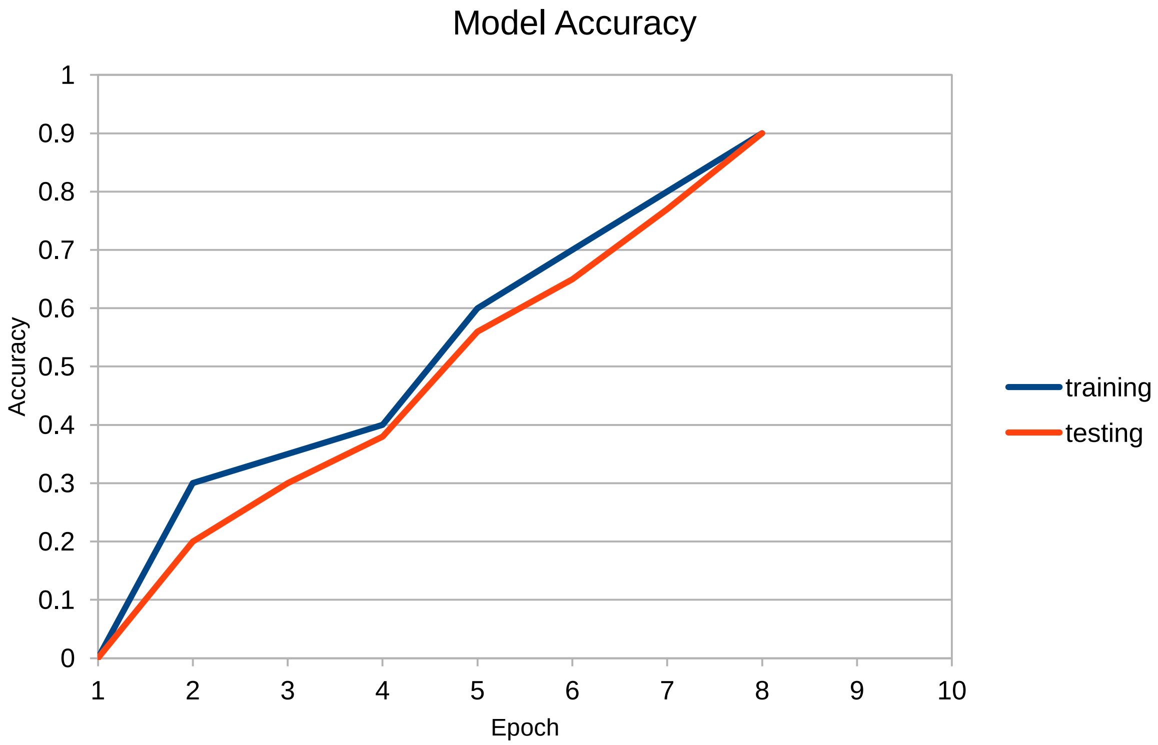

Figure 2. Statistical inference of Hybrid CNN model

If any user likes a particular post, and if he wants to tweet it in his account, then he can compose tweet using this application. The user must have Twitter APIs keys. This will be displayed in his personal twitter account.

A paired t-test was incorporated to verify the statistical significance of the obtained results. The experiments were repeated several times (using 5-fold cross validation, i.e. 80%-20% split) on twitter dataset; and accuracy was reported at 90% confidence intervals as can be observed from Figure 2.

5. Conclusion

Fake news, is one of the biggest modern day problem that has the capacity to influence decisions and opinions. Fake news detection challenge is a global issue. It has already reached most of the people through social networking sites. Prevention is not a solution. But detection can definitely solve the problem. Hence it is very important to develop a system that can help users in detecting the news as fake or real. Here, data mining techniques and semantic search concept are used. Tweets from twitter using Twitter API Keys are collected and once the tweets get loaded, they are pre-processed. Using this system, the fake news can be detected and hence the further circulation of fake news which misleads people can be avoided. This project can be improved by considering the news from different sources. Some related news can be displayed along with the output in the future study.

References

[1] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441, doi: 10.1109/MIPR.2018.00093.

[2] Traore, Issa & Saad, Sherif. (2017). Detecting opinion spams and fake news using text classification. Security and Privacy. 1. e9. 10.1002/spy2.9.

[3] Kai Shu, Suhang Wang, Amy Sliva, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. arXiv preprint arXiv:1708.01967, 2017

[4] Wei, Wang & Barnaghi, Payam & Bargiela, Andrzej. (2008). Search with meanings: An overview of semantic search systems. International Journal of Communications of SIWN. 3.

[5] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[6] Mäkelä, Eetu. (2008). Survey of Semantic Search Research.

[7] H. Bast, B. Buchhold, and E. Haussmann. Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval, 10(2-3):119–271, 2016. URL https://doi.org/10.1561/1500000032.

[8] Jain, Ranjna & Duhan, Neelam & Sharma, Ashok. (2015). Comparative Study on Semantic Search Engines. International Journal of Computer Applications. 131. 4-11. 10.5120/ijca2015907370.

[9] Junaid Rashid, and M. W. Nisar, "A Study on Semantic Searching, Semantic Search Engines and Technologies Used for Semantic Search Engines," International Journal of Information Technology and Computer Science (IJITCS), vol. 8, no. 10, pp. 82-89, 2016. doi:http://dx.doi.org/10.5815/ijitcs.2016.10.10

[10] Malve, Ankita & Chawan, Pramila. (2015). A Comparative Study of Keyword and Semantic based Search Engine. 4. 10.15680/IJIRSET.2015.0411039.

[11] Gilda, Shlok. (2017). Evaluating machine learning algorithms for fake news detection. 110-115. 10.1109/SCORED.2017.8305411.

[12] El-gayar, M.M. & Mekky, Nagham & Atwan, Ahmed. (2015). Efficient Proposed Framework for Semantic Search Engine using New Semantic Ranking Algorithm. International Journal of Advanced Computer Science and Applications. 6. 10.14569/IJACSA.2015.060818.

[13] Miraj Patel, Detection of Maliciously Authored News Articles, December 2017

[14] Tripathi, Milan. "Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM." Journal of Artificial Intelligence 3, no. 03 (2021): 151-168.

[15] Pandian, A. Pasumpon. "Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 123-134.

[16] Anand, C. "Comparison of Stock Price Prediction Models using Pre-trained Neural Networks." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 3, no. 02 (2021): 122-134.

[17] Karthigaikumar, P. "Industrial Quality Prediction System through Data Mining Algorithm." Journal of Electronics and Informatics 3, no. 2 (2021): 126-137.

[18] Kumar, Sanjay, Ryan Bansal, and Raghav Mehta. "A Study of Blending Ensembles for Detecting Bots on Twitter." In Innovative Data Communication Technologies and Application, pp. 29-40. Springer, Singapore, 2021.

[19] Viloria, Amelec, Nelson Alberto, and Carlos Alberto Jiménez Cabarcas. "Bots, Internet of Things and Threats to Personal Data in the Technological Era." In Proceedings of International Conference on Intelligent Computing, Information and Control Systems, pp. 591-599. Springer, Singapore, 2021.

[20] Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis, Fake news detection: A hybrid CNN-RNN based deep learning approach, International Journal of Information Management Data Insights, Volume 1, Issue 1, 2021, 100007, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2020.100007.

[21] T. Hamdi, H. Slimi, I. Bounhas, Y. Slimani, A hybrid approach for fake news detection in twitter based on user features and graph embedding, International conference on distributed computing and internet technology, Springer (2020), pp. 266-280.

Author’s biography

S. Sunil Kumar Aithal works as Assistant Professor in the Department of CSE in NMAM of Technology in Karnataka, India, his area of research includes Data Science, Image and Computer Vision.

Krishna Prasad Rao works as Assistant Professor in the Department of CSE in NMAM of Technology in Karnataka, India, his area of research includes Data Science, Blockchain.

R. P. Puneeth works as Assistant Professor in the Department of CSE in NMAM of Technology in Karnataka, India, his area of research includes Data Science, Blockchain.