

A Survey on Role of Machine Learning and NLP in Fake News Detection on Social Media

Chetan Agrawal
Department of CSE
UIT-RGPV
Bhopal, India

chetan.agrawal12@gmail.com

Anjana Pandey
Department of IT
UIT-RGPV
Bhopal, India

anjanapandey@rgtu.net

Sachin Goyal
Department of IT
UIT-RGPV
Bhopal, India

sachingoyal@rgtu.net

Abstract—Due to the enormous and exponential advancement in the online social network, the triad of Facebook, Twitter and Whatsapp posed a great challenge in the form of fake news in front of us. In recent years many events like false propaganda of the ‘US presidential election’, opinion spamming in ‘Brexit referendum’, and long-tail series of viral rumors after many natural calamities around the world, created a lot of chaos and law and order problem. Simultaneously, this rapid explosion of fake news also attracted the attention of different researchers to investigate the real cause of it and thus to developed some tools and techniques to relieve and discover the Rumors across online media as soon as possible. In this regard, the Machine Learning (ML) algorithms and Natural Language Processing (NLP) algorithms emerged as the remarkably vital and essential tool to detect fake news in the current age. NLP when aided with machine learning produced many remarkable results that were possible just by manual fact-checking or by normal text detection process. We have systematically discussed the role of NLP and machine learning in the fake news detection process, and various detection techniques based on these. Basic terminology of NLP and machine learning too explained in brief. At last, we gave light on the future trends, open issues, challenges, and potential research oriented toward NLP and ML-based approaches.

Keywords— *Fake news detection, Machine Learning, NLP, Social Media*

I. INTRODUCTION

The present era witnessed rapid advancement and drastic evolution of telecommunication technology and with the fusion of online streaming multimedia sites, the fastest internet technology made Facebook, Twitter and Whatsapp are the biggest players of online social networking. The 24*7 availability of online social networking websites generates each day a huge amount of data that is both essential as well as a critical fusion for personal life. Social media posts are highly rich in content and in recent years it has been seen that they messed up through a colossal quantity of bogus content and misleading information. These all activities gave birth to fake news which has become the biggest challenge in front of us. The “news which has no or very less association with genuine content, and intentionally created and circulated just to spread wrong information among societies, which directly or indirectly fulfills the propaganda of someone or to benefits others is called fake news” [1].

Detection of fake news requires a lot of efforts because many times the manual fact-checking process is not a solution due to the fast internet it is very difficult for fact-checking and automated fake news agencies to detect false news by comparing it by searching the real event

corresponding to that fake news in less time. This paper tries to give a complete overview of the Natural language processing (a domain of computer science that is related to text analysis) and machine learning approach (an advanced field of computer science related to the neural networks and artificial intelligence-based phenomenon) to detect fake news which is accumulated and propagated over the world through online social media sites.

From the above section, we have got a crisp overview that ‘detection of fake news is very complex and filled with hectic tasks that requires a lot of computational work not only related to syntax level of processing of text but also related with right and suitable prediction of the semantic meaning of each post that goes viral on social media and found to be fake in general case [2].

This is why the basic methodologies of fake news detection have now shifted towards natural language processing and its associated areas. By exploring those features, many scientists and researchers proposed many solutions to this fake news problem after applying one or more models based on machine learning that improves the result of previous studies.

Moreover, the rapid growth in many areas of machine learning algorithms motivated many scholars to think in this direction and applying a machine learning model along with NLP features so that the accuracy of the fake news detection methodologies can be enhanced.

The organization of this paper is, in section II literature reviews of various works done in field of fake news detection using ML and NLP is discussed, all the famous and remarkable works carried out by researchers have been briefly discussed that will act as the backbone for methods of fake news detection., Role of NLP and its basics for fake news detection is explained in section III, in section IV fake news detection and its various steps like pre processing and features extraction are described, Section V refers to the detailed and comprehensive overview of advanced ML algorithms/ models which be applied to the various dataset. Section VI proposed a framework for detection of Fake News with flow graph representation, results of various ML based approaches are discussed in section VII, future research direction and challenges during associated with fake news detection are discussed in section VIII, finally the paper is conclude in section IX followed by references used.

II. LITERATURE REVIEW

Many researchers, authors, data scientists, scholars have published many articles and papers of their results in the fake

news detection field. Here we are explaining some of the famous work that has been carried out in the recent past.

Chen et al. [7] showed the wasteful conduct for classification saw by N-Grams and Part-Of-Speech (POS) label strategy. In this way, they prompted deep Syntax examination alongside Probabilistic Context-Free Grammars (PCFG) procedure.

Shlok Gilda [8] performed well in examining some key element extraction of text and showed that when text corpus is gathered as 'bi-grams' then TF-IDF execution upon it creates a much powerful model for fake news recognition.

Aldwairi et al. [9] showed a comparative analysis of some machine learning models based on Precision, Recall, F-Measure, and ROC values of given social media datasets.

LIAR [10] is the dataset that fascinated ample interest of various investigators for the research of detection of bogus news content. Authors [9] initially utilized the SVM model and afterward CNN (Convolution Neural Networks) Model, along with some other machine learning tools and algorithm side by side and afterward the performance of SVM, LR, Bi-LSTM is evaluated and it was demonstrated that inclusion of meta-data with text, significantly enhanced the accuracy of detection of false content.

Karimi et al. [11] Did research on LIAR dataset with a different technique by utilizing a superior method called Multi-source Multi-class Fake news Detection scheme (MMFD), which includes three folded mechanisms. MMFD is also later combined with the CNN model and LSTM.

Long et al. [12] worked with the "attention model" and in this procedure initially features were calculated from the text then converted into vector form later it fed into an LSTM feature that will enhance the accuracy of text deception detection.

Kai Shu et al. [2] had shown the data mining viewpoint and its nearby effect on identification of fake news by giving deep perception on content-based and social-context based examination for fake news.

Shu et al. 2018 [13] introduced the new dataset which was proposed to remove all limitations of a previously proposed dataset and named it as FakeNewsNet. This dataset has information related to three different aspects i.e. news context, social context, and dynamic context of information.

One more paper by Shu et al. [14] demonstrated the use of linguistic Inquiry and word count feature (LIWC) mechanism with FakeNewsNet dataset and analyzed that "usage of the words extracted from the FakeNewsNet dataset along with LIWC feature produces 64% accuracy.

Hannah Rashkin et al. [15] did an wide-ranging examination of a variety of linguistic features on documents extorted from various datasets and affirmed that LSTM generated improved results contrasted to other methods.

The another model of machine learning given by Bajaj S. [16] studied many machines and deep learning algorithms, their application with various publicly available datasets with NLP usage and initiate that RNN structure supported with GRUs did executes splendidly and it dawdled the other methods similar to LSTM.

III. NLP IN FAKE NEWS DETECTION

This is an obvious fact that the discovery of the fake news process gives prime importance to the content of news. Under this scheme news content, its structure, language, selection of words, tone of writing all such things play an essential role. And NLP is a very major tool for such news analysis which includes distinguished application tools to extract those features. The scope of NLP is very wide and it is mixed with other domains of computer science (Fig. 1) as well in the current age that can be illustrated below.

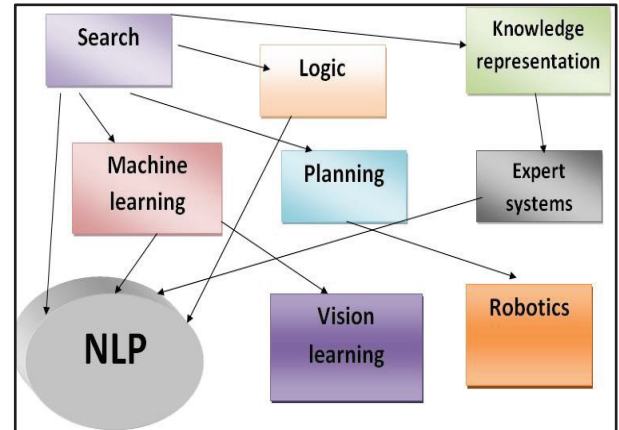


Fig. 1. NLP with different branches of computer science

The NLP is nothing but, a wide research wing of computer science that deals with the manipulation and detection of natural language that we human speaks and this field also tries to build computer model which can understand human recognizable natural languages with the help of computing software

Natural language processing has mainly two goals one is the science goal and another one is of engineering goal. In science goal, the main aim is to understand how language is produced and understood by an intelligent entity and engineering goals motivate to build such models that analyze and generate language that can reduce the machining gap. Initially, there are two prime motivation factors to initiate the research in this area.

TABLE I. THE VARIOUS TASK OF NLP

Task	Execution of Task
Information retrieval	It Finds out the required documents based on various keywords
Information extraction	It Identifies and extracts the statistics from a certain basis
Language generation	Based on the certain explanation about text generating characteristic it gives speech to discover a prototype
Text clustering	Here consortium of related types of speech, content are put in one body
Text classification	Refers to the assigning of different feature based categorizations to content.
Machine translation	This area of NLP used to translate whichever speech content into another speech
Grammar checkers	Refers to the checking of the grammar of any given language

- First, the natural language-based computer can be used in the information retrieval process that can detect, read the normal text and speeches written in the paper.
- Second, natural language processing will increase the understanding of the working process of various human languages and provide the inner implementation details of how any system will be able to learn and recognize the human language.

In today's time, NLP evolved as the main pillar for understanding and implementing the concept of text summarization, text classification sentiment analysis, and opinion mining for numerous advanced projects in computer science. The various task performed by NLP is shown in Table I.

IV. FAKE NEWS DETECTION

The discovery of Fake news is such a vast task that it cannot be easily done without preprocessed and well-synthesized input data. Furthermore, irrelevant and redundant features might and generally degrade the overall accuracy and performance of the classifiers. So, initially preprocessing of text is performed and thereafter by applying many feature extraction algorithms we can reduce the text feature size. For this many times preprocessing of textual data is done to refine the news data content and make it suitable for the further process. Some of the famous text preprocessing features are listed below.

A. Preprocessing

- **Stop Word Removal:** Although the stop words are used to fulfill the structure of sentences but they may lead to a big source of noise in a corpus of text. Articles, prepositions, conjunctions, and pronouns are called stop words so, in preprocessing work such words are removed, and rest words are passed on to the next step [3].
- **Lower or Upper case conversion:** It is also an essential preprocessing task as per user need that transforms all characters to lowercase or in uppercase as the need of situation.
- **Tokenization:** This is the mechanism of transforming the given sentences into smaller pieces called 'chunks of the word' or 'tokens'. Words, numbers, punctuation marks all such lexemes can be treated as tokens.
- **Stemming:** Stemming is performed when the tokenization is done and here tokens are converted into predefined standard form. It is related to transforming the tokenized terms within their initial form which lessens the terms by not counting similar terms, those have an identical stemmed result. For example, the terms "making", "made" and "maker" will be turned into "make" [4].
- **Lemmatization:** Just like the stemming process, this preprocessing step removes some of the inflectional forms of sentences or words into a common character or base form. Somewhat it is more sensible as compared to stemming lemmatization does not simply remove all the inflection property of token but it put all such

token in one correct base form on the lexical knowledge of words.

- **Part of speech tagging (POS):** This preprocessing step is utilized to situate the label of different parts of speech tags to each word of sentences.
- **Collocation extraction:** Collocations are the frequent types of words pattern that occurs together in any sentences often. In NLP such word combinations are very important to predict the nature of the text and also predicting the next word in a sentence.

B. Feature extraction

After the text pre-processing procedure to obtain the right and suitable features from the chunks of text data it cumbers some tasks. There is a much probability that inappropriate and unnecessary features could lessen the accuracy and degrade the classifier's overall performance. In the fake news detection process although many types of features are extracted that include News Context, Social Context, Source context, and Similarity-based feature since here we dealing with fake news so linguistic-based and visual-based features details are provided [2].

Linguistic Features: These are the most essential feature of any news content that is extracted and widely used in various classifiers for fake news detection and we are discussing more the linguistics feature of fake news.

- **Term Frequency-Inverted Document Frequency (TF-IDF) [25]:** it is one of the most essential and widely used feature extractions in the text analysis and preprocessing phase which is further used in establishing many results it consists of two main features and their basic discussion are following.

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t)$$

- **Term Frequency (TF):** It is, widely used in calculating and predicting the existed relationship among the word in a sentence based on calculating the probabilities of each word count in the sentence. Moreover, it is also true that all text dimensions differs, a term might emerge further in a long-sized text than a tiny one.

$$TF(t, d) = \frac{\text{no. of times 't' occurred in 'd' document}}{\text{total word count of document 'd'}}$$

- **Inverse Document Frequency (IDF):** it is based on the fact that sometimes a speech isn't of a lot exploit if it's subsist in all the texts. Like "a", "an", "the", "on", "of" etc. Materialize several times in a text but are of modest significance. IDF feature chooses such text or words and scales them down and side by side it also increases the importance of a rarely used word in the document ones. Here, more values of IDF signify the more uniqueness of the word. It is that technique which calculates the similarity index among the all the word occurred in given corpus of sentences.

$$IDF(t) = \log_e \left(\frac{\text{total no of documents}}{\text{no. of documents with term 't' in document}} \right)$$

- **Word level TF-IDF:** this is responsible for computing the occurrence of any particular word or term in different documents.

- ***N-gram level TF-IDF***: it is based on a calculation of terms associated in a document as per N- terms together in different documents.

We can say that TF-IDF strived by imprisoning the mainly frequently stirring speeches by conveying them fewer weights whereas giving significant weight to terms, which are exist in the appropriate subset of the corpus, and has in some text. Term Frequency & Inverse Document Frequency is the amalgamation of TF-IDF. One more thing is that it can be calculated as following two main levels.

- ***N-grams***: It is closely related to the clustering phenomenon in which a chunk of a token is arranged or extracted in one group based on searching the next word sequence pattern in given text or speech choosing n such grams (token) together. It is generally counted as unigram when $n=1$; bigram when $n=2$ and so on. N here generally represents unigram or bigram. After extraction of n-grams, these are used with the TF-IDF feature.
- ***Punctuation***: Proper usage of punctuations many times resulted in a lot of great extraction tools for detecting fake news. Punctuation feature compiles normally in 11 categories of punctuation that includes for example commas, dashes, exclamation marks periods and question marks, etc.
- ***Word Embedding***: It attempts to plot a statement to a vector utilizing a thesaurus. The frequency-based word embedding vectors was utilized for learning the statistics. They have too been classified into Linguistic based features.
- ***Count-Vectorizer***: It is another important feature extraction in a text in which a matrix is used to represent documents of the corpus, and a term from the corpus and in which row and column intersection known as cells signify the occurrence of the specific term in a specific text. By this, a special type of dictionary is made by collecting such unique tokens or words in the corpus [26].
- ***Psycho-linguistic features***: Apart from the basic linguistic feature the psycholinguistic features have also a big role in the context of the news. Many times words filled with positive, negative features can easily determine the tone of the language. And LIWC (Linguistic Inquiry and Word Counts) is one such valuable tool to detect psycholinguistic features from the corpus of text [5].
- ***RST and Vector Space Model (VSM)***: It has also got a very special place in natural language processing and is often used for deceptive content detection. Rubin et al. [8] showed in their paper that their approach could be an impressive technique of counterpart the actual lexical and semantic analysis tool in the process of fake content detection. This feature-based analysis related to human sentiment and thus, it finds out truthful and fake news in stories. RST framework provides the coherency in news detection and applies some calculative measures on the news along with the VSM technique.

V. MACHINE LEARNING MODEL

Machine Learning has although a major responsibility in preprocessing and feature extraction of assorted texts for fake news detection. Although NLP has prime importance in preprocessing, feature extraction of text and another vital role in linguistic text analysis it has been seen that many times this alone is not enough to detect a required pattern of fact-checking in many cases. So, that is why there is a need for some mechanism that can explain the predicted labels and extracted patterns and features in a more rigorous way which can be manipulated in a multi-dimensional way to discover and predict the unknown result in a more comprehensive manner. To accomplish such a task easily and lucidly we need machine learning models. It has the requisite and apt ability to train the extracted feature more deeply and predict the result in a much understandable manner. Machine learning with distinguished models and classifiers [6] can detect fake news and can produce optimum results. That is why it is widely used in many research areas including fake news detection area shown in fig. 2 :

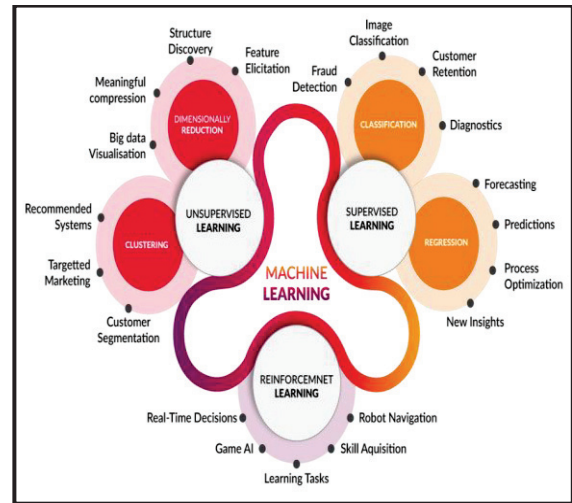


Fig. 2. Machine learning & its associated areas

- ***Linear Regression [17]***: This machine learning model is used to establish and predict that relationship that is linear among the target value and other predictors. There are two kinds: Simple and Multiple regression models.
- ***Logistic Regression [18]***: Basically, it is not a regression but it is a type of classification machine learning algorithm that is beneficial to produce Boolean or true-false value when input is given to it and thus it then classifies them into two domain of true and false algorithm. This input data in the logit function and then predicts the probability of any event that is why it is called logistic regression.
- ***Decision Tree [19]***: It is another famous algorithm of supervised learning that is utilized for clustering and as prediction models of ML. The decision tree algorithm deals with two standards in particular decision examination and from that point the decision-making stage.
- ***Random Forest [20]***: It consists of a collection of many decision trees of a machine learning model with an advanced approach to designing a random forest machine learning tool. This model is utilized

to calculate the group prediction. Here the groups having the most numbers of counts turn into the model's calculation. The wisdom of the crowd model and its co-relation with a lot of decision trees is the main concept behind this random forest model

- **Support Vectors machine [21]:** SVM is a sort of classifier model of ML model. We by and large gap the different class marks in a discriminative classifier by isolating hyper-plane in a manner all class names are consistently positioned on each side of the hyper-plane any uncertainty and having two planes basic to one name.
- **Naive Bayesian model [22]:** The Naïve-Bayes another famous algorithm based on a probabilistic approach to classifying the class label by calculating probability based on Baye's theorem. After that, it predicts the nature of the related domain. Since training inputs are given in form of predicted probability so, it is a kind of supervised learning approach. Along with simplicity, Naïve-Bayes is known for being the best among many algorithms, as sometimes it outperforms many other sophisticated classification methods machine learning model.
- **k- NN model [23]:** it is another type of machine learning model that might be utilized as together classification and regression difficulties, but additionally utilized in a classified-based problem. As per its mechanism at first, this algorithm tries to store available cases and after deployment of the nearest neighbor algorithm it model classifies them in new cases to further analysis based on majorities of the vote count of its k-neighbors. Here, as per the distance parameter Euclidean, Manhattan, Minkowski, and hamming distance functions are used.
- **K-means model [24]:** It is a widely used unsupervised machine learning algorithm that solves mainly clustering problems by following the approach of classifying a given number of a dataset into a certain number of clusters. Here, one notable thing is data sets or values residing in one group are somehow related to each other and on the other hand they are very different from all such data points which are situated in a different cluster. It iteratively by assigning each data point to one of K groups based on the given features, finds a new cluster.

VI. FRAMEWORK FOR DETECTION OF FAKE NEWS

It is clear that the diverse and complex nature of news content, the linguistic feature alone is not enough for fake news deception. There is a need for some mechanism that can explain the pattern and extract the features more deeply. Machine learning along with NLP, Artificial intelligence is a prerequisite model and classifiers, based on that fake news detection technique can produce the optimum result. The framework for Fake News Detection has the following steps:

- Dataset selection for the collection of news corpus
- Dataset mining work: Data preprocessing
- Deployment of feature extraction and Count vector on chunked data
- Selection of training and testing data from the featured matrix
- Applying machine learning classifiers model on the training dataset
- Build a model for proposed method
- Predicting result analysis
- Performance comparison

The overall scheme can be depicted with the help of the flowchart below in fig 3:

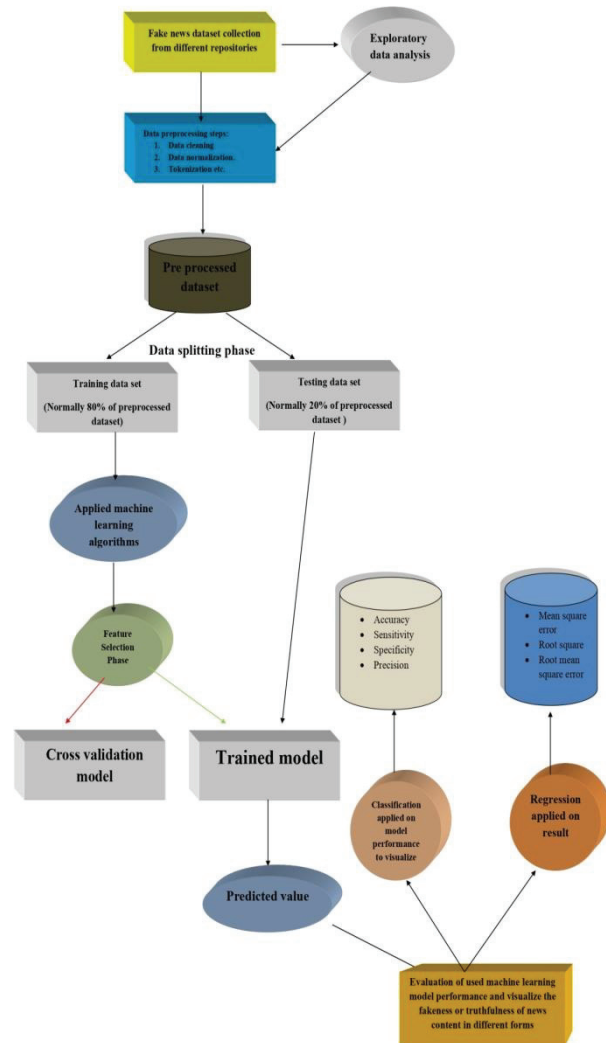


Fig. 3. Flow graph of farmework of ML Method

VII. RESULTS

In this sub-section, the performance evaluations of PCFG and TF-IDF for BI-Gram are discussed in Table II. The results obtained by implementing GB, SGD, RF, DT, SVM and baseline classifiers on the corpus dataset are shown in Table II.

TABLE II. RESULTS USING PCFG AND TF-IDF BI-GRAM

Model	Precision	Recall	ROC	Accuracy
Gradient Boosting	40.2%	16.1%	75.6%	65.7%
Stochastic Gradient Descent	74.1%	71.7%	87.5%	65.7%
Random Forests	84.2%	18.4%	80.0%	64.8%
Bounded Decision Trees	66.9%	37.9%	65.9%	67.6%
Support Vector Machine	80.9%	44.5%	84.3%	73.6%
Baseline	32.18%	32.18%	-	67.89%

The performance evaluations of TF-IDF for BI-Gram are discussed in Table III. The results obtained by implementing GB, SGD, RF, DT, SVM and baseline classifiers on the corpus dataset are shown in Table III.

TABLE III. RESULTS USING ONLY TF-IDF BI-GRAM

Model	Precision	Recall	ROC	Accuracy
Gradient Boosting	41.0%	22.3%	79.4%	68.7%
Stochastic Gradient Descent	88.8%	45.3%	88.3%	77.2%
Random Forests	82.9%	25.3%	78.8%	67.6%
Bounded Decision Trees	58.5%	23.3%	60.7%	66.1%
Support Vector Machine	81.3%	48.1%	85.6%	76.2%
Baseline	32.18%	32.18%	-	67.89%

The performance evaluations of various Machine Learning algorithms are discussed in Table IV. The results obtained by implementing Byes Net, naïve Bayes, Logistic regression and Random Trees on the corpus dataset are shown.

TABLE IV. RESULTS USING VARIOUS ML ALGORITHMS

Classifier	Precision	Recall	ROC	F-Measure
Bayes Net	94.4%	97.3%	100%	97.2%
Naive Bayes	98.7%	98.7%	100%	98.6%
Logistic	99.4%	99.3%	99.5%	99.3%
Random Trees	99.3%	99.3%	97.3%	99.3%

VIII. FUTURE RESEARCH SCOPE AND CHALLENGES

We have seen the remarkable research of detection of fake news utilizing NLP aided with distinguished ML tools. One more thing is that result of different detection of fake news models rely upon the quality and property of various repository of composed fake news by some sources. Coming to the future goal of fake news detection with ML and NLP based approach is:

- Here, from this above figure mentioned in the paper [2], we observed that generally previous trends of research, associated with the detection of fake news associated to the feature-oriented attributes of fake news that deals with providing deep observation to the news contents and social features of the texts for finding deception

pattern

- But in the current years with the headway in ML models the discovery philosophy, this culture is moved towards the model-arranged strategies and information oriented procedures where ML algorithm is applied on the different dataset of information as information sources and there are numerous noteworthy works done in this space that we have found in the past portions
- Coming to the application-oriented exploration of fake news, right now, it is the first and foremost stage and it continues to advance step by step. This region needs more examination and investigation. Since online media is brimming with dynamic exercises along these lines, toward this path, to alleviate the bits of rumors related exercises found on person to person communication destinations some interdisciplinary-based exploration approaches are being utilized by experts.
- Sentiment analysis is another unmistakable space of things to come of fake news identification since everything relies upon one's plan, one's outlook. How somebody needs to spread phony substance in which setting, so investigating such occasion in a superior way and foresee nature of information that could be viral in not so distant future dependent on some significant occasion occurring in ebb and flow time is enormous spaces of examination.
- A fact-checking mechanism in much more depth work analysis than that of fake news detection is nowadays required as per one of its future research.

Challenges associated with fake news detection are:

- Need for a more advanced multi-dimensional dataset:** We have seen that BuzzFeedNews, LIAR kinds of datasets have just a couple of attributes. FakeNewsNet despite the fact that is an all around organized dataset that has a much huge element of information types. Yet, we need a greater amount of this sort of dataset.
- Increased traffic of rumor, social bots, and Sybil account:** gossips, Social bots, and Sybil profiles on online media [6] in present era has constructed numerous awareness between many online media platforms in the last three years.
- Advancement in research methodology:** Most assuredly, the idea of online media is exceptionally powerful and in the previous five years the example and kinds of fake news additionally change a ton in this span thus, how it will in the following five years and how to distinguish and relieve them at a beginning phase is a major challenge before researchers

IX. CONCLUSION

In this paper, we have explored distinguished mechanisms of NLP and designs of detection of fake news. We explicitly investigated some of the previous remarkable results obtained in fake news detection along with some brainstorming analysis of the mitigating effect of dynamic fake news proliferation. We have elaborated all the prerequisite terminology associated with various models of machine learning found its great relevance in the discovery

of fake news. Then we looked upon the result and various outcomes obtained by many scholars, researchers, and fake news activists in due course of time. At last, we also touched on the futuristic goals of fake news detection by pointing out some of the famous challenges in front of us in this field.

REFERENCES

- [1] Burkhardt, Joanna M. "Combating fake news in the digital age". Vol. 53, no. 8. American Library Association, 2017
- [2] Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD explorations newsletter* 19, no. 1 (2017): 22-36.
- [3] James, H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition." *Person Education, Inc.* (2000).
- [4] Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT Press, 1999.
- [5] Mihalcea, Rada, and Carlo Strapparava. "The lie detector: Explorations in the automatic recognition of deceptive language." In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 309-312. 2009.
- [6] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumors in social media: A survey. *arXiv preprint arXiv:1704.00656*, 2017
- [7] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news, Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1-4, 2015.
- [8] Shlok Gilda. Evaluating machine learning algorithms for fake news detection. In *Research and Development (SCORED)*, 2017 IEEE 15th Student Conference on, pages 110-115. IEEE, 2017
- [9] Aldwairi, Monther, and Ali Alwahedi. "Detecting fake news in social media networks." *Procedia Computer Science* 141 (2018): 215-222.
- [10] Wang, William Yang. "liar, liar pants on fire": A new benchmark dataset for fake news detection." *arXiv preprint arXiv:1705.00648* (2017).
- [11] Gautam J., Atrey M., Malsa N., Balyan A., Shaw R.N., Ghosh A. (2021) Twitter Data Sentiment Analysis Using Naive Bayes Classifier and Generation of Heat Map for Analyzing Intensity Geographically. In: Bansal J.C., Fung L.C.C., Simic M., Ghosh A. (eds) *Advances in Applications of Data-Driven Computing. Advances in Intelligent Systems and Computing*, vol 1319. Springer, Singapore. https://doi.org/10.1007/978-981-33-6919-1_10
- [12] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp 252-256.
- [13] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. 2018 preprint *arXiv:1809.0128*
- [14] Kai Shu, Suhang Wang, and Huan Liu.. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*. 2017
- [15] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017 pp 2931-2937.
- [16] Bajaj, Samir. "The pope has a new baby! fake news detection using deep learning." *CS 224N* (2017): 1-8.
- [17] Maulud, Dastan, and Adnan M. Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning." *Journal of Applied Science and Technology Trends* 1, no. 4 (2020): 140-147.
- [18] Rymarczyk, Tomasz, Edward Kozłowski, Grzegorz Kłosowski, and Konrad Niderla. "Logistic regression for machine learning in process tomography." *Sensors* 19, no. 15 (2019): 3400.
- [19] Somvanshi, Madan, Pranjali Chavan, Shital Tambade, and S. V. Shinde. "A review of machine learning techniques using decision tree and support vector machine." In *2016 international conference on computing communication control and automation (ICCUBE)*, pp. 1-7. IEEE, 2016.
- [20] More, A. S., and Dipti P. Rana. "Review of random forest classification techniques to resolve data imbalance." In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, pp. 72-78. IEEE, 2017.
- [21] Shawe-Taylor, John, and Shiliang Sun. "A review of optimization methodologies in support vector machines." *Neurocomputing* 74, no. 17 (2011): 3609-3618.
- [22] Al-Aidaroos, Khadija Mohammad, Azuraliza Abu Bakar, and Zalinda Othman. "Naive Bayes variants in classification learning." In *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pp. 276-281. IEEE, 2010.
- [23] Mezquita, Yeray, Ricardo S. Alonso, Roberto Casado-Vara, Javier Prieto, and Juan Manuel Corchado. "A Review of k-NN Algorithm Based on Classical and Quantum Machine Learning." In *International Symposium on Distributed Computing and Artificial Intelligence*, pp. 189-198. Springer, Cham, 2020.
- [24] Bodapati S., Bandarupally H., Shaw R.N., Ghosh A. (2021) Comparison and Analysis of RNN-LSTMs and CNNs for Social Reviews Classification. In: Bansal J.C., Fung L.C.C., Simic M., Ghosh A. (eds) *Advances in Applications of Data-Driven Computing. Advances in Intelligent Systems and Computing*, vol 1319. Springer, Singapore. https://doi.org/10.1007/978-981-33-6919-1_4
- [25] Jalilifard, Amir, Vinicius Fernandes Caridá, Alex Fernandes Mansano, Rogers S. Cristo, and Felipe Penhorate Carvalho da Fonseca. "Semantic sensitive TF-IDF to determine word relevance in documents." In *Advances in Computing and Network Communications*, pp. 327-337. Springer, Singapore, 2021.
- [26] VEL, S. SAKTHI. "Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries." In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 879-884. IEEE, 2021.