Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200146

Analyzing Fake News Based on Machine Learning Algorithms

Pawar A B[a], Jawale M A[a], Kyatanavar D N[a]
[a] *Sanjivani College of Engineering, Kopargaon, India*

Abstract. Usages of Natural Language Processing techniques in the field of detection of fake news is analyzed in this research paper. Fake news are misleading concepts spread by invalid resources can provide damages to human-life, society. To carry out this analysis work, dataset obtained from web resource OpenSources.co is used which is mainly part of Signal Media. The document frequency terms as TF-IDF of bi-grams used in correlation with PCFG (Probabilistic Context Free Grammar) on a set of 11,000 documents extracted as news articles. This set tested on classification algorithms namely SVM (Support Vector Machines), Stochastic Gradient Descent, Bounded Decision Trees, Gradient Boosting algorithm with Random Forests. In experimental analysis, found that combination of Stochastic Gradient Descent with TF-IDF of bi-grams gives an accuracy of 77.2% in detecting fake contents, which observes with PCFGs having slight recalling defects

Keywords. Classification Algorithms, Context Free Grammar, Fake News detection, Natural Language Processing, Machine Learning

1. Introduction

In this section, we discuss about the relevant theory, scope and objectives. Relevant theory briefs about basic concepts like what is fake news, introduction about NLP, use of machine learning in proposed research work. Fake news are a kind of such journalism where false information spread over traditional news media's platforms too. Generally, this kind of deliberate disinformation prepared by reporters with unethical ways like paying for news, highlighting or breaking stories. It known as checkbook journalism. These kinds of news penetrated in main media as well as social media too. The intent of creating such junk news is always to mislead and to damage reputation of any entity, organization, or individual, and or gain financial profit or political benefit. In addition, spreading such unethical, dishonest or sensitive pseudo news increases readership of media. On social media platforms, it helps to earn revenue by such click bait news. Today, significance of such news has increased in politics, advertising agencies, viewers to websites and it competes with legal news stories with great impact. Even during election times such pseudo news implicated greatly. Therefore, studying on such fake news content is necessary and concern for legitimate user for correct information.In today's informative world, natural language processing (NLP) plays its vital role during informative collection, processing and retrieval. NLP is sub variant of computer science. Artificial Intelligence deals with machine and human interaction in order to program machine in such way that it analyses large amount of data and takes decision. Major challenges in NLP frequently

[1] Pawar A B, Sanjivani College of Engineering, Kopargaon, SPPU, India;
Email: anil.pawar1983@gmail.com

consist of natural language understanding, natural language processing and natural language generation.

Machine Learning (ML) is study and research of algorithms and computational models that computer applications use to function specific tasks without external interventions. This execution runs mainly on patterns and inference knowledge. It is sub variant of Artificial Intelligence. It creates mathematical prototyping of sample input data, which is known as training data. This training data is used predict results or decisions without being externally programmed to do its operations.

Machine Learning algorithms are applicable for multiple applications such as e-mail filtering, computer vision, computational statistics, and data mining. To do analysis of fake news contents, collecting of right corpus is challenging task. So, during research work data collection following factors were considered

1. Can truthful and deceptive instances will be available for study?
2. Can it verifiable with ground truth?
3. Can we find homogeneity in collected contents in terms of length or in writing contents?
4. What will be manner of posting i.e. is it based on sensitivity, political etc.

In this regard, to resolve these kind of challenges the source like OpenSources.co website is used. This website mainly compiles list of ongoing fake and trusted news list. Main challenge observed in collecting fake news corpus is copyright issue. From [6] a dataset containing 1 million articles is used during study. It contains sources like Reuters, local news sources, blogs, etc.

To make data ready for further analysis, it is required to clean data. To do same task, data pre-processing is carried out which is one of the technique available in Data mining. This is well known and testified technique used to make data understandable and ready to use for further analysis into better decision making.

In the next stage i.e. Feature generation, after getting collection of pre-processed data, there is always need to identify potential data elements for statistical analysis. Therefore, feature generation helps in decreasing unwanted data processing and increases chances of getting potential data for proper prediction. In context of fake news also, it proves its importance by removing common words by selecting certain size of block of data, by applying set of patterns etc. Generated features then can play vital role in research analysis.

Feature extraction and selection of specific feature set is critical since it directly effects on results of research work. Therefore, testing such derived values of generated features is required as they may contain vital information or relationship patterns useful for analysis.

2. Literature Review and Proposed System

In this section, different Literature Reviews, motivation Outcomes from it and the proposed system architecture discussed,

S. Feng, R. Banerjee and Y. Choi [1], “Syntactic stylometry for deception detection”, have proposed novel approach than existing methods of deception detection defined as ‘syntactic stylometry’. For testing it, four types of datasets considered from product review to essay collection. In this investigation, features obtained from Context Free Grammar (CFG), which in turn fed into parse trees for each iteration found improved detection over predefined thresholds depend on shallow lexico-syntactic features. It reports 91.2 % accuracy along with 14% error reduction on hotel review dataset analysis.

N. J. Conroy, V. L. Rubin, and Y. Chen [2] in the paper, "Automatic deception detection: Methods for findings fake news", have given a description of two assessment methods employed in deception detection. These approaches are linguistic cue and network analysis. Performance of these assessment methods shown significant improvement than a random guess (more than16 percentage) in deception detection of business communication contents. In comparative study of human judgement and SVM Classifier, SVM Classifier shown 86% performance accuracy for detecting negative spam.V. L. Rubin and T. Lukoianova in [3], "Truth and deception at the rhetorical structure level", have described an analytical framework for detecting fake deceptive stories and truthful stories based on coherence and applied structure. This framework is termed as RST (Rhetorical Structure Theory).In [4], "Fake News Detection Using Naive Bayes Classifier", have shown a simple approach for fake news detection using naive Bayes classier. Using this approach, they built their own classifier and tested against a data set of Facebook news posts. They have achieved classification accuracy of 74%, which is very good considering the simplicity of their model. Even more details studied from rest of reference given in this research paper [5-21] and found need of analyzing the use of NLP in field of fake news detection. Based on this study, following system model proposed for analysis work as shown in Figure 1.

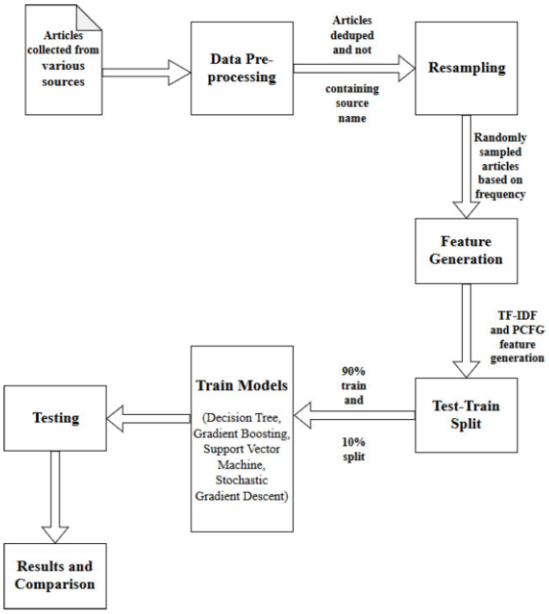

Figure 1. System Architecture

2.1 Data Preprocessing

In this step, any article mentioning the name of the source are removed. This is necessary. In order to see that selected prototype will not perform mapping from known resources while categorizing news articles as reliable and or unreliable from obtained resources. Twitter handless and e-mail addresses are removed for the same reason.

2.2 Resampling

For selected model, it is required to ensure that selected model will identify fake and original news, the system do resampling of large dataset to get expected results. In this system , sample contains 500 articles as maximum value of n which may be seem less but worth to get desired results.In addition, the low frequency sources were not dropped to maintain some heterogeneity of sources. The accurate value of maximum sample corpus is concern and it can be studied in research. If faced with similar corpus difficulties, then in order to achieve an optimal result, one can consider variable sizes of samples to experiment the research. A slight drop in precision was noticed before and after re-sampling the distributions, which indicated, that fitting happened.

2.3 Feature Generation

The feature set which is vectored bi-gram. Vectored bigram is the term based on TF-IDF, which is more relative to particular document corpus among document collection. Another technique used is Probabilistic Context Free Grammar for feature generation.Feature generation subtask such as generation of tokenized keywords, performing part of speech tagging of news article statements with syntactic parsing, identification of entities are performed with the help of python programming using Spacy package which is found better than NLTK NLP package. This selected python package is part of Cython [20].

2.4 Test-Train Split

In this step, the dataset is split into two parts 90% for training the model and 10% for testing the dataset after model is trained.

2.5 Train Model

Training model makes use of machine learning algorithms as selected Stochastic Gradient descent, Support vector machine, and bounded decision tree used in order to train the model.

2.6 Testing

Once the model is trained remaining 10%, dataset is tested on it and results are compared for various combination of algorithms.

3. Implementation Details

This section explore implementation details of research work and executed on computer machine with following dataset and tools,
1. News Articles (reliable and unreliable)
2. Spacy (python package)
3. NLTK
4. SciKit Learn
The major algorithmic steps followed for system implementation are

1. Collect news articles from various sources as described in earlier sections
2. Preprocess these articles as per prosed system model preprocessing strategy for
(i) Removal of duplicate copies of articles.
(ii) Clearing the source name from article.
3. Resample articles based on frequency of each source.
4. Generate PCFG and TF-IDF features for each article as per specified algorithm requirements.
5. Divide the given dataset of articles into two parts. One part consist 90% of total articles for training model and remaining 10% into other part for testing.
6. Train the models using various machine-learning algorithms and 90% of article dataset, for classifying them as authentic or fake. These algorithm details are described in subsequent section.
7. Test the model trained on remaining 10% dataset.
8. Collect the results from all models and compare them.
Following machine-learning algorithms were used to train model for proposed system.

3.1 Support Vector Machine

A Support Vector Machine (SVM) is a classification algorithm, which separates data element based on hyperplane. It is based on discrimination. It gives output of new entities classification based on supervised learning. In 2-D space, it generates two classes of entities, which are separated by hyperplane.

3.2 Random Forests

Random Forest is a learning algorithm, which makes use of multiple decision trees involved in decision making as per selected random forest. The decision trees are trained using bagging method, which based on the idea that by combining learning models so the overall results will increase.

3.3 Bounded Decision Tree

Under supervised learning algorithms, Decision tree is one of the classification algorithm, which can be applied for categorical, as well as for continuous variables in classification problems. In this algorithm, unbounded set of decision tree will end up giving 100% accuracy because in worst case there will be one leaf for each observation. However, bounded decision tree will impose constraints on this tress to improve its performance.

3.4 Gradient Boosting

Mainly, it is used in regression as well as in classification wherein it gives results based on prediction in stepwise approach, which later on generalized to produce optimum results by reducing an arbitrary differential loss function.

3.5 Stochastic Gradient Boosting

Being an improved version of gradient boosting, it uses bagged base learners instead of original learners and replaces ordinary residuals by out-of-bag one.

4. Result Analysis

The obtained results during analysis of ML algorithms discussed in following section.

4.1 Combining PCFG and bi-gram TF-IDF model

After combining the feature sets, it is seen that the performance of the algorithms is well above the baseline (Figure 2). Stochastic Gradient Descent not only retains a high recall but also gives high precision indicating that these models can be used for both analyzing important articles as well as "fake news" filtering.

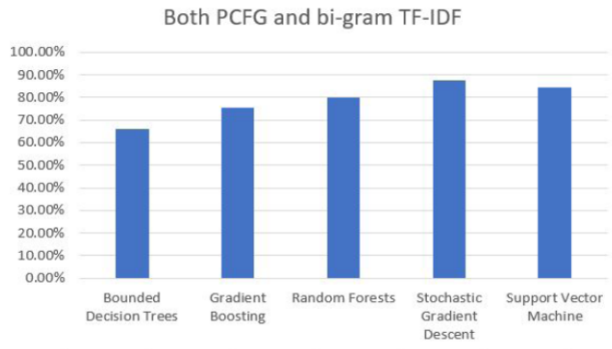

Figure 2. Results Obtained on Combination of PCFG and bi-gram TF-IDF

4.2 Only TF-IDF model

The value of the features responsible for achieving the combined results can be understood in more details after removing the PCFG features as shown in Figure 3.

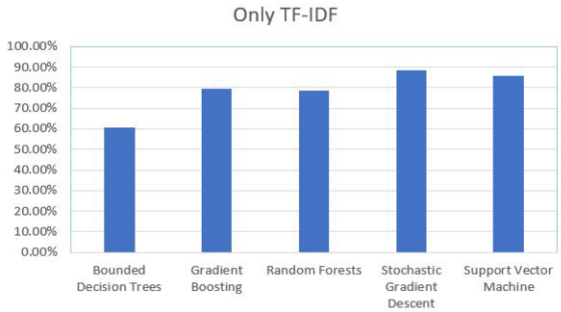

Figure 3. Results obtained on when only TF-IDF is used

4.3 Only PCFG model

The separation of PCFG's prediction results from TF-IDF bi-gram features is possible. In isolation, PCFG cannot be best one for classification. To understand why results from all models are common, prediction produced by each one were taken under consideration and it was found that the produces rank order of scores was same for all models. Authors used a top-k of 0.05 for classification rather than a 0.70 threshold as the distribution score. For these models, using low mean values with tightly coupled

ranges makes classification as tiresome task as the 0.70 threshold were not that illuminating. This indicates that PCFGs do not consider prominent resource information for classification. Although classifier performed really well, it is found improvement regions. Author evaluated models based on the absolute probability thresholds; however, for models where probability scoring is not well calibrated it may not be much reliable. Although TF-IDF performs better, it overfits to topics important in the ongoing news cycle. In addition, the analysis is hampered since a vectored approach makes it technically hard to predict importance of features. Due to such issues, analysis is limited and broader generalizability is prevented.

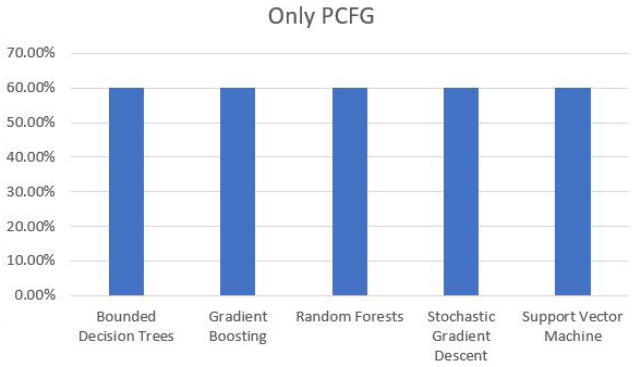

Figure 4. Results obtained on when only PCFG is used

5. Conclusion and Future Scope

Overall, the proposed system provided promising results. Considering the use of machine classification for identification, this method shows that the Term Frequency is potentially predictive of fake news. Model based on TF-IDF feature sets, the Stochastic Gradient Descent models are the best performing models considering the overall receiver operating characteristic (ROC). On the other hand, PCFG found equivalence for recall and precision in top model performance rather than additional decisive values. It indicates that use of PCFG is better for fake news article classification. Though TF-IDF is observed useful but can produce more promising results if more complete corpus will be available for experimentation and use of absolute probability score may not be more significant in case of improper calibration of detection models.

References

[1] S. Feng, R. Banerjee, and Y. Choi, Syntactic stylometry for deception detection, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171-175.

[2] 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, 2015, pp. 1-4.

[3] V. L. Rubin and T. Lukoianova, Truth and deception at the rhetorical structure level, Journal of the Association for Information Science and Technology, vol. 66, no. 5, 2015, pp. 905-917.

[4] Mykhailo Granik, Volodymyr Mesyura, Fake News Detection Using Naïve Bayes Classifier, PloS one, vol. 10, no. 6, e0128193, 2015.
[5] V. W. Feng and G. Hirst, Detecting deceptive opinions with profile compatibility, in IJCNLP, 2013, pp. 338-346.
[6] Open sources. [Online]. Available: http: / / www. opensources.co/
[7] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009, PP.1-7.
[8] K. Chellapilla and D. Chickering. Improving cloaking detection using search query popularity and monetizability. In AIRWeb, 2006.
[9] J. Chung, C¸ Gulc¸ehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.
[10] Z. Gyongyi and H. Garcia-Molina. Link spam alliances. In VLDB, 2005.
[11] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In VLDB, 2004.
[12] M. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. In SIGIR Forum. 2002.
[13] G. Hinton. A practical guide to training restricted Boltzmann machines. In Neural Networks: Tricks of the Trade (2nd ed.). 2012.
[14] Kingsbury. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 2012.
[15] G. Hinton, S. Osindero, and Y. The, A Fast Learning Algorithm for Deep Belief Nets. Neural Comput., 2006.
[16] N. Immorlica, K. Jain, M. Mahdian, and K. Talwar, Click fraud resistant methods for learning click-through rates. In WINE, 2005.
[17] H. Jaeger. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. Technical report, Fraunhofer Institute for Autonomous Intelligent Systems (AIS), 2002.
[18] Y. Kim. Convolutional neural networks for sentence classification. In EMNLP, 2014.
[19] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In AIR Web, 2006.
[20] Shlok Gilda. Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection .2017 IEEE 15th Student Conference on Research and Development (SCOReD), 2017.
[21] Jerome H. Friedman.Stochastic gradient boosting.Computational Statistics & Data Analysis, 2002.
[22] V.D. AmbethKumar, Vijaya Rajasekar, V.D.AshokKumar, "EFFICIENT DAILY NEWS PLATFORM GENERATION USING NATURAL LANGUAGE PROCESSING", International Journal of Information Technology, June 2019, Volume 11, Issue 2, pp 295–311.