Term Weight Measure based Approach for Fake News Detection

SaiPrasanth K
Asst Professor, Information technology
Bapatla engineering college, Bapatla, A.P , India

Kiran Kumar N
Asst Professor, Dept.of MCA
Bapatla engineering college, Bapatla, A.P , India

Dr SrinivasaRao K
Asst Professor, Information technology
Bapatla engineering college, Bapatla, A.P , India

Ravi Kumar K
Asst Professor, Information technology
Bapatla engineering college, Bapatla, A.P , India

Abstract:Digital platforms have dramatically changed the process of news production, consumption and dissemination in our society. People are using digital platforms for campaigning of misinformation or false information that change the trustworthiness of the entire news system.The main purpose of spreading such fake news is to misguide or misinform the common reader,damage fame of any person, firm or a nation, create confusion, gain financially or to gainpolitically from sensationalism. The fake news detection becomes hot research topic in recent times. Several approaches proposed based on machine learning and deep learning techniques for detecting fake news. In this work, a machine learning based approach is proposed based on term weight measures. The authors are used different type of terms in real news and fake news. This is the motive to develop an approach based on the terms used in the dataset. The most frequent terms in the dataset are used for representing the documents as vectors. The value of a term in vector is determined by using different term weight measures. Various machine learning algorithms are used in this experimentation to construct the learned model. This model is used to predict the accuracy of the proposed system as well as to predict the class label like real news or fake news of a test document. The Random Forest algorithm obtained good accuracy for fake news detection when compared with other algorithms.

Key Words: Fake News Detection, Fake News, Machine Learning Algorithms, Term Weight Measures,

1. Introduction

Digital platforms like twitter, Facebook, blogs and social media websites enhancethe communication and interactionsamong users [1]. These platforms are used by the people forpropagation of misinformation or fake news which are oftenintended to deceive people, especially in contexts such as health and politics. Fake news is a news article or message published and propagatedthrough media, carrying false information regardless the means and motives behindit [2]. In recent times, the detection of fake news is attracted by several researchers and proposed different types of solutions for fake news detection. This type of fake news spread among millions of people online, has attracted much attention from researchers, and many studies have been made in order to examine fake news and its impact on society and democracy.

Theautomatic solutions for fake news detection are used asan auxiliary tool for fact-checkers to identify whether the content is fake orreal.An effective way to detect fake news disseminated on digital platforms is the directfact-checking, typically performed by expert journalists. A fact-checking task verifies the correctness ofthe information by comparing them with one or more reliable sources. Examplesof such organizations include “Snopes.com”, “PolitiFact”, “FactCheck.org” etc.However, fact-checking is a time-consuming process since it commonly requiresa detailed analysis to support the verdict. Therefore, some studies are emerging toward computational fact-checking[3] including automatic detection of fake news.

Currently, there are several approaches to perform automatic fake newsdetection based on artificialintelligence techniques such as supervised, weakly supervised viareinforcement, active and deep learning and also basedon specific strategies such as block chain technology.In this work, the Most Frequent Terms (MFT) in the data set

are stored as bag-of-words. The bag-of-words model represents statements, sentences as multiset of words andall occurrences of an element is stored in a vector but order and grammar of words are ignored. In bag of words model, spatial information is not captured such as word co-occurrence. The document vectors are represented as vectors with these bag-of-words. The vector value of a term is computed by using different Term Weight Measures (TWMs). In this approach, different term weight measures are used in the experiment. The document vectors are passed to Machine Learning Algorithm (MLA) to determine the accuracy of proposed approach. The ISOT fake news dataset is used in the work for fake news detection.
The details of dataset are presented in section 2. The performance metrics are explained in section 3. Section 4 describes the need of TWMs and explained different TWMs used in this work. The proposed approach is explained in section 5. The experimental results are presented in section 6. The conclusions are specified in section 7 with future improvements.

1. **Characteristics of Dataset:** In this work, the dataset named as ISOT Fake News Dataset[7]is used for fake news detection. The dataset is divided into two types of articles such as real news and fake news. Real world resources are used to collect this dataset. The real news articles are collected from news website Reuters.com. The fake news articles collected from Wikipedia and Politifact tool which is a fact checking organization. The majority articles of dataset focused on World news and political topics. The information related to dataset is presented in table 1.

Table 1. The details of the dataset

	News type	Number of Articles	Total Number of Articles
Fake News	News	9050	23481
	Politics	6841	
	Left-news	4459	
	Government - News	1570	
	US news	783	
	Middle - East	778	
Real News	Politics – News	11272	21417
	World - News	10145	

2. Evaluation Measures

The researchers widely used F1-score,precision,recall and accuracy as a performance metrics in imbalanced domains. These metrics are derived by using the confusion matrix. Table 2 shows the confusion matrix.

Table 2. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Where, TP is count of postive class documents predicted as positive, TN is count of negative class documents predicted as negative, FN is count of positive class documents predicted as negative, FP is count of Negative class documents predicted as positive.
Accuracy is a performance measure and it is ratio of rightly predictedvalues to the total values. Higher the accuracy indicates better the model.Accuracy is represented in equation (1).

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

Precision is rightly predicted positive values in the total positive predictions. Higher the value ofprecision indicates false positive rate is lower. Precision is calculated by using equation (2).

$$Precision = TP/(TP+FP) \quad (2)$$

Recall is rightly predicted positive values in the all values predicted correctly. Equation (3) is used to compute the recall.

$$Recall = TP/(TP+FN) \quad (3)$$

F1 – score is the harmonic mean of recall and precision. F1-score is represented in equation (4).

F1 Score = 2 (Recall * Precision)/(Recall + Precision) (4)

3. Term Weight Measures

In Vector Space Model, the documentsare represented as vectors. Every document vector consists of numerical weights of terms or featuresthat are extracted from the dataset. Text categorization mainly depends on the term weights in the document but the procedure of determining the term weight affects the classification accuracy. The TF-IDF is the most popular term weight measure for assigning the weights to the words in information retrieval domain. Sebastiani (2003) et al., proposed [8] the thought of Supervised Term Weight (STW) measures, which are used to allocate weight to terms by considering the known class membership information of documents.

Majority of STW measures used the information about the distribution of terms in two classes such as negative and positive while assigning weight to the terms [9]. These measures performance were not good for multiclass text classification when the dataset contains more than two classes of textual documents. When the experiments were performed on multiclass datasets, the positive and negative classes were divided into “one-versus-all” or “one-against-rest” [10]. In one-versus-all, one class of documents was considered as positive class, remaining was considered as negative class.

Most of the STW measures used term’s intra class distribution information which is the distribution of a term within documents of a class. One researcher observed [11] that both inter-class distribution (distribution of term among the classes of documents) and intra-class distribution information of terms is equally important in STW measures. Generally in text classification, the importance of a weight of a term mainly depends on the discriminative power of a term. The term inter-class distribution information is more crucial than intra-class distribution to decide the term’s distinguishing power.

3.1 TF

The TF (Term Frequency) is number of times a term is occurred in a document. The TF allocates more weight to the terms which are frequently present in a document but some of the words that are frequently used by the authors are not having any distinguishing power.

3.2 TFIDF

To overcome the drawback of TF measure, researchers introduced [12] a new measure TF-IDF with the combination of TF and a global factor known as IDF (Inverse Document Frequency). The measure IDF is inversely proportional to the document frequency. In TF-IDF, the words which were appearing rarely in a corpus was assigned with more weights to the terms considering two factors such as the total number of documents contain this term and the frequency of term in a document. So, TF-IDF measure was widely used and it is more suitable than the TF measure. Equation (5) shows the TF-IDF measure.

$$TFIDF\left(T_i\right) = TF_i \times \log\left(\frac{N}{DF_i}\right) \quad (5)$$

Where, TF_i represents the frequency of a given term t_i in the document, N is total documents count in the dataset and DF_i is the number of documents contain the term t_i at least one time.

3.3 Normalized Document Length Term Weight (NDLTW) Measure

In this work a normalized measure is used to normalize the lengthy and small text base document which uses the information of the inner document distribution to get the weight of the terms. This technique is name as the NDLTWmeasure[13] and represented in equation (6).

$$W_{tij} = W(t_i, C_j) = \sum_{k=1}^{m} \frac{(LOG(tf_i)+1)/(LOG(avgtf_i)+1)}{SLOPE * ut_k + (1-SLOPE) * avgut_k} \quad (6)$$

Where, $\{D_1,\ldots,D_m\}$ represents the set of documents, t_{ij} represents the weight of a term t_i in a classC_j, tf_i is the number of times the term t_i appears with in a document k, $avgtf_i$ is average term frequency of termt_i to the total number of terms in a document k . The SLOPE value is 0.2.$ut_k$is unique terms count, $avgut_k$ is average of unique terms.

3.4 Relevance Frequency based Term Weight (RFTW) Measure

Researchers analysed broadly and proposed some new and modified STW measures. The good example of those measures is Term Frequency-Relevance Frequency (TF-RF) [9]. In this measure, the weight of a term is computed based on its Relevance Frequency (RF), which is the ratio among the positive class of documents contain the term (a)and the negative class of documents contain the term (c). The RFTW measure is explained in equation (7).

$$TF * RF = TF * LOG\left(2 + \frac{a}{MAX(1, c)}\right) \tag{7}$$

TF-RF measure assigned more weight to the terms which are specific to a positive class. The basic plan of TF-RF measure is the terms which are occurred in more positive class of documents when compared with negative class of documents were more useful to select positive text from a negative text [A3].

3.5 TF – Prob

A modified STW measure was proposed in [11] which is based on the probability information of terms known as TF-Prob. The TF-Prob measure is a combination of A/B and A/C. The weight of a term t_i with respect to C_j in TF-Prob measure is shown in equation (8).

$$TF - \Pr ob(t_i, C_j) = TFi \times \log\left(\frac{A}{C} * \frac{A}{B} + 1\right) \tag{8}$$

TF-RF measure includes only the term's inter-class distribution and is represented by A and B. But TF-Prob measure includes the intra-class distribution and inter-class distribution of a given term, represented by A and C. The reason for introducing the intra-class distribution in TF-Prob measure is that the terms which were appeared in most of documents in a positive class i.e., A and C obtained good weight to represent the positive class.

4. Proposed Model

In this work, five TWM's are used to determine the feature value in the vector representation of documents. The proposed approach is displayed in fig. 1. In the proposed approach, the first step is collection of dataset. After identification of suitable dataset for fake news detection, next step is applying different preprocessing techniques to remove the unwanted data. In this work, the preprocessing techniques applied on the dataset are tokenization, stopword removal and stemming. Tokenization is technique of splitting the content into terms. Stopwords are words which are frequently used by the author but they don't have any distinguishing power. Stemming converts the word into its root form. After cleaning the data, next step is extracting the most frequent 8000 terms. Then, the TWMs are used to compute the weights of terms. The documents are represented as vectors by using the weights of the terms. Finally, MLAs are used to generate the model by training with document vectors. The test document vectors are passed to the model to predict the fake news or real news. The learned model also gives the accuracy of the proposed system.

4.1 Machine Learning Algorithms (MLA)

The MLA creates a model by training on the dataset. The prediction is made when the test data is passed to the learned model. The algorithm examines the training data and reproduces the approximate function, whichcan be termed mapping function. This function can then be used for predicting newinput dataset. The Machine Learning algorithm faced the problem of overfitting when the function tends todescribe outliers and noise instead of capturing the underlying relationship in the trainingdata.A model that overfits on the training data shows poor predictive performance on unseenexamples, implying a generalization error. Overfitting occurs when a model is toocomplex.
Fourmachine learning algorithms such asNaive Bayes (NB),k-Nearest Neighbors (KNN),non-linear Support Vector Machinewith the Radial Basis Function (SVM) and Random Forests (RF) are used in this work.

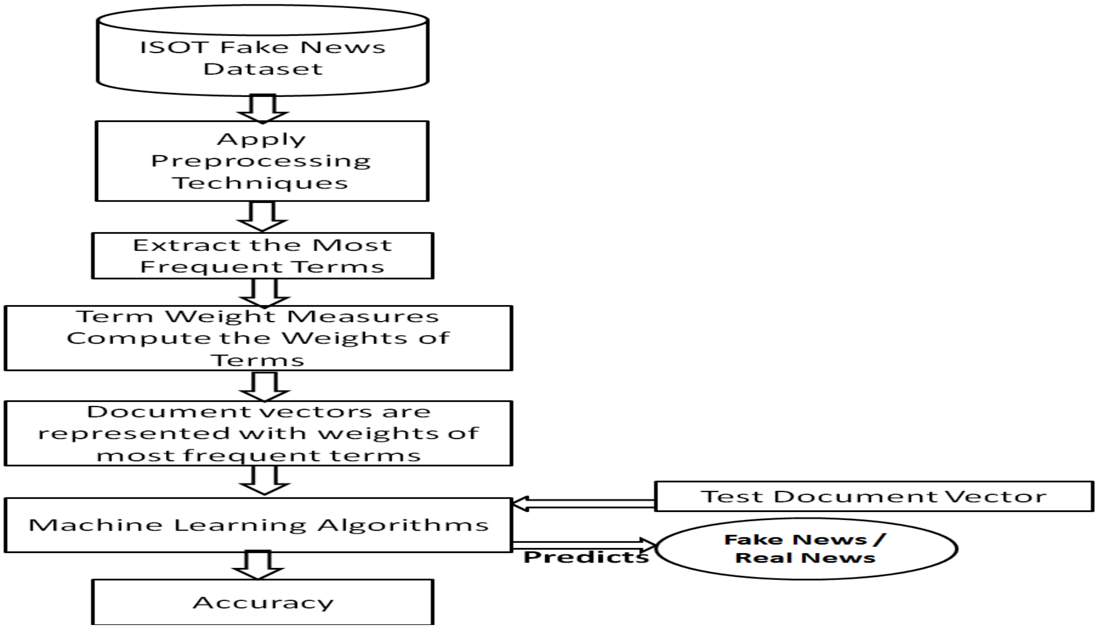

Fig. 1. The Proposed Approach for Fake News Detection

5. Experimental Results and Analysis

In this work, the experiment carried out with weights of MFTs and different MLAs for fake news detection. The Table 3 displays the accuracies of fake news detection when experimented with most frequent 8000 terms.

Table 3. The accuracies of Fake News Detection when experimented with most frequent 8000 terms

Term Weight Measures / machine Learning Algorithms	KNN	NB	SVM	RF
TF	0.5512	0.5885	0.6124	0.6688
TFIDF	0.5946	0.6253	0.6543	0.7073
NDLTW	0.6187	0.6561	0.6877	0.7339
TFRF	0.6630	0.6992	0.7205	0.7827
TF-Prob	0.6854	0.7136	0.7513	0.8132

The combination of RF classifier with TF-Prob term weight measure achieved highest accuracy of 0.8132 for fake news detection. The fig. 2 shows the diagrammatic representation of accuracies of fake news spreaders detection.

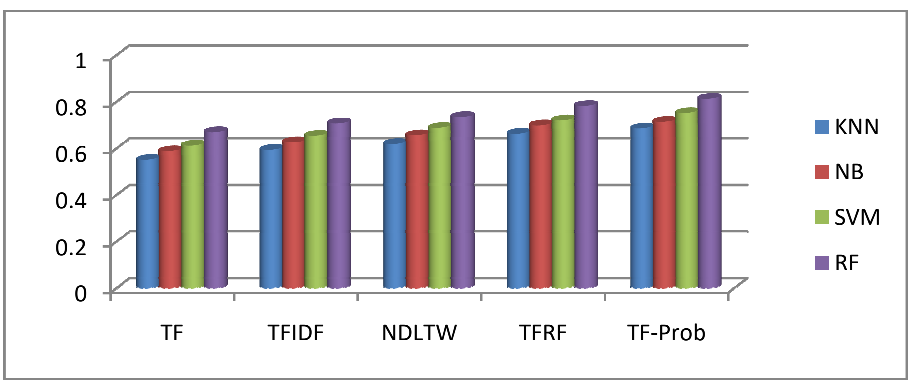

Fig. 2. The Accuracies of Fake News Detection

From fig. 2, it was identified that the RF classifier attained best accuracies for fake news detection when compared with KNN, NB and SVM. It was also identified that the TF-Prob performance is good in the prediction of fake news detection accuracy when compared with other TWMs accuracy.

6. Conclusion and Future Scope

The advent of social media and mass-information on the internet has led to fake news taking on a new form compared to its previous iterations. Fake News representstypically fabricated news or false hype comprising information broadcasted throughvarious forms of media.In this work, the experiment conducted with content based features of MFTs in the dataset for detecting fake news. The TWMs are used to allocate suitable weight to the terms. Five term weight measures are used in the experiment. The documents are represented as vectors by using most frequent 8000 terms and each vector value of a term is computed by using term weight measure. Various MLAs are used for performance evaluation of proposed approach for fake news detection. The RF classifier attainedbest accuracy of 0.8132 for fake news detection.
In future, it was planned to develop a new TWM to increase the accuracy of fake news detection. It was also planned to implement suitable deep learning techniques with appropriate hyperparameters.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017a). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1):22--36.
2. Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., and Glass, J. (2019). Automatic fact-checking using context and discourse information. Journal of Data and Information Quality (JDIQ), 11(3):1--27.
3. Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., and Glass, J. (2019). Automatic fact-checking using context and discourse information. Journal of Data and Information Quality (JDIQ), 11(3):1--27.
4. Alam, S., &Ravshanbekov, A. (2019). Sieving Fake News From Genuine: A Synopsis. arXiv preprint arXiv:1911.08516.
5. Rohit Kumar Kaliyar, AnuragGoswami, Pratik Narang, SoumenduSinha, “FNDNet – A deep convolutional neural network for fake news detection”, Cognitive Systems Research 61 (2020) 32–44
6. Huang, Y-F., Chen, P-H., Fake News Detection Using an Ensemble Learning Model Based on Self-adaptive Harmony Search Algorithms, Expert Systems with Applications (2020), doi: https://doi.org/10.1016/j.eswa.2020.113584
7. https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php
8. Debole, F. ,&Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In Proceedings of the ACM symposium on applied computing (pp. 784–788)
9. Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (4), 721–735. http://doi.org/10.1109/TPAMI.2008.110
10. Nguyen, T. T., Chang, K., &Hui, S. C. (2013). Supervised term weighting centroid- based classifiers for text categorization. Knowledge and Information Systems, 35 (1), 61–85.
11. Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weight- ing approach. Expert Systems with Applications, 36 (1), 690–701. http://doi.org/10. 1016/j.eswa.2007.10.042
12. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comput. Surv.,34, 1–47.
13. Singhal, A., Buckley, C. and Mitra, M., "Pivoted document length normalization", in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM., (1996), 21-29