MJRET

OPEN ACCESS

Multidisciplinary Journal of Research in Engineering and Technology

www.mjret.in

eISSN:2348-6953

False information detection in Social Media Network by Using Mining Techniques in Machine Learning

[1]Shruti Verma, [2]Dr. Hari Om

[1]M.tech Final year, Rama University, India

[2]Dean of Computer Science & Engineering Department, Rama University, India

Abstract:*The spotting fake news is a relevant problem solving operation. Nowadays Social media is also being used for the news consumption task. Because of some factors or features like- of low cost, easy access and rapid dissemination of the information. Social media lead people to seek out and monopolize the news from the social network. But in another term Social Media network gives a chance for quick spreading of the fake news. So there is probability that low Data Mining techniques play a very essential role in fake news detection. This framework uses classification technique like- Support Vector Machine (SVM), Nave Bayes and Decision trees to detect and classify the news into fake or genuine classes.*

Keywords: *Fake News, Data Mining, Support Vector Machine, Nave bayes, Decision Tree.*

1. INTRODUCTION

Social media may be a very fast-growing thing from the last decade. Most of the knowledge generating today come from social media. In some cases, social media can have the potential of spreading the news more quickly than newspaper Media, TV media. It can cover news that was unable to hide by other media. Generally, this type of faux news is made to market some agenda. Fake news also caused issues like sarcastic articles or fabricated news or some cases pretending to plan government propaganda. It's very possible that two different articles that are similar in their number of words could also be opposite in their meaning.

The detection of false news may be a challenging research issue that ought to be addressed completely and quickly. There's the supply of tons of online websites to see the authenticity of faux news along side the lot of articles that guide to research the suspicious news. Even after the availabilities of those facilities, it's hard to completely control the false news entities. The rationale behind this uncontrolled situation is that the increasing social media platforms. Moreover, the main factor is that fake news also disseminated by bots and cyborgs. These entities can spread and make the content automatically during a quick manner as compared to human. During this manner, fake news are disseminated by both the human and robotic entities. The target of those entities depends on the agenda of disseminates. The fake advertisements or fake reviews can target the web customers, fake health news can target the adulthood people, fake educational news can target the scholars, and forays, etc. during this way, the fake news directly or indirectly can affect human life and may cause suspicious activities.

These suspicious activities are often reduced by improving the prevailing or designing the new research methodologies. The info Mining techniques based decision tree and Support vector machine concept are ensembled to detect the suspicious news.

2. LITERATURE REVIEW

The online social media platforms have the power to influence the lives of human in various positive and negative aspects. The fake news greatly affect the general public opinion and able the change the general final resulting scenario. At the time of sensitive situations, the dissemination of faux news stories can leave the harmful impact on people. There are many existing theory and procedure for the suspicious news detection. A number of the standard contributions within the field of faux and suspicious news are discussed here.

A various number of knowledge mining techniques are applied for fake news detection earlier. Each technique has its advantages. So comparing all techniques is important . Dataset is tested using Support Vector Machines, Bounded Decision Trees, and Nave Bayes.

We propose a way for "fake news" detection and ways to use it on Facebook, one among the foremost popular online social media platforms. This method uses the Naive Bayes classification model to predict whether a post on Facebook are going to be labeled as REAL or FAKE.

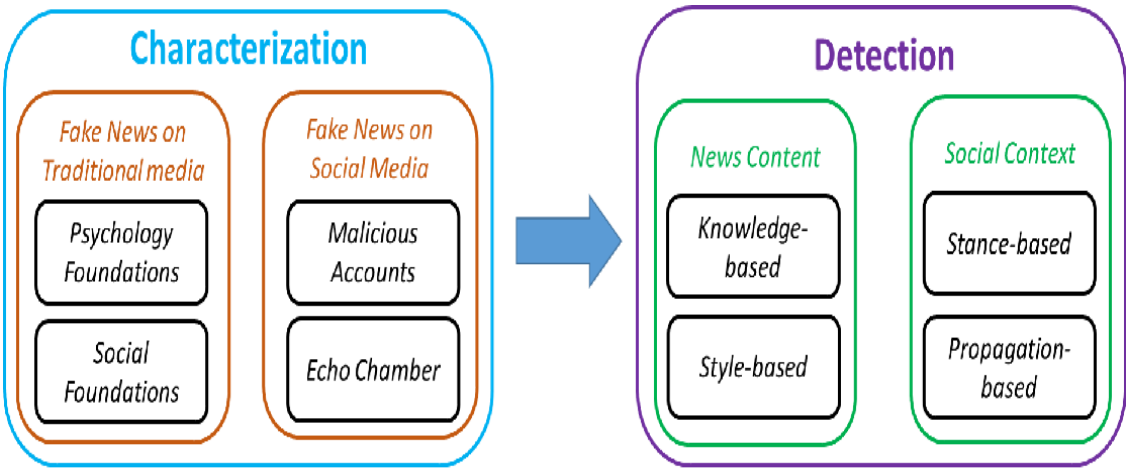

Figure1: List of Fake news based on different detection method

Raw datasets collected for fake news detection usually contains some noise like missing values. The performance of any data processing process depends on the input file . So data preprocessing step plays an important role before applying a knowledge mining concept. Data processing Data preprocessing handles missing values efficiently. Specifically, we've successfully handled the missing value's problem by using data imputation for both categorical and numerical features. Selecting a dataset is a crucial step because the complete process depends on the fields, records, and data of the dataset. The dataset we used is from kaggle fake news challenge -1. The information springs from the Emergent Dataset created by Craig Silverman. It's three fields namely headline, body text, and label. The label shows whether the news is assessed as fake or real.

3. PROPOSED METHODS

Classification

Classification is that the process of learning a target function f that maps each record, x consisting of set of attributes to at least one of the predefined class labels, y. A classification technique may be an approach of building classification models from an input file set. This system uses a learning algorithm to spot a model that most closely fits the connection between the attribute set and sophistication label of the training set. The model generated by the training algorithm should both fit the input file correctly and properly predict the category labels of the test set with as high accuracy as possible. The key objective of the training algorithm is to create the model with good generality capability. The subsequent figure shows the overall approach for building a classification model.

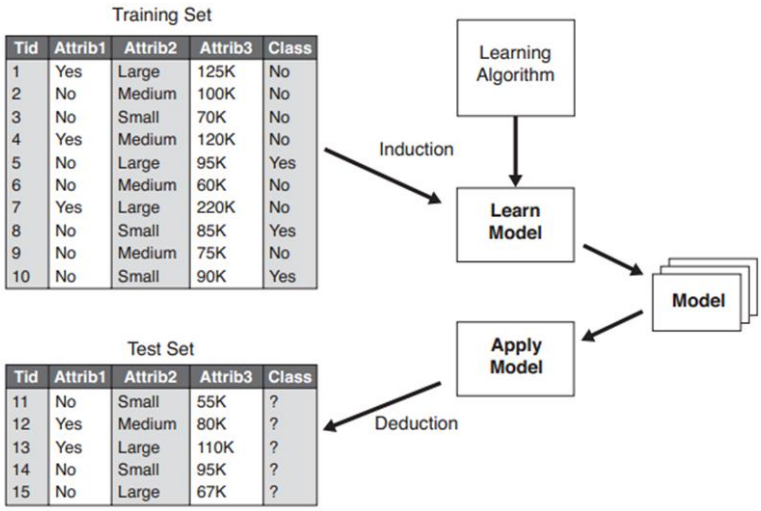

Figure 2.A: General approach for building a classification model

The classifiers that we've implemented for classifying the news are:

- Support Vector Machine
- Naive Bayes Classification
- Decision Tree Classification

All these algorithms are the quality algorithm and is widely utilized in problems like detecting spam email messages, categorizing cells as malignant or benign based upon the results of MRI scans, classifying galaxies based upon their shapes etc.

Support Vector Machine

Support Vector Machine (SVM) is often used for regression and classification problems. In regression, SVM predicts a value, whereas in, classification, it's wont to predict a category label. SVM may be a supervised machine learning algorithm meaning that machine trained with training examples and later trying to predict for brand spanking new test samples. In SVM, an n-dimensional space is employed to plot each data item. Then, a hyper plane which differentiates the classes are used to perform a classification task. An SVM classifies data by finding the simplest hyperplane that separates all data points of one class from those of the another class. The simplest hyperplane for an SVM means the one with the most important margin between the two classes. An SVM classifies data by finding the simplest hyperplane that separates all data points of one class from those of the another

class. The support vectors are the info points that are closest to the separating hyperplane. The figure illustrates linear classification, with + indicating data points of type 1,and indicating data points of type 0.

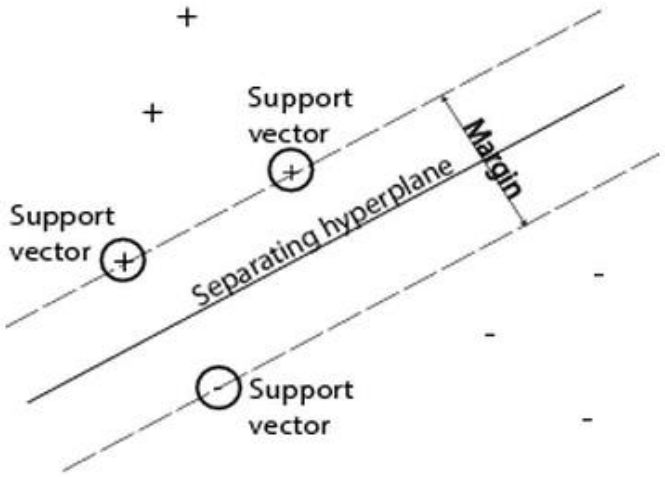

Figure 3.B: Support Vector Machine Classification for two Dimensional data

The datasets that we've used can't be classified using linear classifer. So, non-linear classifier with Gaussian kernel is used. The implementation of SVM is completed on Matlab. Moreover, the benefits of using the SVM method are that it tends to be very accurate and performs extremely well on datasets that are smaller and more concise.

Additionally , this method is extremely flexible since it are often accustomed classify or maybe determine numbers. Also, support vector machines have the potential to handle high dimensional spaces and have a tendency to be memory efficient.

On the contrary, the disadvantages of using the SVM approach are that it's difficulty with large datasets since "the training time with SVMs are often high" and it's "less effective on noisier datasets with overlapping classes." additionally, the SVM method won't "directly provide probability estimates".

Naive Bayes Classification

Naïve Bayes may be a sort of classifier considered as a supervised learning algorithm, which belongs to the Machine Language class and works by predicting, "membership probabilities" for every individual class, as an example, the likelihood that the given evidence, or record, belongs to a particular class. The category with the best, or highest probability, shall be determined the "most likely class," which is additionally referred to as Maximum a Posterior (MAP). Naïve Bayes classifier is that this method uses the "naïve" notion that each features are unrelated. In most cases, this assumption of independence is outrageously false. Suppose Naïve Bayes classifier is scanning a piece of writing and comes across "Barack," in many cases an equivalent article also will have "Obama" contained in it. Albeit these two features are clearly dependent, the tactic will still calculate the possibilities"as if they were independent," which does find yourself overestimating "the probability that a piece of writing belongs to a particular class ". Since Naïve Bayes classifier overestimates the possibilities of dependencies, it gives the impression that it might not work well for text classification.

On the contrary, Naïve Bayes classifier still features a high performances rate even with "strong feature dependencies," since the dependencies will actually find yourself cancelling out one another for the foremost part.

In addition, what makes Naïve Bayes classifier desirable is that it's relatively fast and a highly accessible technique. It is often used for binary or multi class classifications, making it a superb choice for "Text Classification problems". Also, Naïve Bayes classifier may be a straightforward algorithm that only really relies on performing many counts. Thus, it is frequently "easily trained on a little dataset". However, the most important downfall of this method is that it deems all the features to be separate, which cannot always be the case. Hence, there's no relationship learned among the features.

Bayes Theorem is defined as:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where P(A) refers to the probability that event A will occur. P(A|B) stands for the probability that event A will happen, as long as event B has already happened. In Bayesian classification we've a hypothesis that the given data belongs to a specific class. We then calculate the probability for the hypothesis of being true. This is often among the foremost practical approaches surely sorts of problems. The approach requires just one scan of the entire data. Also, if at some stage additional training data is added then each training example can incrementally increase or decrease the probability that the hypothesis is correct.

Decision Tree Classification

A decision tree is a famous classification method that generates tree structure where each node denotes a test on an attribute value and every branch represents an outcome of the test. The tree leaves represent the classes. The figure shows the choice tree evaluated from our training dataset utilized in the project. It displays the relationships found within the traning dataset. this system is fast unless the training data is extremely large. It doesn't make any assumptions about the probability distribution of the attributes value. the method of building the tree is named induction.

One of the foremost widely used classifiers is Decision Tree Classifier. it's also a strong classifier. almost like SVM, Decision Tree also can perform both regression and classification. it's also a supervised learning algorithm. Decision Tree classifiers are more popular because tree analysis is straightforward to know. It divides the given data set into small parts and a choice tree is incrementally constructed. The leaf nodes of a choice tree represent the classification. Decision trees are comfortable with numeric and categorical data.

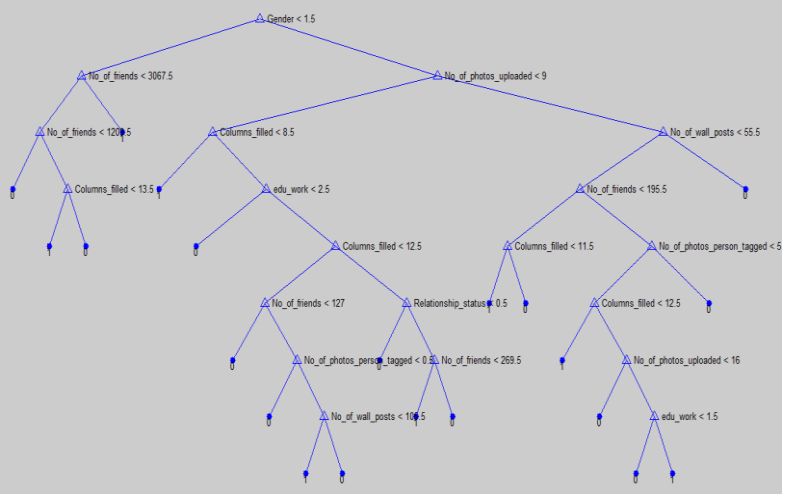

Figure 4.D:Defined Nodes in Tree Structure in Decision Tree Classification Method

4. RESULT AND SOLUTION

We find that the efficiency of the SVM is highest when the data is well trained and the efficiency of the Nave Bayes is lowest which don't change much when the training dataset increases. As the no. of attributes increases for the training dataset the efficiency of all the algorithms increases. The false positive rate of the SVM is least that means if news is detected fake then the chance of being fake is very high in SVM, whereas Nave Bayes shows high false positive rate. The false negative rate on the other hand is very low for Naive Bayes and the SVM has average false negative rate is the algorithm is well trained. So, from the results we find that SVM is well suited for classification of the fake news in the social networks. We discussed different approaches that have been defined in the last few years to overcome the problem of fake news detections in social networks. Most of the approaches based on supervised or unsupervised methods. Those approaches are not providing good results due to non availability of gold standard data set that can help to train and evaluate the classifier and produce good results. Different groups are working now to combat this hot issue and for that purpose they are thinking to utilize actual dataset rather than opinions, blogs. To tackle the problem fake news detection we need to incorporate both behavioral and social entities and to combine knowledge and data.

5. CONCLUSION

Misinformation and fake news always lead to threatening and suspicious outcomes. The increasing social media platforms and growing users may lead to hit the target users with their agenda. The research on fake news is still growing and expanding for the better prediction of such kind of fake news. The most of the existing work is based on individually machine learning and other classification methods which lacks in some cases.

In this research process the concepts of decision tree and ant colony optimization are ensembled and used for the detection of suspicious news. We have given a framework using which we can detect

fake news in any online social network with a very high efficiency as high far around 95%. Fake news detection can be improved by applying NLP techniques to process the posts and the information.

REFERENCES

[1]. 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.

[2]. A. Bondielli, and F. Marcelloni, "A survey on fake news and rumour detection techniques." Information Sciences, Vol. 497, 2019, pp. 38-55.

[3]. J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised Learning for Fake News Detection." IEEE Intelligent Systems, Vol. 34, no. 2, 2019, pp. 76-81.

[4]. Wang, W.Y., 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. CoRR abs/1705.00648.

[5]. C. Wagner, S. Mitter, C. K□orner, and M. Strohmaier. When social bots attack: Modeling

[6]. Susceptibility of users in online social networks. In Proceedings of the WWW, volume 12, 2012.

[7]. Saxena, R. (2017). How the Naive Bayes Classifier works in Machine Learning. Retrieved October 20, 2017, from https://dataaspirant.com/2017/02/06/naive-bayes-classifiermachine-learning/

[8]. ham, L. Transferring, Transforming, Ensembling: The Novel Formula of Identifying Fake News. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019.