International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Fake News Detection using Machine Learning

Nikhil Sharma

Department of Computer Engineering, DCE, Gurugram, Haryana, India

ABSTRACT

Indian politics suffered from a great set back due to fake news. Fake news is intentionally written to mislead the audience to believe the false propaganda, which makes it difficult to detect based on news content. The fake news has hindered the mindset of the common people. Due to this widespread of the fake news online it is the need of the hour to check the authenticity of the news. The spread of fake news has the potential for extremely negative impact on society. The proposed approach is to use machine learning to detect fake news. Using vectorisation of the news title and then analysing the tokens of words with our dataset. The dataset we are using is a predefined curated list of news with their property of being a fake news or not. Our goal is to develop a model that classifies a given article as either true or fake.

General Terms

Fake News, Self Learning, Pattern Matching, Response Generation, Artificial Intelligence, Natural Language Processing, Context Free Grammar, Term Frequency Inverse Document Frequency, Stochastic Gradient Decent, Word2Vec.

KEYWORDS: *Natural language processing, Machine learning, Classification algorithms, Fake-news detection, Filtering*

How to cite this paper: Nikhil Sharma "Fake News Detection using Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.1317-1320, URL: www.ijtsrd.com/papers/ijtsrd31148.pdf

IJTSRD31148

1. INTRODUCTION

This project emphasises on providing solutions to the community by providing a reliable platform to check the Authenticity of the news. The project Fake News Detection using Machine Learning revolves around discovering the probability of a news being fake or real, Fake News mainly comprises of maliciously-fabricated News developed in order to gain attention or create chaos in the community.

In 2016 American election the propaganda carried on by the Russian hackers had the drastic effect on the country, few had supported for President Trump while others didn't but still, due to the spread of the fake news against both presidential candidates Trump and Clinton there was an uproar in the public and moreover the spread of these fake news on the social media had a drastic impact on the lives of the Americans.

After the election results, these fake news had made its prominent way into the market. These have also led into the exclusion of Britain from the European Union i.e. Brexit. During the Brexit time the same fake news propaganda was carried on the internet and due to this a mentality is developed among people that one option is better than another thus leading into the manipulation of the decision of the public and hindering the importance of the democracy. Thus the very foundation on which the countries are operating is disturbed and people don't know whom to believe and whom to not thus the belief system of democratic countries are compromised and people began to think on their own decision whether they took the decision was right or not or the influence of this news was the cause of it? Thus the paper deals with tackling the situation of fake news which has the power to shatter the whole economy of the world and create a "Great Fall".

2. MOTIVATION

Fake news mostly spreads through the medium of social networking sites such as Facebook, Twitter and several others. Fake news is written and published with the intent to mislead in order to damage a person, and/or gain financially or politically. A litany of verticals, spanning national security, education and social media are currently scrambling to find better ways to tag and identify fake news with the intention of protecting the public from deception. Our goal is to develop a reliable model that classifies a given news article as either fake or true. Recently Facebook has been at the centre of much critique following media attention. They have already implemented a feature for their users to check fake news on the site itself, it is clear from their public announcements that they are actively researching their ability to distinguish these articles in an automated way. Indeed, it is not an easy task. A given algorithm should be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. We need to determine what makes a new site 'legitimate' and a method to determine this in an objective manner.

3. LITERATURE SURVEY

A. Mykhailo Granik, Volodymyr Mesyura, "Fake News detection using Naïve Bayes, 2017 " , proposed an approach for detection of fake news using Naïve Bayes classifier with accuracy of 74% on the test set.

B. Sohan Mone, Devyani Choudhary, Ayush Singhania, "Fake News Identification, 2017" proposed system calculates the probability of a news being fake or not by applying NLP and making use of methods like Naïve Bayes, SVM, Logistic Regression.

C. Sholk Gilda "Evaluating Machine Learning Algorithms for Fake News Detection, 2017" proposed system make use of available methods like Support Vector Machines, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees, and Random Forests in order to calculate best available way to achieve maximum accuracy.

D. Sakeena M. Sirajudeen, Nur Fatihah a. Azmi, Adamu I. Abubakar, "Online Fake News Detection Algorithm, 2017" The proposed approach is a multi-layered evaluations technique to be built as an app, where all information read online is associated with a tag, given a description of the facts about the contain.

E. Ver´onica P´erez-Rosas, Bennett Kleinberg, Alexandra Lefevre Rada Mihalcea, "Automatic Detection of Fake News, 2017", proposed system does comparative analyses of the automatic and manual identification of fake news.

4. GAP ANALYSIS

Table 1. Comparison of existing and proposed system

Sr No.	Existing System	Proposed System
1	This system uses tf-idf encoding with statistical machine learning.	This system will use wikipedia Fast Text Word2Vec Embeddings
2	Machine Learning concepts such as Self Learning along with Matching are not used.	Long Short Term Memory (Recurrent Neural Networks).
3	This system performed well but lacks performance with complex news	The systerm performs well on basic as well as complex news

Table 1 - Gap Analysis

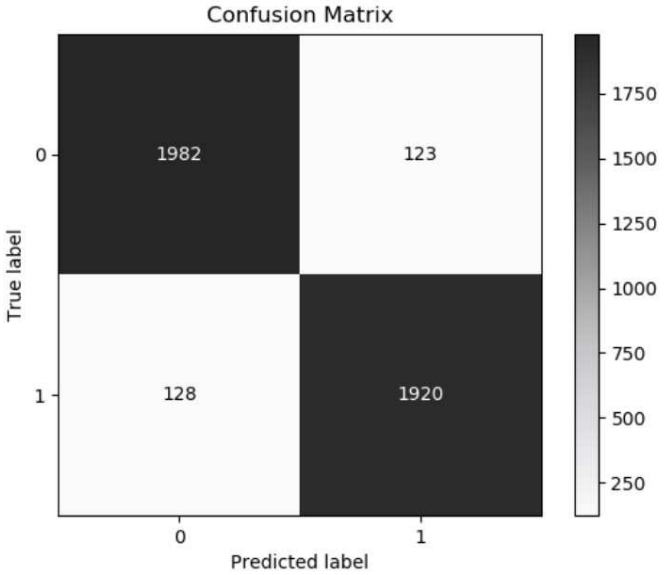

Fig 2 - LSTM

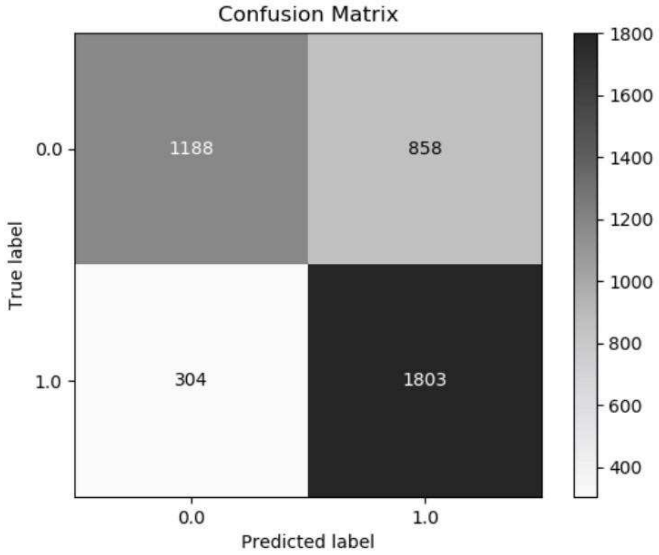

Fig 3 - Naïve Bayes

5. PROPOSED WORK

SYSTEM FEATURE 1 – NEWS GATHERING

We gathered random news on various articles with different subjects to train our model. By studying these, System detects news intent using machine learning algorithm. Pre Labelled news are used to train our models. The Accurate and Best performing model is selected for our predictions. The pre labelled data that we collected is form a reliable resource such as Kaggle. The news collected also contains the class attribute with its corresponding values either true or false on the basis of which it will be determined whether the news is true positive, true negative or false positive, false negative. The class attribute helps in producing the confusion metrics through which attributes like precision, recall etc are calculated in order to evaluate accuracy of the model. The proposed model initially consist of 10,000 different news articles and their corresponding class attributes. Once the news is gathered the model goes to the next feature.

SYSTEM FEATURE 2 - COMPLEX NEWS HANDLING

System will analyse complex news which can be difficult for traditional model. Following steps are required for handling of the complex news, which are as follows Tokenising, padding, encoding, Embedding matrix formation, Model Formation, Model Training and finally predicting the model. The process starts with the tokenising of the input news which is present in the LIAR dataset. The dataset we are using consists of 10,000 news articles with class attribute of each article. In the next process each article/news is taken and is tokenised, in the tokenisation process all the stop words are removed as well as stemming and lemmatisation is also performed.

Second Stage is the Padding the tokens of variable length for this, pad_sequences() function in the areas deep learning library can be used to pad variable length sequences. The default value is 0.0, which is suitable for almost every application, although this can be changed by specifying the preferred value via the "value" argument. The padding to be applied at first or the end of the sequence, called pre- or postsequence padding, can be called the "padding" argument.

Text data requires special preparation before you can start using it for predictive modelling. The text must be parsed to remove words, called tokenisation. Then the words need to be encoded as integers or floating point values for use as input to a machine learning algorithm, called Text Encoding. Once this process of encoding is completed then the text or tokens gets ready for the embedding process.

Embedding is representation for text where words that have the same meaning have similar representation. It is a approach to represent words and documents that may be considered one of the key development of deep learning on challenging natural language processing problems. This transformation is necessary because many machine learning algorithms require their input to be in vectors of continuous values; they just won't work on strings of plain text. So natural language modelling techniques like Word Embedding which is used to map words and phrases from vocabulary to a corresponding vector of real numbers. Word2Vec model is used for learning vector representations of a particular words called "word embedding" . This is typically done as preprocessing step, after which the learned vectors are feed into a model mostly RNN inorder to generate predictions and perform all sort of interesting things. We will be filling the values in such a way that the vector somehow represents the word and its context, meaning, or semantics.

One way is to create co-occurrence matrix. A co-occurrence matrix is a matrix that consist of number counts of each word appearing next to all the other words in the corpus (or training set). Let's see the following matrix.

I	0	1	0	1	1	0
Love	1	0	1	0	0	0
NLP	0	1	0	1	0	0
And	1	0	1	0	0	0
Like	1	0	0	0	0	1
Dogs	0	0	0	0	1	0

Table 2 - Word Embedding Table

We are able to gain useful insights. For example, take the words 'love' and 'like' and both contain 1 for their counts with nouns like NLP and dogs. They also have 1's for each of "I", which indicates that the words must be some sort of verb. These features are learnt by NN as this is a unsupervised method of learning. Each of the vector has several set of characteristics. For example let's take example, V(King) V(man) + V(Women) ~ V(Queen) and each of word represents a 300-dimension vector. V(King) will have characteristics of Royalty, kingdom, human etc. in the vector in specific order. V(Man) will have masculinity, human, work in specific order. When V(King)-V(Man) is done, masculinity, human characteristics will get NULL and when added with V(Women) which having femininity, human characteristics will be added thus resulting in a vector similar to a V(Queen). The interesting thing is that these characteristics are encoded in the vector form in a specific order so that numerical computations such as addition, subtraction works perfectly.
This is because of the nature of unsupervised learning.

SYSTEM FEATURE 3 – FAST TRAINING OF NEW DATA ON GPU

The Proposed System uses Nvidia GPU using CUDA architecture and thus the training of complex real time news becomes easy and faster. Aresa automatically uses the GPU wherever and whenever possible with the help CuDNNLSTM, which is a high level deep learning keras and tensor-flow neural network which runs the model on GPU (Nvidia gpu) using CUDA technology. CUDA is NVIDIA's parallel computing architecture that enables dramatic increases in computing performance by harnessing the power of the GPU (graphics processing unit).Fast LSTM implementation backed by CuDNN. The execution of model training gets faster by 12 to 15 % depending on data.

5.1. FIGURES/CAPTIONS

This diagram depicts the actual working of the proposed system and all the functionalities it will perform. Model formation for fake news detection make use of the training and the test data set and some other parameters like the dimensions of the vector space where it hold the relation between the two or more news entities. All these data is set to pass into the main function which is thought to generate the confusion metrics and present the result in terms of percentage.

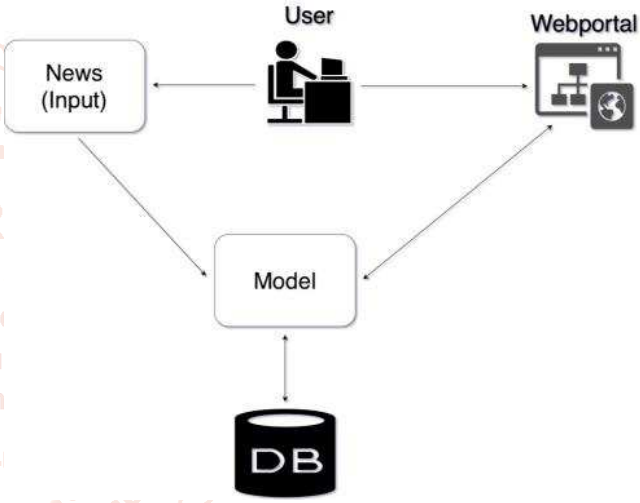

Fig 5 - Working of proposed model

Initially the system stores the gathered news in database which is then retrieved by the model, which then processes the training data and produces the classifier. The user is supposed to enter the news manually which is thought to be unverified, once the input is given via Web-portal it then reaches out to the model in backend who process and gives output. The news given by the user is taken as a test set or test case and is sent to classifier which classifies it.

6. CONCLUSION

The circulation of fake news online not only jeopardises News Industry but has been negatively impacting the user's mind and they tend to believe all the information they read online. It has power to dictate the fate of a country or even whole world. Daily decision of public also gets affected. Applying the projected model would definitely help in differentiating between Fake and Real news.

REFERENCES

[1] Sadia Afroz, Michael Brennan, and Rachel Green- stadt. Detecting hoaxes, frauds, and deception in writ- ing style online. In ISSP'12.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[3] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. Communication Research, 41(3):430–454, 2014.

[4] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 US presidential election online discussion. First Monday, 21(11), 2016.

[5] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In AAAI'16.

[6] Thomas G Dietterich et al. Ensemble methods in machine learning. Multiple classifier systems, 1857:1–15, 2000.

[7] kaggle Fake News NLP Stuff. https://www.kaggle.com/rksriram312/fake-news-nlp-stuff/notebook.

[8] kaggle All the news. https://www.kaggle.com/snapcrack/ all-the-news.

[9] Mykhailo Granik, Volodymyr Mesyura, "Fake News detection using Naïve Bayes, 2017"

[10] Sohan Mone, Devyani Choudhary, Ayush Singhania, "Fake News Identification, 2017".