FAKE NEWS DETECTION USING NLP

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

THE AWARD OF

BACHELOR OF ENGINEERING DEGREE IN COMPUTER SCIENCE AND ENGINEERING BY

EZHILKAVIYA.K

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

STARLION COLLEGE OF

ENGINEEERING AND TECHNOLOGY

MANANKORAI,THANJAVUR.

-614206

OCTOBER -2023

ABSTRACT

With the recent social media boom, the spread of fake news has become a great concern for everybody. It has been used to manipulate public opinions, influence the election - most notably the US Presidential Election of 2016, incite hatred and riots like the genocide of the population. A 2018 MIT study found that fake news spreads six times faster on Twitter than real news. The credibility and trust in the news media are at an all-time low. It is becoming increasingly difficult to determine which news is real and which is fake. Various machine learning methods have been used to separate real news from fake ones. In this study, we tried to accomplish that using Passive Aggressive Classifier, LSTM and natural language processing. There are lots of machine learning models but these two have shown better progress. Now there is some confusion present in the authenticity of the correctness. But it definitely opens the window for further research. There are some of the aspects that has to be kept in mind considering the fact that fake news detection is not only a simple web interface but also a quite complex thing that includes a lot of backend work.

INTRODUCTION

These days‟ fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is “fake news “ but lately blathering social media ‟s discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. The importance of disinformation within American political discourse was the subject of weighty attention , particularly following the American president election . The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper ,it is to produce a model that can accurately predict the likelihood that a given article is fake news. Facebook has been at the critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees ‟s it ; they have also said publicly they are working on to distinguish these articles *in* an automated way. Certainly, it is not an easy task. A given algorithm must 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. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News.

MOTIVATION

We will be training and testing the data, when we use supervised learning it means we are label the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be pre processing i.e. the null values which are not readable are required to be removed from the data set and the data is required to 2 be converted into vectors by normalizing and tokening the data so that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sic kit Learn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

Materials and Methods

In the following, we describe our proposed framework, followed by the description of algorithms, datasets, and performance evaluation metrics.

2.1. Proposed Framework. In our proposed framework, as illustrated in Figure 1, we are expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake.)e ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach.)ere are numerous reputed websites that post legitimate news contents, and a few other websites such as Snopes which are used for fact checking. In addition, there are open repositories which are maintained by researchers [11] to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles.)e datasets are available online and are extracted from the World Wide Web.)e first dataset is ISOT Fake News Dataset [23]; the second and third datasets are publicly. A detailed description of the datasets is provided in Section 2.5.)e corpus collected from the World Wide Web is Pre processed before being used as an input for training the models.)e articles' unwanted variables such as authors, date posted, URL, and category are filtered out. Articles with no body text or having less than 20 words in the article body are also removed. Multicolumn articles are transformed into single column articles for uniformity of format and structure.)ese operations are performed on all the datasets to achieve consistency of format and structure. Once the relevant attributes are selected after the data cleaning and exploration phase, the next step involves extraction of the linguistic features. Linguistic features involved certain textual characteristics converted into a numerical form such that they can be used as an input for the training models. These features include percentage of words implying positive or negative emotions; percentage of stop words; punctuation; function words; informal language; and percentage of certain grammar used in sentences such as adjectives, preposition, and verbs. To accomplish the extraction of features from the corpus, we used the LIWC2015 tool which classifies the text into different discrete and continuous variables, some of which are mentioned above. LIWC tool extracts 93 different features from any given text.

As all of the features extracted using the tool are numerical values, no encoding is required for categorical variables. However, scaling is employed to ensure that various feature's values lie in the range of (0, 1). This necessary as some values are in the range of 0 to 100 (such as percentage values), whereas other values have arbitrary range (such as word counts). The input features are then used to train the different machine learning models. Each dataset is divided into training and testing set with a 70/30 split, respectively. The articles are shuffled to ensure a fair allocation of fake and true articles in training and tests instances.

The learning algorithms are trained with different Hyper parameters to achieve maximum accuracy for a given dataset, with an optimal balance between variance and bias. Each model is trained multiple times with a set of different parameters using a grid search to optimize the model for the best outcome. Using a grid search to find the best parameters is computationally expensive [26]; however, the measure is taken to ensure the models do not over fit or under fit the data.

Novel to this research, various ensemble techniques such as bagging, boosting, and voting classifier are explored to evaluate the performance over the multiple datasets. We used two different voting classifiers composed of three learning models: the first voting classifier is an ensemble of logistic regression, random forest, and KNN, whereas the second voting classifier consists of logistic regression, linear SVM, and classification and regression trees (CART). The criteria used for training the voting classifiers is to train individual models with the best parameters and then test the model based on the selection of the output label on the basis of major votes by all three models. We have trained a bagging ensemble consisting of 100 decision trees, whereas two boosting ensemble algorithms are used, XG Boost and Ada Boost. A *k*-fold (*k*� 10) cross validation model is employed for all ensemble learners. The learning models used are described in detail in Section 2.2. To evaluate the performance of each model, we used accuracy, precision, recall, and F1 score metrics as discussed in Section 2.6.

2.2. Algorithms. We used the following learning algorithms in conjunction with our proposed methodology to evaluate the performance of fake news detection classifiers.

OVERVIEW OF PROJECT

With the advancement of technology, digital news is more widely exposed to users globally and contributes to the increment of spreading and disinformation online. Fake news can be found through popular platforms such as social media and the Internet. There have been multiple solutions and efforts in the detection of fake news where it even works with tools. However, fake news intends to convince the reader to believe false information which deems these articles difficult to perceive. The rate of producing digital news is large and quick, running daily at every second, thus it is challenging for machine learning to effectively detect fake news

FAKE NEWS DETECTION IN SOCIAL MEDIA

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers". Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through click baits. Click baits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyse the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier.

Automatic Online Fake News Detection Combining Content and Social Signals

The proliferation and rapid diffusion of fake news on the Internet highlight the need of automatic hoax detection systems. In the context of social networks, machine learning (ML) methods can be used for this purpose. Fake news detection strategies are traditionally either based on content analysis (i.e. analyse the content of the news) or - more recently - on social context models, such as mapping the news" diffusion pattern. In this paper, we first propose a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing their already high accuracy by up to 4.8%. Second, we implement our method within a Facebook Messenger chat bot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%.

THE SPREAD OF FAKE NEWS BY SOCIAL BOTS

The massive spread of fake news has been identified as a major global risk and has been alleged to influence elections and threaten democracies. Communication, cognitive, social, and computer scientists are engaged in efforts to study the complex causes for the viral diffusion of digital misinformation and to develop solutions, while search and social media platforms are beginning to deploy countermeasures. However, to date, these efforts have been mainly informed by anecdotal evidence rather than systematic data. Here we analyse 14 million messages spreading 400 thousand claims on Twitter during and following the 2016 U.S. presidential campaign and election. We find evidence that social bots play a key role in the spread of fake news.

Accounts that actively spread misinformation are significantly more likely to be bots. Automated accounts are particularly active in the early spreading phases of viral claims, and tend to target influential users. Humans are vulnerable to this manipulation, retweeting bots who post false news. Successful sources of false and biased claims are heavily supported by social bots. These results suggests that curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

thousand claims on Twitter during and following the 2016 U.S. presidential campaign and election. We find evidence that social bots play a key role in the spread of fake news. Accounts that actively spread misinformation are significantly more likely to be bots. Automated accounts are particularly active in the early spreading phases of viral claims, and tend to target influential users. Humans are vulnerable to this manipulation, retweeting bots who post false news. Successful sources of false and biased claims are heavily supported by social bots. These results suggests that curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

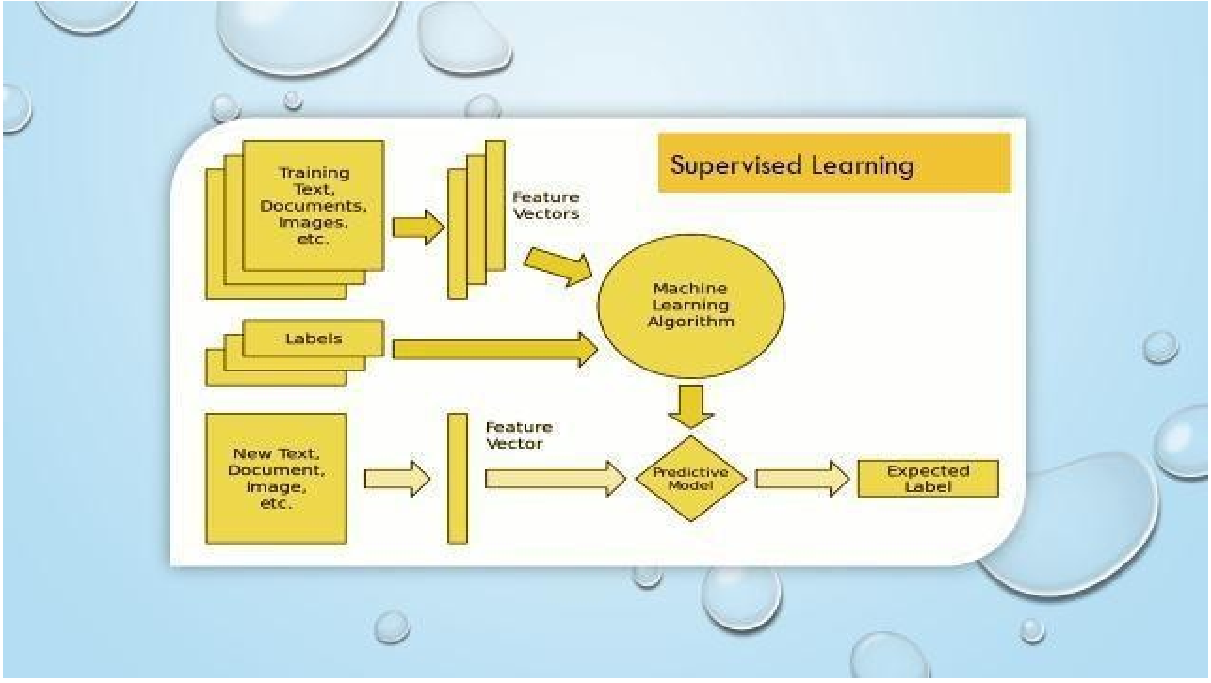
MISLEADING ONLINE CONTENT

Tabloid journalism is often criticized for its propensity for exaggeration, sensationalise, scare-mongering, and otherwise producing misleading and low quality news. As the news has moved online, a new form of tabloidization has emerged: „click baiting.‟ „Clickbait‟ refers to “content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page” [„clickbait,‟] and has been implicated in the rapid spread of misinformation online. This paper examines potential methods for the automatic detection of clickbait as a form of deception. Methods for recognizing both textual and non-textual click baiting cues are surveyed, leading to the suggestion that a hybrid approach may yield best results.

Big Data Analytics and Deep Learning are two high-focus of data science. Big Data has become important as many organizations both public and private have been collecting massive amounts of domain-specific information, which can contain useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. Companies such as Google and Microsoft are analyse large volumes of data for business analysis and decisions, impacting existing and future technology. Deep Learning algorithms extract high- level, complex abstractions as data representations through a hierarchical learning 7 process. Complex abstractions are learnt at a given level based on relatively simpler abstractions formulated in the preceding level in the hierarchy. A key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely un label and un-.categorized

SYSTEM ARCHITECTURE

Fig:


Supervised Learning
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

SCRIPT MODE PROGRAMMING

Invoking the interpreter with a script parameter begins execution of the script and continues until the script is finished. When the script is finished, the interpreter is no longer active.

Let us write a simple Python program in a script. Python files have extension**.py**. Type the following source code in a test.py file –

```
Print "Hello, Python!"
```

We assume that you have Python interpreter set in PATH variable. Now, try to run this program as follows –

```
$ python test.py
```

This produces the following result –

```
Hello, Python
```

Conclusion

Task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques.)e data we used in our work is collected from the World Wide Web and contains news

Table 5: F1-score on the 4 datasets.

	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.98	0.91	0.92	0.87
Linear SVM (LSVM)	0.98	0.32	0.7	0.87
Multilayer perceptron	0.98	0.34	0.95	0.9
K-nearest (KNN)	0.89	0.23	0.83	0.77
Ensemble learners				
Random forest (RF)	**0.99**	0.32	0.95	**0.91**
Voting classifier (RF, LR, KNN)	0.97	0.88	0.94	0.88
Voting classifier (LR, LSVM, CART)	0.96	0.86	0.92	0.86
Bagging classifier (decision trees)	0.98	**0.94**	0.94	0.9
Boosting classifier (Ada Boost)	0.98	0.92	0.92	0.86
Boosting classifier (XG Boost)	**0.99**	**0.94**	0.95	0.9
Benchmark algorithms				
Perez-LSVM	**0.99**	0.8	**0.96**	0.9
Wang-CNN	0.87	0.67	0.31	0.73
Wang-Bi-LSTM	0.84	0.44	0.35	0.57

Logistic
regression
(LR)
Linear SVM Multilayer
perceptron
K nearest
(KNN)
Random
forests (RF)
Voting
classifier
(RF,
LR, KNN)
Voting
classifier
(LR,
LSVM, CART)
Bagging
classifier
(Decision
trees)
Boosting

classifier
(Ada Boost)
Boosting
classifier
(XG boost)
Perez-LSVM Wang-CNN Wang-Bi-
LSTM
Algorithms
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Precision/recall/F1-score
Precision
Recall
F1-score

Figure 3: Precision, recall, and F1-score over all datasets.

articles from various domains to cover most of the news rather than specifically classifying political news.)e primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles using an LIWC tool and used the feature set as an input to the models.)e learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others. We used multiple performance metrics to compare the results for each algorithm.

)e ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.