International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 04, 2023
www.ijirse.com
ISSN: 2454-9665

Fake statements detection made by luminaries byusing NLP and Machine learning

P.Navya, P.Kusuma, P.Ajay

*(Under the guidance of Mr. **M.Chennakesava Rao**, Associate Professor, Head of theDepartment,Department of Computer Science and Engineering, Tirumala Engineering College)*

ABSTRACT

In this project shows an approach for fake statements detection made by luminaries by means of NLP and Machine Learning.Several approaches were implemented as a softwaresystem and tested against a data set of statements. The best achieved result in binary classification problem (true or false statement) is 86%. The results may be improved in several ways that are described in the article as well. The progress in modern informational technologies brings us to the era whereinformation is as accessible as ever. It is possible to find the answers to the questions we are interested ina matter of seconds. Availability of mobile devices makes it even more convenient for the users. This factor changed the way of how people get the news information a lot. Every mainstream mass media hasits own online portal, Facebook account, Twitteraccount etc., so people can access news information really quickly.

Keywords— *Fake statements ,Luminaries ,SVM, Logisticregression, Dcecision Tree, N*

I. INTRODUCTION

The progress in modern informational technologies brings us to the era where information is as accessible as ever. It is possible to find the answers to the questions we are interested in matter of seconds. Availability of mobile devices makes it even more convenient forth users. This factor changed the wayof how people get the news information a lot. Every main stream mass media has its own online portal, Facebook account, Twitter account etc., so people can access news information really quickly. Unfortunately, the news information that we get is not always true. Paradoxically, the Internet makes it harder to factcheck the available information, because there are too many sources that often even contradict each other. All of this caused the emergence of fake news.

Each entry in the data set, besides the statement itself, also contains a lot of metadata. It contains the datewhen the statement was made, the job of the publicfigure who made that statement, the source where the statementwas taken from,some keywords that characterize the content of the statement and many moreother features.

The data set consists of 10460 entries in total (7569 ofthem were provided for training and 2891for testing). There are more than 2000 different sources of the statements. The RAMP studio team collected the data set using Polity Fact website. The Polity Fact is a project operated by Tampa Bay Times in which reporters from the Times and affiliated media fact check statements by members of the United States Congress, the White House, lobbyists and interests groups.

They publish original statements and their evaluations on the Polity Fact.com website, and each "Truth-O-

ISSN: 2454-9665

Meter" rating.

PolitiFact.com was awarded the Pulitzer Prize for National Reporting in 2009for "its fact- checking initiative during the 2008 presidential campaign that used probing reporters and the power of the World Wide Web to examine more than 750 political claims, separating rhetoric from truth to enlighten voters".

There is a belief that fake news problem may be solvedautomatically, withouthuman interference, by means of artificial intelligence. This cause by the rise of deeplearning and other artificial intelligence techniquesshowed us that they can be very effective in solving complex, sometimes even non-formal classification tasks. This article describes a way for classification of short

political statements by means of artificial intelligence.

Several approaches were implemented and tested ona data set of a statement made by real-life politicians.The data set that was used for training and testing wascollected by a RAMP studio team. It contains of short statements made by famous public figures. Two possiblelabels were available for the statement. They are: 1)True

2)False

The steps that were used for the pre-processing are thefollowing:

1) Splitting the statements into separate tokens (words).
2) Removing all numbers.
3) Removing all punctuation marks.
4) Remove all other non-alpha characters
5) Applying the stemming procedure to the rest of thetokens.

A. Classification with logistic regression Logistic regression is a statistical method foranalyzing a data set in which there are one or more independent variables that determine an outcome [8]. The outcome is measuredwith a dichotomous variable (inwhich there are only two possible outcomes). For the cases when there are more thantwo labels, the strategy, which is called “One versusall”, issued.

In this strategy every category is binary classified against its inverse a fictional category that states that the example does not belong to the current category). The category with the highest score is picked as a result of a classification.

Logistic regression is one of the simplest machine learning techniques. It is easyto implement and easy to interpret. It is usually a good idea to implement logistic regression classifier before proceeding with a more complex approach because it gives you an estimate of how well machine learning algorithms will perform on this specifictask.

It also helps to eliminate some basic implementation bugs regarding data set treatment. The results that were achieved for logistic regression classifier are the following:

Classification accuracy-67% Binary classification accuracy-75%

Models that assign class labels to problem instances, represented as vectors offeature values, where the class labels are drawn from some finite set.

ISSN: 2454-9665

It is not a single algorithm for training such classifiers, but a family ofalgorithms based on acommon principle:

1. K-Nearest neighbor algorithm
2. Support vector machine
3 Logistic Regression
4 Decision Tree
5 XgBoost

II. LITERATURE SURVEY

A simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. We achieved classification accuracy of approximately 74% onthe test set which is a decent result considering the relative simplicity of the model. These results may be improved in several ways thatare described in the article as well. Received results suggest, that fake news detection problem can be addressed with artificial intelligence methods. The news information can be easily accessed through Internet and social media. It is convenient for user to follow their interest events available in online mode [1].

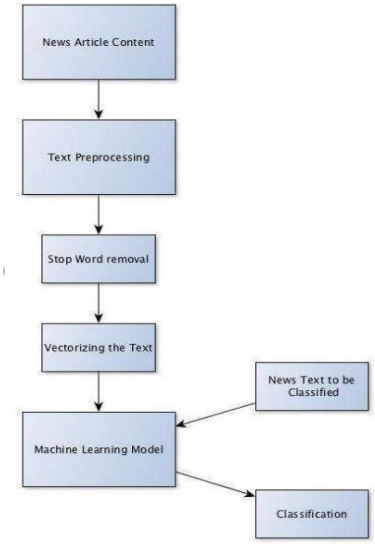
Fact-checking journalism is the heart of Polity Fact. Our core principles are independence, transparency, fairness, thorough reporting and clear writing. The reasonwe publish is to give citizens the information they need togovern themselves in a democracy. Since our launch in 2007, we've received many questions about how we choose facts to check, how we stay nonpartisan, how we go about fact-checking and other topics. This document attempts to answer those questions and many more [8].

The exhaustively of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined; experiments with three test collections showing, in particular, those frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on lessfrequent, more specific, terms are of greater value than matches on frequent terms [7].

In artificial intelligence (AI), new advances make it possible that artificial neural networks (ANNs)learn to solve complex problems in a reasonable amount of time (Lacuna et al., 2015). To the computational neuroscientist, ANNs are theoretical vehicles that aid in the understanding of neural information processing (van Gerven).These networks can take the form of the rate- based models that are used in AI or more biologically plausible models that make use ofspiking neurons (Brette, 2015).The objective of this special issue is to explore the use of ANNs in the contextof computational neuro science from variousperspectives[11].

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 04, 2023
www.ijirse.com
ISSN: 2454-9665

III. Proposed System



First of all it was decided to use only the statements themselves for classification purposes. This means that none of the metadata provided is used for classification. The classification algorithm might actually be improved in the future by taking into accountthis metadata. Splitting the statements into separate tokens (words). Removing all numbers. Removing all punctuation marks. Remove all other non-alpha characters. Applying the stemming procedure to the restof the tokens. In linguistic morphology and informationretrieval, stemming (or lemmatization) is the process of reducing inflected or derived words to their word stem, base or root form – generally a written word form. This helps to treat similar words (like “write” and “writing”)as the same words and might be extremely helpful for classification purposes.

Advantages:

Stop words are the words occur in basically all types of texts. These words arecommon andthey do not really affect the meaning of the textual information. The count vectorizer in NLP made it possible to know the frequency of words occurring.

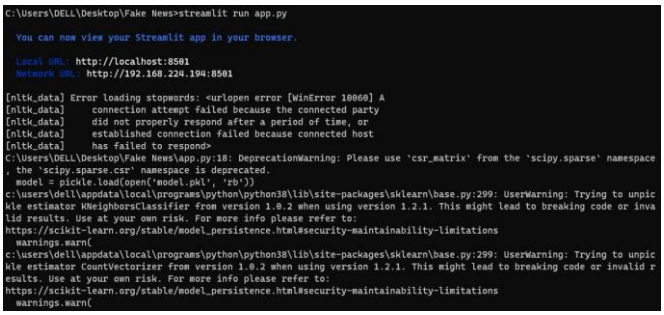
IV. Training and testing the dataset

The dataset is given to the machine learning algorithms to analyse their performance in prediction. The best algorithm is selected for the creation of webpage. Before training, the data should be finite in order to get accurate output..

The machine learning algorithms are given with datasets and their performance is measured. The datasets are trained with their respective parameters and are tested. The best algorithm is furtherly decided for précised output. The algorithms are KNN, Xgboost, Logistic Regression, SVM, and Decision Tree. For lung cancer, Xgboost has high accuracy and hence selected.

International Journal of Innovative Research in Science and Engineering
Volume 9, Issue 04, 2023
www.ijirse.com
ISSN: 2454-9665

V. Results



At first we will open the command prompt as inthe path of the project code file. And then we run our web interface in order to create a network URL and a local URL which will be used for further improvisations.

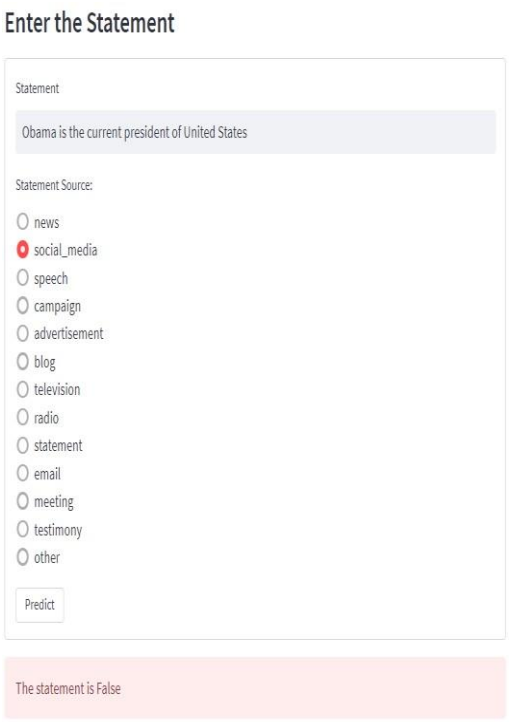
After that the web page will be opened and it displays a user friendly page in order to pass our statements and know whether they are true or false.

Then we will find a place where we need to paste

ISSN: 2454-9665

our statement in order to predict whether it is true or false and get an idea about the statement that is passed in any kind of social media.

If the statement is true it will be displayed as above and in the same way ,if the statement is false the the page will be as follows.



VI. Conclusion

In this paper, several algorithms for classifying statements made by public figures were implemented unsurprisingly, deep neural networks showed the best results both in classification accuracy based on six categories and binary classification. This encourages future research with extensive usage of deep neural networks. Achieved results might be significantly improved. It is possible to both improve the data which is used for training as well as the machine learning models themselves. This might be a subject for future research. Together with the text summarization (the problem that also can be solved by means of artificial intelligence),thisapproach might be used for classification of news articlesas fake or true. This might also be a subject for future research.

VII. Future Enhancements

The major enhancements that can be made in our projectare as follows:

- This system can be trained with more data to givesolution to a wide range of statements.
- We can also improve the accuracy and efficiency moreby using Deep Learning techniques.
- The system can be trained with the data of differentcountries which can be more accurate.