Abstract:
In today's day and age, where data is readily available, not only can we know what is going on around the world with a touch on our smartphones but it is equally easy to be fed lies about that very world. That what fake news is in a nutshell, false information that has been propagated in various ways. This situation has worsened more so since nowadays people tend to receive news from social media rather than a verified news channel for example or even the internet for that matter. In our study we are primarily concerned about the propagation of fake news via social media only. We would be using machine learning and several NLP methods to recognize and isolate such fake news that most probably has been published by unreliable sources. Here we would be collecting data from various sources to construct our very own dataset containing both fake and genuine news and using several algorithms to compare this data and evaluate the legitimacy of the news.

Keywords: Fake news, machine learning, nlp

Introduction:

Perhaps the best way or rather the simplest way of defining fake news is that it is a set of misinformation or lies for that matter. Here we would be essentially using machine learning and natural language processing tools to categorize and identify fake news that come from unreliable sources.

In the pursuit of trying to fit the world in our palms it seems we have lost the ability to differentiate between the truth and the lie. As the internet connects more and more people not only do the number of people that perhaps are helpful and competent enough to help you and others are of more availability than ever before but also are the ones who have aggressive and harmful intentions of spreading such false information sometimes to harm one’s reputation or perhaps just for fun, regardless, this is a problem that just seems to be growing and it doesn’t take alot of thinking to come to the conclusion that this in no way will ameliorate in the future if left as it is.

To combat this problem we have developed this model. Initially several websites are used to collect data by scrapping and these data are used to create a dataset. These data then go through pre-processing and later various models and algorithms are used to differentiate between fake and genuine news that have been stored in the dataset.

LIterature Review:

Since this is a topic that’s growing in its popularity by the day many papers have been published about it.

1. The paper named "Fake News Detection via NLP is Vulnerable to Adversarial Attacks" written by Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat and Justin Hsu demonstrates and speaks about certain techniques to detect such fake news, at the same time speaking about ways that these fake news might already be capable of bypassing the modern NLP detection approaches
2. Another work regarding this topic is "Fake News Detection using Machine Learning and Natural Language Processing" written by Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema. This paper deals with the issue by creating a database of news

marked as either real or fake, this is then used to train the model for future detection of fake news. It uses certain feature extraction techniques such as bag-of-words and TF-IDF and algorithms such as Naive Bayes and others to search for such fake news.

3. A slightly unique take on this topic is found in the paper named "Fake Media Detection Based on Natural Language Processing and Blockchain Approaches" written by ZEINAB SHAHBAZI and YUNG-CHEOL BYUN. In this paper the author proposes the use of NLP to detect texts while the blockchain can be used to verify the authenticity of the data given, by a news channel for example. Using certain blockchain features namely hash values and timestamps and the textual data that has been extracted a model is built to eventually detect fake news.
4. Another quite similar work is " Detection of Fake News Using Machine Learning and Natural Language Processing Algorithms " by Noshin Nirvana Prachi, Md. Habibullah, Md. Emanul Haque Rafi, Evan Alam, and Riasat Khan. This approach also uses a database containing several articles that are already labeled as either fake or real, using this useful features such as sentiment, readability and other factors are compiled and later used to train the ML model which is later used to detect sample articles.
5. Lastly another work that we came across named "FAKE NEWS DETECTION USING MACHINE LEARNING" written by Ms.CH.UMA DEVI, R.PRIYANKA, P.S.SURENDRA, B.S.PRIYANKA, CH.N.D.L.NIKHILA also uses a dataset that contains marked news articles. A model is built on the count vectorizer, a Naive Bayes approach along with Logical Regression is used for text processing.

Fake news detection using Natural Language Processing

A dataset cannot usually be used just as it is and this is true because it contains certain redundancies such as misspellings, repetition or inappropriate words and hence before the dataset is used to build the model it must first be processed and cleaned up of these unnecessary data.

A. Description of the dataset
   An ISOT dataset is used. The dataset was compiled from real-world sources. As we know the dataset primarily contains fake and true articles. The true articles, meaning the ones that are proven to be legit, have been gathered from Reuters.com whereas the fakes ones have been taken from Wikipedia. The dataset contains two CSV files: true.csv and fake.csv, each containing 12,600 files.
B. Pre-Processing
   Even though the dataset has been compiled according to our needs it is anything but ready for implementation, we now move on to pre-processing the data. This is done in a few steps: firstly tokenization is used to break down streams of words into tokens, after this the "stop words" are removed, these are words that add little or no value to the dataset, a list of all the stop words are found in the NLTK library. Capitalization is the

next step, contrary to the name usually all the data is converted to lowercase, at the end stemming and lemmatization are used to complete the pre-processing part.

C. Dimensionality Reduction

Datasets usually come with various columns and features, since most of these features are redundant it can be overwhelming to deal with all these columns and these very columns are what represents the dimensionality of a dataset and so filtering out the necessary columns or “reducing” them is the next step. This is done using the Single Value Decomposition method.

D. Classification Techniques

We use Supervised Learning Technique here. Firstly we use Rocchio Classification which essentially creates “centroids”, which is just an average node calculated from all the nodes in the group, these centroids are then used to set boundaries concerning other nodes. Gradient boosting is one of the most, if not the most, important concepts used here, this essentially uses several weak models along with a strong model to design a final powerful model. Models usually have flaws and this is exactly what the gradient boosting system realizes and uses the next model to compensate for the lacking in the old model. Lastly a Passive Aggressive Classifier is used, unlike other conventional classifiers, these online-learning classifiers can take chunks of data, instead of the whole thing at once, which makes it more convenient for such a huge dataset.

Results

Finally in order to test the accuracy and the practical effectiveness of the model we worked with we ran several simulations and tests to confirm its accuracy. The dataset was divided into two groups namely the training set and the test set, this was done in a 80:20 manner meaning the training set consisted of 80% of the dataset and the test set contained 20%. The results are shown below:

Conclusion

Among an age where information is more available in the history of mankind, the affect of fake news is perhaps the most prominent and significant and therefore must be taken more seriously and in order to actually make a change in this field, we obviously need a deeper understanding in this field and surely require more modified datasets, such as the ones containing only attribute that are needed, and our model actually uses a dimension-reduction approach for this very lack of an ideal dataset or perhaps one even resembling such. After refining the data, if you will, we used classification models such as Rocchio Classification, Bagging, Gradient Boosting Classifier, and Passive Aggressive Classifier to forecast the fake data. Among the tests conducted, we received the highest accuracy of 94.67 percent of the bagging classifier.