International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Fake News Detection Using Machine Learning

[1]M SURESH, [2]THUBATI SRIRAMYA, [3]SAVANI PRAVALLIKA, [4]TIRUMALASETTY SRAVANI, [5]POLA JITHENDRA SAI
[1,2,3,4,,5]Assistant professors, Department of CSE in Narasaraopet Institute Of Technology

ABSTRACT:

This Project comes up with the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news', that is, misleading news stories that comes from the non- reputable sources. Only by building a model based on a count vectorizer (using word tallies) or a (Term Frequency Inverse Document Frequency) tfidf matrix, (word tallies relative to how often they're used in other articles in your dataset) can only get you so far. But these models do not consider the important qualities like word ordering and context. It is very possible that two articles that are similar in their word count will be completely different in their meaning. The data science community has responded by taking actions against the problem. There is a Kaggle competition called as the "Fake News Challenge" and Facebook is employing AI to filter fake news stories out of users' feeds. Combatting the fake news is a classic text classification project with a straight forward proposition. Is it possible for you to build a model that can differentiate between "Real "news and "Fake" news? So a proposed work on assembling a dataset of both fake and real news and employ a Naive Bayes classifier.

1. 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 view points. 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 seeked to produce a model that can accurately predict the likelihood that a given article is fake news.Facebook has been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees'sit ; 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 is. Later, it is needed to look into how the technique.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

2. LITERATURE SURVEY:

N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news,"

This research surveys the current state-of-the-art technologies that are instrumental in the adoption and development of fake news detection. "Fake news detection" is defined as the task of categorizing news along a continuum of veracity, with an associated measure of certainty. Veracity is compromised by the occurrence of intentional deceptions. The nature of online news publication has changed, such that traditional fact checking and vetting from potential deception is impossible against the flood arising from content generators, as well as various formats and genres.

The paper provides a typology of several varieties of veracity assessment methods emerging from two major categories – linguistic cue approaches (with machine learning), and network analysis approaches. We see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network-based behavioral data. Although designing a fake news detector is not a straightforward problem, we propose operational guidelines for a feasible fake news detecting system.

S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," Most previous studies in computerized deception detection have relied only on shallow lexico- syntactic patterns. This paper investigates syntactic stylometry for deception detection, adding a somewhat unconventional angle to prior literature. Over four different datasets spanning from the product review to the essay domain, we demonstrate that features driven from Context Free Grammar (CFG) parse trees consistently improve the detection performance over several baselines that are based only on shallow lexico-syntactic features. Our results improve the best published result on the hotel review data (Ott et al., 2011) reaching 91.2% accuracy with 14% error reduction.

3. System analysis

3.1 Existing System

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature. Conroy, Rubin, and Chen outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech (POS) tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars (PCFG) have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.Feng and First implemented a semantic analysis looking at

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

'object:descriptor' pairs for contradictions with the text on top of Feng's initial deep syntax model for additional improvement. Rubin, Lukoianova and Tatiana analyze rhetorical structure using a vector space model with similar success. Ciampaglia et al. employ language pattern similarity networks requiring a pre-existing knowledge base.

Disadvantages:

 It is not possible to find whether the given data is Real or Fake.

 Fakedata will be increases.

3.2 Proposed System:

In this paper a model is build based on the count vectorizer or a tfidf matrix (i.e) word tallies relatives to how often they are used in other artices in your dataset) can help . Since this problem is a kind of text classification, Implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizervstfidfvectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for countvectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as "the", "when", and "there" and only using those words that appear at least a given number of times in a given text dataset.

Advantages:

 It is possible to find whether the given data is Real or Fake.

 Fake data will be decreases.

4. SYSTEM REQUIREMENTS

Hardware Requirements:

RAM : 4GB and Higher
Processor : Intel i3 and above
Hard Disk : 500GB: Minimum

Software Requirements:

OS : Windows
Python IDE : python 2.7.x and above
: Pycharm IDE
Setup tools and pip to be installed for 3.6.x and above

5. MODULES:

5.1 Algorithms:

 Multinomial Navies Bayes

 Passive Aggressive Classifier

Multinomial Navies Bayes:

Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong(naive) assumption, that every feature is independent of the others, in order to predict the category of a given sample. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest probability will be output. Naive Bayes classifiers have been successfully applied to many domains, particularly Natural Language Processing(NLP).We do have other alternatives when coping with NLP problems, such as Support Vector Machine (SVM) and neural networks. However, the simple design of Naive Bayes classifiers make them very attractive for such classifiers. Moreover, they have been demonstrated to be fast, reliable and accurate in a number of applications of NLP.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Passive Aggressive Classifier:

Step 1

Let's say you're doing a regression for a single data point didi. If you only have one datapoint then you don't know what line is best. The yellow line will pass through the point perfectly and so will the blue line.

Step 2

We can describe all the lines that will go through the line perfectly though. If we make a plot of the weight space of the linear regression (w0w0 is the constant and w1w1 is the slope) we can describe all the possible perfect fits with a line.Note that the blue dot corresponds to the blue line and the yellow point corresponds to the yellow line.

Step 3

Any point on that line is as good as far as didi is concerned so which point has the most value? How about we compare it to the weights that the regression had before it came across this one point? Let's call these weights worigworig. In that case the blue regression seems better than the yellow one but is it the best choice?

Step 4

This is where we can use math's to find the coordinate on the line that is as close as our original weights worig.

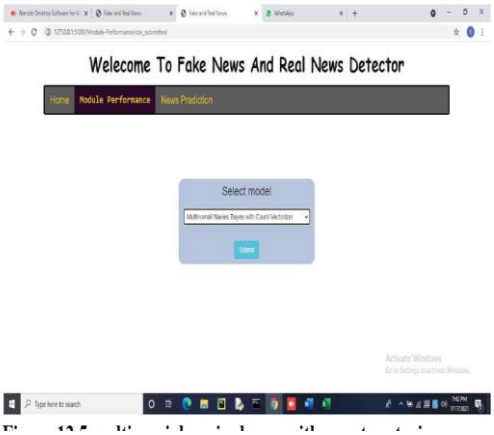
In a very linear system you can get away with linear algebra but depending on what you are trying to do you may need to introduce more and more math's to keep the update rule consistent.

Step 5

To prevent the system from becoming numerically unstable we may also choose to introduce a limit on how large the step size may be (no larger than CC).This way, we don't massively over fit to outliers. Also, we probably only want to update our model if our algorithm makes a very large mistake. We can then make an somewhat aggressive update and remain passive at other times. Hence the name! Note that this approach will for slightly different for system that do linear classification but the idea of passive aggressive updating can still be applied

6. Results

Fig.1.1.fake news detector



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Figure 6.1 multinomial navies bayes with count vectorizer

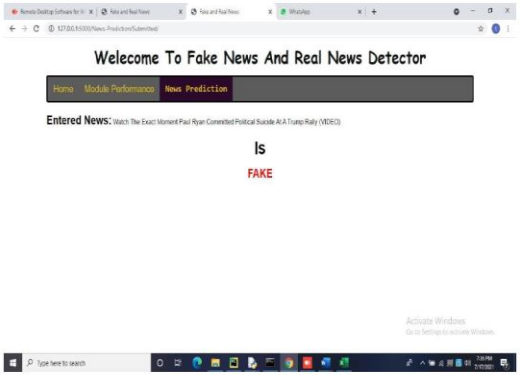


Figure.6.2 5.news prediction fake

Figure.6.3.news prediction real

7. Conclusion

With an increasing focus of academic researchers and practitioners alike on the detection of online misinformation, the current investigation allows for two key conclusions. First, computational linguistics can aide in the process of identifying fake news in an automated manner well above the chance level. The proposed linguistics-driven approach suggests that to differentiate between fake and genuine content it is worthwhile to look at the lexical, syntactic and semantic level of a news item in question. The developed system's performance is comparable to that of humans in this task, with an accuracy up to 76%. Nevertheless, while linguistics features seem promising, we argue that future efforts on misinformation detection should not be limited to these and should also include meta features (e.g., number of links to and from an article, comments on the article), features from different modalities (e.g., the visual makeup of a

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) S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.

3)ShlokGilda,Department of Computer Engineering, Evaluating Machine Learning Algorithms for Fake News Detection,2017 IEEE 15th Student Conference on Research and Development (SCOReD)