International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

Fake News Detection Using Kaggle Dataset and Machine Learning Algorithm

Mohammad Ehzaam[1], Sania Fareed[2], Tahura Nikhath[3], T. Anitha[4],
[1,2,3]UG Scholar, Dept. of IT, ISL Engineering College, Hyderabad,
[4]Assistant Professor, Dept. of IT, ISL Engineering College, Hyderabad.

Abstract

This Project comes up with the applications of Random Forest 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 Random Forest classifier in order to create a model to

classify an article into fake or real based on its words and phrases. The main objective is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news.

Keywords: vectorization, confusion matrix, random forest, Machine Learning

INTRODUTION

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 seemed 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's it; they have also said publicly they are working on to 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 techniques in the fields of machine learning, natural language processing helps us to detect fake news.

LITERATURE SURVEY

With the widespread of social networks, the risk of information sharing has become inevitable. Sharing a user's particular information in social networks is an all-or-none decision. Users receiving friendship invitations from others may decide to accept this request and share their information or reject it in which case none of their information will be shared. Access control in social networks is a challenging topic. Social network users would want to determine the optimum level of details at which they share their personal information with other users based on the risk associated with the process. In this paper, we formulate the problem of data sharing in social networks using two different models: (i) a model based on \emph{diffusion kernels}, and (ii) a model based on access control. We show that it is hard to apply the former in practice and explore the latter. We prove that determining the optimal levels of information sharing is an NP-hard problem and propose an approximation algorithm that determines to what extent social network users share their own information. We propose a trust-based model to assess the risk of sharing sensitive information and use it in the proposed algorithm. Moreover, we prove that the algorithm could be solved in polynomial time. Our results rely heavily on adopting the super modularity property of the risk function, which allows us to employ techniques from convex optimization. To evaluate our model, we conduct a user study to collect demographic information of several social networks users and get their perceptions on

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

risk and trust. In addition, through experimental studies on synthetic data, we compare our proposed algorithm with the optimal algorithm both in terms of risk and time. We show that the proposed algorithm is scalable and that the sacrifice in risk is outweighed by the gain in efficiency.

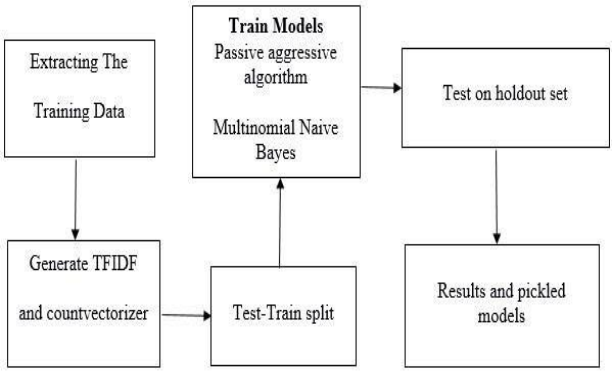
Understand the role ethics plays in standards development and application as well as what questions to ask when deciding whether or not a decision is ethical. Learn what options you have when faced with a standards related ethical dilemma and what resources are available to assist you in making ethical decisions. Gain an understanding of the different possible types of standards abuse and learn about real-life examples of the effects of bad ethical decisions in standards related situations.

by learning the origin of the word distributed representation, knowing the distributed representation is one of the bridges of natural language processing mapping to mathematical calculations. Through the learning distributed representation model: neural network language model, CBOW model and Skip-gram model, the advantages and disadvantages of each model are clarified. Through the experiment, we can understand the mapping relationship between distributed representation calculation and natural language processing, and also clarify the research direction of the next step, that is, the use of distributed representation to sentence, paragraph, article modeling.

PROPOSED METHOD

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 articles in your dataset) can help. Since this problem is a kind of text classification, implementing a Random Forest 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 tfidfvectorizer, 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.

BLOCK DIAGRAM



MODULES DESCRIPTION

Collecting Data

there must be two parts to the data-acquisition process, "fake news" and "real news". Collecting the fake news was easy as Kaggle released a fake news dataset consisting of 13,000 articles published during the 2016 election cycle. Now the later part is very difficult. That is to get the real news for the fake news dataset. It requires huge work around many Sites because it was the only way to do web scraping thousands of articles from numerous websites. With the help of web scraping a total of 5279 articles, real news dataset was generated, mostly from media organizations (New York Times, WSJ, Bloomberg, NPR, and the Guardian) which were published around 2015 – 2016.

Data Pre-Processing

This file contains all the pre-processing functions needed to process all input documents and texts. First we read the train, test and validation data files then performed some pre-processing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Feature Extraction

In this file we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-tdf weighting. We have also used word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

Classification

Here we have built all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. We have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest classifiers from sklearn. Each of the extracted features were used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, 2 best performing models were selected as candidate models for fake news classification. We have performed parameter tuning by implementing GridSearchCV methods on these candidate

models and chosen best performing parameters for these classifier. Finally, selected model was used for fake news detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidfvectorizer to see what words are most and important in each of the classes. We have also used Precision Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

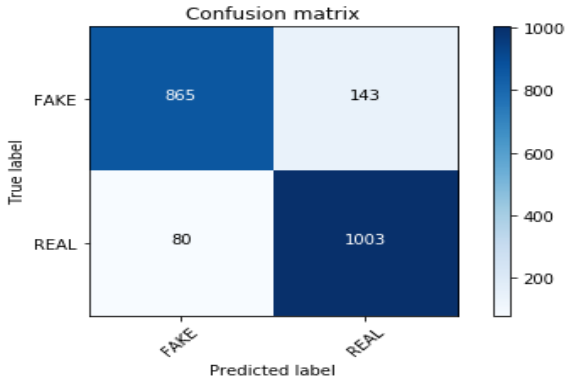
Prediction

Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.sav. Once you close this repository, this model will be copied to user's machine and will be used by prediction.py file to classify the fake news. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of truth.

RESULTS

1. RESULT1

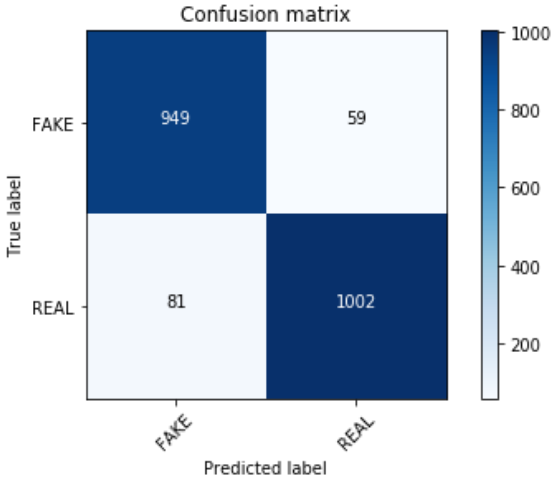
Count Vectorization:



Accuracy: 0.893 Confusion matrix, without normalization

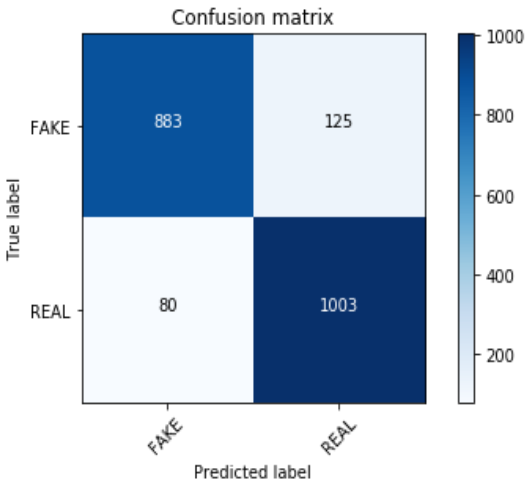
2. RESULT2

Count Vectorizer with Passive Aggressive Classifier:



Accuracy: 0.933, Confusion matrix, without normalization

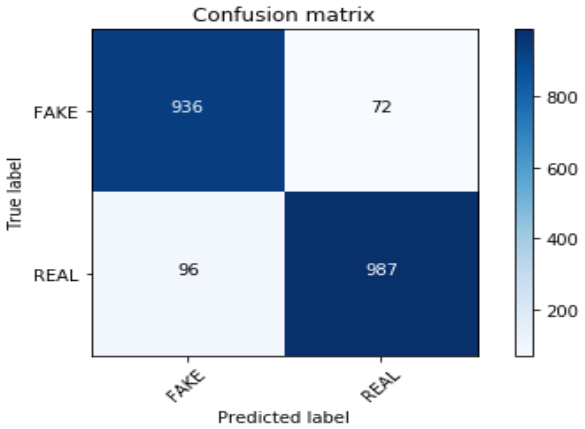
3. RESULT3



Accuracy: 0.902

Confusion matrix, without normalization Hash Vectorization with Passive

4. RESULT4



Accuracy: 0.932

CONCLUSION

For testing the performance, the Sci-kit Learn'sGridSearch functionality is utilized to efficiently execute this task. The optimal parameters for count vectorizer are no lowercasing, two-word phrases no single words, and to only use words that appear at least three times in the corpus. This model's cross-validated accuracy score is 91.7%, true positive score is 92.6%, and its AUC score is 95%.

REFERENCES

[1] Fake news websites. (n.d.)Wikipedia.[Online].Available:https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017.

[2] Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news. [Online]. Available: https://www.wired.com/2016/12/bittersweet-sweepstakes-build-ai-destroysfake-news/

[3] Conroy, N., Rubin, V. and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.

[4] Markines, B., Cattuto, C., &Menczer, F. (2009, April). Social spam detection. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48)

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 10
September 2019

[5] RadaMihalcea , Carlo Strapparava, The lie detector: explorations in the automatic recognition of deceptive language, Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, August 0404, 2009, Suntec, Singapore

[6] Spamming. (n.d.)Wikipedia.[Online]. Available:https://en.wikipedia.org/wiki /Spamming. Accessed Feb. 6, 2017.

[7] Naive Bayes classifier. (n.d.)Wikipedia.[Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier. Accessed Feb. 6, 2017.

[8] Naive Bayes spam filtering. (n.d.)Wikipedia.[Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering. Accessed Feb. 6, 2017.

[9] Craig Silverman, Lauren Strapagiel, HamzaShaban, Ellie Hall, Jeremy Singer-Vine. (2016, Oct.

10). Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. [Online]. Available: https://www.buzzfeed.com/craigsilverman/partisan-fbpagesanalysis?utm_term=.twM44ywz1B#.cxEnnGWD6g