Mining for Credibility

Mahdokht Afravi
Jonathan Avila
mmafravi@gmail.com
jeavila6@miners.utep.edu
University of Texas at El Paso
El Paso, Texas

Cristian Ayub
Gerardo Cervantes
cayub@miners.utep.edu
gcervantes8@miners.utep.edu
University of Texas at El Paso
El Paso, Texas

ABSTRACT

We used the publicly available FakeNewsCorpus dataset that consists of articles labelled as one of 11 types. We used articles labelled as 'fake', 'reliable', and 'conspiracy' to extract features. With this feature set, we train a model using linear regression that would predict whether an article is 'fake' or 'reliable'. Our project site is a GitHub web page that can be found at https://mahdafr.github.io/19s_cs5362-dm/ which contains the source code, presentation slides at http://bit.ly/2YgdAlu, and a PDF copy of this report.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction.*

KEYWORDS

data mining, natural language processing, datasets, fake news, DBSCAN, kmeans, linear regression

ACM Reference Format:
Mahdokht Afravi, Jonathan Avila, Cristian Ayub, and Gerardo Cervantes. 2018. Mining for Credibility. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 111, 2 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Fake news has been a constant concern for as long as news became widespread with the invention of the press in 1439 [2]. As there is not a concrete definition of what fake news means, we will use the data set named FakeNewsCorpus [4]. This data set will allow us to demonstrate that fake and reliable news can be identified despite the meaning people may give to such a concept.

2 THE CORPUS

The FakeNewsCorpus dataset is publicly available. It contains over 9.4 million articles from 745 domains [4]. The dataset is categorized into 11 types, including *Satire*, *Extreme Bias*, *State News*, *Junk Science*, *Hate News*, *Clickbait*, and *Political*. Four of the subsets are the focus of our project:

- ***Fake News***, which contains 928,083 articles. This subset is built from sources that create or distribute false information, or distort factual news reports.
- ***Conspiracy Theory***, which contains 905,981 articles. This subset is built from sources that are "well-known promoters of kooky conspiracy theories".
- ***Proceed with Caution***, which contains 319,830 articles from domains that may or may not be factual.
- ***Credible***, which contains 1,920,139 articles from domains that follow traditional, ethical journalism practices.

Each article contains 15 fields, including *author*, *domain*, *title*, *type*, and *keywords*, which will prove useful to the project's focus.

3 THE PROBLEM

There is a lot of data that will be used for these experiments. One of the problems is going through the data and making sure that it is adequate for use in this project. The data is noisy, which means it might not be suitable to use for the problem and may require further automated cleaning. Data pre-processing might need to be done because it can at times improve the results. A few pre-processing techniques when working with text are: *stopword removal*, *lemmatization*, and *stemming*.

4 THE APPROACH

There are many available methods for analyzing large datasets. However, the most suitable method for this problem at hand is the DBSCAN (*density -based spatial clustering of applications with noise*) algorithm [1]. This algorithm has a few key advantages:

- does not require a specific number of clusters,
- finds shaped clusters,
- has notion of noise, and
- is robust to *outliers*.

After doing a preliminary analysis of the data, we will then work towards creating a model that can predict whether an article is credible or not. We will also evaluate other subsets of data: *Conspiracy Theory*, *Proceed with Caution*, and *Credible*. We will hand-annotate a small sample of this data with whether it is credible or not and then evaluate over it.

5 PHASE I

5.1 Data Clean Up

Working with a smaller dataset will make future

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

5.2 Preprocessing

The FakeNewsCorpus dataset needs to be in numerical representation since it is in text format. Therefore, we implemented the following techniques: stop word removal, stemming, lowercasing, tokenization, and punctuation removal. These techniques are used to reduce the dimension of the given dataset so that is easier to analyze later [3]. For instance, stop work removal consists of identifying and eliminating meaningless words within the dataset, such as articles, prepositions, and conjunctions. However, for this specific technique, we need a so-called stop-list, which is a dictionary of stop words.

6 PHASE II

6.1 Kmeans

We used kmeans clustering [5] to separate the dataset into subsets for further analysis. The labels used for the subsets were: conspiracy, fake, and reliable. The given dataset had 30,000 articles and a dictionary of 20,000 words was used to identify all the words an average student knows and classify the given articles properly in a different cluster. Table 1 shows the clusters obtained from kmeans and how the data is distributed in them.

Articles	Conspiracy	Fake	Reliable
19671	5499	6209	7885
7174	3168	2981	1025
1845	845	411	589
833	367	299	167
314	83	78	153
181	18	1	162
42	14	14	14
18	6	7	5

Table 1: Eight clusters, with number of articles for each label

Table 1 Shows that 8 clusters were created out of the 30,000 articles and these are distributed in all of the clusters. Also, the table shows the different amounts of articles each cluster contains from each of the labels we selected. This clustering technique allowed us to analyze the given dataset by identifying how many articles are indeed reliable and belong to their corresponding label. For this technique we obtained an *average silhouette coefficient* of 0.217, meaning it is accurate but not as much as we would like to.

6.2 DBSCAN

We used DBSCAN [1] to analyze how accurate the clusters obtained with kmeans were since the average silhouette coefficient was not as good as expected and sometimes data might have different shapes and too much noise. The following table shows the results obtained by implementing DBSCAN.

We were able to understand that a higher *eps* allows us to generate better results. We were able to decrease the amount of noise and to measure correctly the distance from noise points to the nearest clusters' core points. Also, we realized that both homogeneity and silhouette coefficient were bad metrics to use with DBSCAN.

Matrix	Clusters	Noise
$15k \times 20k$	2	4.6%
$30k \times 20k$	3	4.3%

Table 2: Clusters created with a given dataset (articles × features)

7 PHASE III

In this phase, the objective a binary output on whether an article is credible or not. The features used for creating our machine learning model were our vocabulary with term frequency. Our training data used has a $10k$ reliable articles as well as $10k$ fake articles. We used 75% of the articles as training data and 25% as testing data.

7.1 Linear Regression

Linear regression was used as a machine learning model in order to classify a given article as reliable or fake. Linear regression

7.2 Results

7.3 Naive-Bayes

8 CONCLUSION

REFERENCES

[1] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering* 60, 1 (2007), 208–221.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[3] C. Silva and B. Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, Vol. 3. 1661–1666 vol.3. https://doi.org/10.1109/IJCNN.2003.1223656

[4] Maciej Szpakowski. 2018. FakeNewsCorpus. https://github.com/several27/FakeNewsCorpus. (2018).

[5] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. 2001. Constrained k-means clustering with background knowledge. In *Icml*, Vol. 1. 577–584.