Possible Misleading Information Identification from News Using Text Analysis

Dobariya Vipula V.

Institute Name & Code: V.V.P Engineering College Rajkot, India

Abstract: *The large use of social media has tremendous impact on our society. Culture, business with potentially positive and negative effects. Now a days, due to the increase in use of online social networks, the fake news for various commercial and political purposes has been emerging in large numbers and widely spread in the online world. The existing systems are not efficient in giving a precise statistical rating for any given news. Also, the restrictions on input and category of news make it less varied. This develops a method for automating fake news detection for various events. We are building a classifier that can predict whether a piece of news is fake based on data source, thereby approaching the problem from purely knowledge based perspective.*

Keywords: Misinformation, Fake news, Fake reviews

1. Introduction

Fake news detection topic has gained a great deal of interest from researchers around the world. When some event has occurred, many people discuss it on the web through the social networking. They search or retrieve and discuss the news events as the routine of daily life.

Some type of news such as various bad events from natural phenomenal or climate are unpredictable. When the unexpected events happen there are also fake news that are broadcasted that creates confusion due to the nature of the events.

Very few people knows the real fact of the event while the most people believe the forwarded news from their credible friends or relatives. These are difficult to detect whether to believe or not when they receive the news information. So, there is a need of an automated system to analyze truthfulness of the news.

During the 2016 US president election, various kinds of fake news about the candidates widely spread in the online social networks, which may have a significant effect on the election results.

According to a post-election statistical report [4], online social networks account for more than 41.8% of the fake news data traffic in the election, which is much greater than the data traffic shares of both traditional TV/radio/print medium and online search engines respectively. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely [4].

This provides an insight into the procedure of detecting fake news. In order to reach a conclusion on the authenticity of the news article, we first take the news event, analyze related data from data sources and then use various classification algorithms to classify the news as legitimate or fake.

2. Literature Review

1) Creating task-generic features for fake news detection [4]
This paper is used for machine learning algorithm.

Feature

a) Statements Domains Score
We propose to use the ranking in order to create a feature we call "statement domains score".
First of all we normalized all URLs. While URLs capture the full path to the contents, we wanted to get information about the information provider.

b) Similarities of title and snippets
• Similarities among all titles for each item in the result set of the query about a statement
• Similarities among all snippets for each item in the result set of the query about a statement
• Similarities between the title and snippet for each item in the result set of the query about a statement.

c) Linguistic inquiry & word count framework
To create groups of features, we used the Linguistic Inquiry and Word Count (LIWC) framework [1], a well-known tool for automatic text analysis. LIWC counts, for a given text, the percentage of words that reflect different emotions, thinking styles, social concerns, and parts of speech, using a validated dictionary.

d) Term frequency and inverse document frequency
To create features based on text analysis, we used TF-IDF (see above) as it is a well-established approach that has proved to be efficient for text classification [2] [3]. We computed term-frequency and inverse document-frequency, with the following parameters
• stop words='english'
• max features=200
• tokenizer=word tokenize
• n-grams=(1,5)
• analyzer='word'.

e) Speaker credit history

In our approach, we analyzed the credit history individually for each speaker, only considering data from the training set. Credit history vectors contain the counts of accurate and inaccurate statements over the history for each class.

2) Fake News or truth? Using satirical cues to detect potentially misleading news. [3]

This paper is used to natural language processing methods in combination with machine learning deal with content directly by detecting language patterns.

Feature:-
a) Word level feature

It is used to lexical and semantic features such as headlines and positive and negative polarity.

b) Sematic validity

True news story found on the web will contain differences in the presence of co-occurring named entities than those entities found in satire.

c) Recognizing irony & sarcasm

They proposed surface features frequent words to represent sarcastic texts.
Ironic expressions often use such markers to safely realize their communicative effects.

3) Unsupervised fake news detection on social media: a generative approach. [5]

This paper are used to two algorithm.
a) Gibbs sampling algorithm: Widely used markov chain monte carlo (MCMC) method to used probability distribution (sample matches to actual desired distribution).
b) Fake news detection algorithm

First randomly initialize the truth estimation of each news to either 0 or 1.

4) Polarization and fake news: early warning of potential misinformation targets[6]

This paper is used to four main phases
a) Data collection

First,we identify two categories of news sources: (1) official and (2) fake— i.e., aiming at fake information. Then, for each category, we collect all data available on the platform under analysis.

b) Sentiment analysis

We associate news and users' comments in the dataset to a sentiment score ranging from totally negative to absolutely positive.

c) Feature definition

We now use the information collected in the previous steps to derive a series of features that take into account how information is presented and perceived on the platform.

d) Classification

Finally, we perform the classification task using different state-of-the-art machine-learning algorithms and compare their results. Once the best algorithms (and the related feature sets) are detected, we are ready to classify entities and, thus, identify potential targets for fake news, i.e., what we define as *early warning*.

5) Supervised learning for fake news detection [7]

This paper detect whether the news is true, false or neutral and gives the results in percentages.

a) Language features

Sentence-level features, including bag-of-words approaches, "n-grams" and part-of-speech (POS tagging) were explored in previous efforts as features for fake news detection.2,6 Here, we implemented 31 features from this set including number of words and syllables per sentence as well as tags of word categories (such as noun, verb, adjective). In addition, to evaluate writers' style as potential indicators of text quality, we also implemented features based on text readability.

b) Lexical feature

Typical lexical features include character and word-level signals, such as amount of unique words and their frequency in the text. We implemented linguistic features, including number of words, first-person pronouns, demonstrative pronouns, verbs, hash tags, all punctuations counts, etc.

3. Methodology

This provides an insight into the procedure of detecting fake news. In order to reach a conclusion on the authenticity of the news article, we first take the news event, analyze related data from data sources and then identify the certain part of the news is misleading or not.
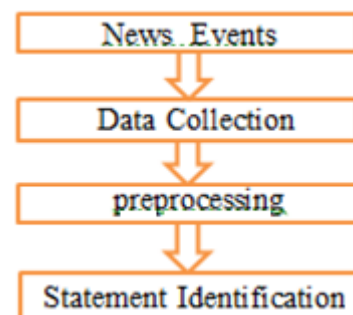
- The basic idea of our method is to build a model that can predict the credibility of real time news events.
- The proposed framework consists of four major steps:

1) News Events
2) Data collection
3) Preprocessing
4) Statement identification

News Events:
- Here we are taking trending news articles as a input.

Data collection:
- Here we are collecting articles from there resources for the same news.

Preprocessing:
- Here we are removing special symbols, hash tags, emojis from all articles (data collection + input news articles).

Statement identification:
- Here we are using NER (Named Entity Recognition) to detect person with statement whichever relate (person warned, person told, person said, person comments).
- Then we compare that statement with other article to see whether they are reflect similar or not.
- If there is the dissimilarity in statement is possible misleading information.

4. Conclusion

Many people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has negative impacts on individual people and society. In this paper, an innovative model for fake news detection using knowledge based algorithms has been presented. This model takes news events as an input and based on twitter reviews and classification algorithms it predicts the percentage of news being fake or real.

References

[1] Shloka Gilda,"Evaluating Machine Learning Algorithms for Fake News Detection" ,2017 IEEE 15th Student Conference on Research and Development (SCOReD).

[2] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON).

[3] Cody Buntain, Jennifer Golbeck, "Automatically Identifying Fake News in PopularTwitter Threads", 2017 IEEE International Conference on Smart Cloud.

[4] Marco L. Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, Luca de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals", ISSN 2305-7254,2017.

[5] Arushi Gupta, Rishabh Kaushal, "Improving Spam Detection in Online Social Networks", 978-1-4799-7171-8/15/$31.00 ©2015 IEEE.

[6] Saranya Krishanan, Min Chen, "Identifying Tweets witj Fake News",2018 IEEE Interational Conference on Information Reuse and Integration for Data Science.

[7] Julio C. S. Reis, Andre correia, Fabricio Murai, Adriano Veloso and Fabricio Benevenuto, "Supervised learning for fake news detection ",2019 IEEE.