Data Mining Project: Proposal

Fake News Detection and Analysis

Team Members:

Saixiong Han: SAH178

Jayashankar Malepati: JAM485

Liju Robin George: LRG37

Proposal:

(1) The problem we plan to do:

We plan to come up with an approach to detect fake news and posts coming from various websites. We plan to use techniques like supervised, unsupervised learnings and other data mining approaches to come up with as good as a model as possible. The task we are going to do can be summarized as following:

a. Predict spam_score according to meaningful features in the dataset, like website, author, etc
b. Identify if the news is fake or not by setting appropriate thresholds.
c. Improve the model performance by feature engineering, topic modeling and ensemble.
d. Find the specific pattern for fake news, like keywords, length or source website.
e. Identify news fields, such as music, sports or politics by text automatically classification and find out the fake news pattern in each field. (Optional, depend on time)

(2) Why the problem is interesting:

Fake news has been quite prevalent during the recent times and considerable amount of issues arise due to the spread of unverified and unacknowledged news. The previous presidential elections showed a great influence of such news in the outcome of the elections. Every country right now is facing the problems caused due to spread of fake news. This is mainly due to the increased involvement of online resources like social media sites and news sites in our daily lives. Even though this is one of the dangerously rising problem in today's highly connected environment, there does not exist any proven approach or methodology to figure out fake news. In this project, we plan to use the data and resources we should, to come up with an approach to battle this issue.

(3) The general approach we plan to take:

The data that we have, has a machine learned score to identify a news being spam or not. This feature is called spam_score and we use this as our target variable to identify whether a post is fake or not. We plan to follow the below given approach to design our model and this project:

a. *Cleaning and Preprocessing:*
The data is from https://www.kaggle.com/mrisdal/fake-news and is in csv format. The data however has considerable amount of overlaps and junk data. We will clean the data, identify and handle missing values and perform necessary conversions. Apart from this, we plan to include a text preprocessing of the title and text (content of post) to enable us to do feature engineering. We plan to perform word normalization, stemming and stop word removal on these two columns. The spam_score feature can be made categorical using a threshold of 0.5: where >0.5 would be considered as fake and < 0.5 would be considered as real.

b. *Descriptive Statistics:*

In this section, once we have all the cleaned and pre-processed data, we perform descriptive analysis on the data and identify all relevant and non-relevant features, correlations, summaries, visualization, normalizations etc. to better understand the data we have in hand.

c. *Feature Selection and Engineering*
We have identified two additional resources; one from Wikipedia and the other from Duran.com, which shows a list of sources that have published fake news or the similar news in the past. We use these resources and create new features for our dataset and use them to come up with our models. Let us call these features as **S**. In addition, from these we would take the text and title columns and try to perform text mining and if possible topic modelling to come up with additional features, we will call these features as **T**. We will create a set of models with and without these features and compare them. Details are given in the next section.

d. *Modelling:*
In this step we plan to create various models from the data and new features we have. We would perform various supervised/unsupervised/ensemble methods to perform the classification:

i. Baseline model: This model will only take the cleaned data and no text/title or the new features. We would see how well this baseline model can perform using training, testing and cross-validation.
ii. Baseline + S Feature model: This model will have the addition of the new features we came up in the previous feature engineering step.
iii. Baseline+ T Feature model: This model will have the addition of the text and title features we came up with.
iv. Full model: This will have all the features

After we have the results from all these models, we pick the best model and try to improve it using various techniques like parameter tuning, ensemble etc.

(4) The kind of data we plan to use and how we are getting the data:
We are planning to use the csv data from https://www.kaggle.com/mrisdal/fake-news coupled with http://theduran.com/updated-list-of-false-misleading-clickbait-y-andor-satirical-news-sources/ and https://en.wikipedia.org/wiki/List_of_fake_news_websites

These datasets are openly available and we already have access to them.

Reference data source:

Kaggle Fake News: https://www.kaggle.com/mrisdal/fake-news

ClickBait: http://theduran.com/updated-list-of-false-misleading-clickbait-y-andor-satirical-news-sources/

Wikipedia Fake News: https://en.wikipedia.org/wiki/List_of_fake_news_websites