Social Computing:Fake News Detection

1. Ajay Jadhav - 2018201095
2. Giridhari Lal Gupta - 2018201019

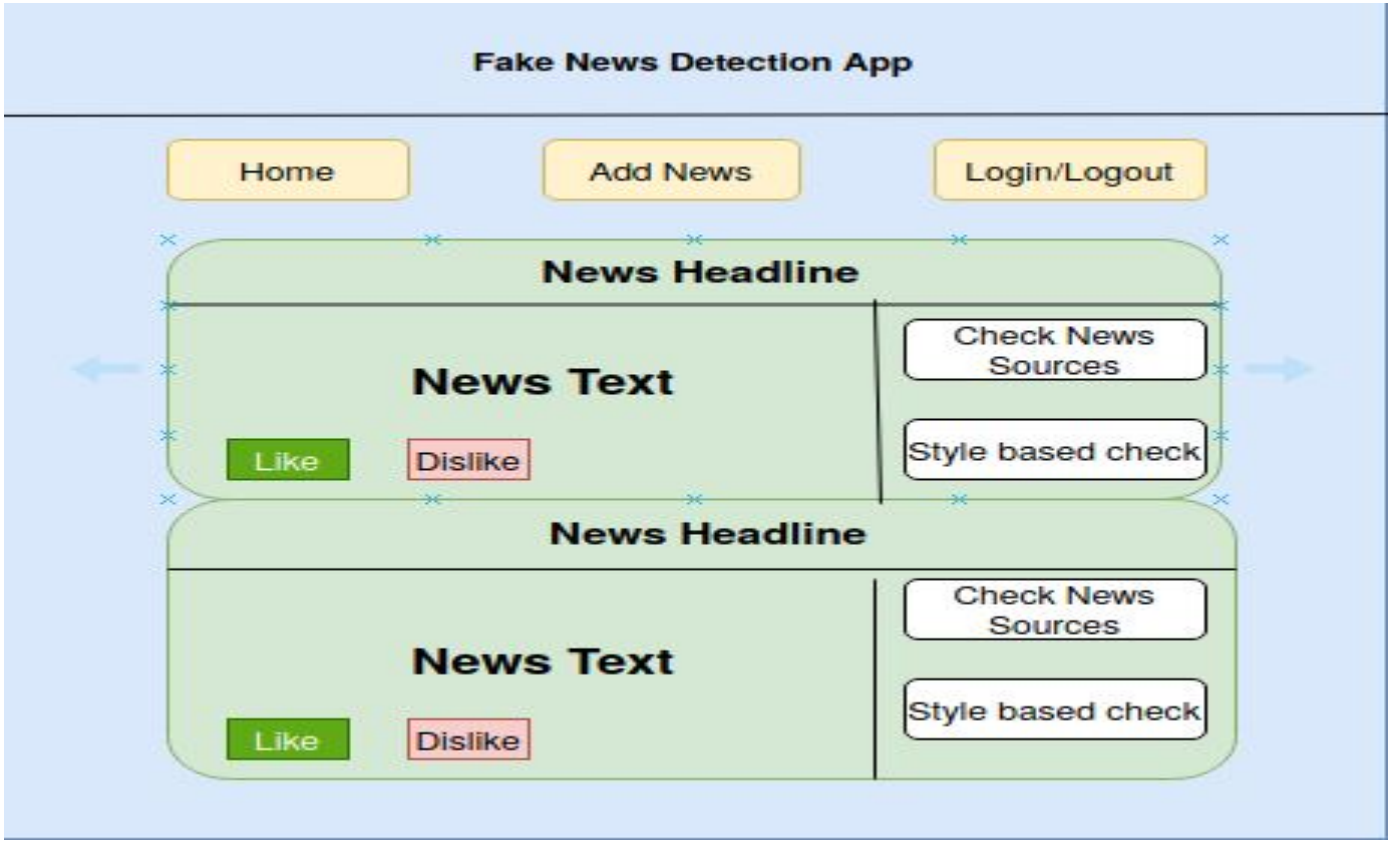
Goal:

A WebApp which performs two way analysis on presented news and presents verdict to user of its genuineness. These methods are based on different paradigm as follows:

1. Content based analysis.
2. Machine Learning based analysis.

Outcome of above two methods will aid user with not only verifying the validity of the information but will also encourage him/her to promote it via votes. Ultimately this multiway filtering will allow genuine information to reach users while fake news and information is

WebApp page and news article design:



Content based method:

- Requirement and achieved outcome
- Methods.
- Data and references.
- Demo

Requirement and achieved outcomes:

R: Whenever a user reads a news article he has to verify the truth of that from multiple sources.

O: We present user with all sources on web where the news has been found.

R: Simply checking news from multiple sources is not enough, to further verify the nature of those sources is equally important. E.g. CNN is left wing, Fox news is Right wing.

O: Many news sources have been tagged by their nature of leaning, e,g left,right,left center,right center,fake. It is saved as domains beforehand in data.

Additional visual aids is also provided to user where overlapped news sources are displayed using graph nodes.

Method:

1. Extract text snippet from news.
2. Break text snippet into n-grams using Pattern module.
3. Reject N-grams having just the stop words
4. A sample of the remaining n-grams is reconstructed into the original strings by re-inserting punctuation.
5. The reconstructed string is passed through Google API.
6. The returned domains are then rated by the amount of queries that returned that domain (more than 6 out of 10 = "high overlap", 3 to 6 = "some overlap", less than 3 = "minimal overlap"), and matched against our database of news sources.
7. The graph is rendered using the Pattern Graph module.

Data and references:

Domain data is extracted from :

1. Opensources.co (github)
2. https://mediabiasfactcheck.com/

The above data is used for calculating overlap and redering overlap graph.

Style based method:

- Need of style based method.
- Achieved outcome.
- Method.
- Data and references.
- Results.
- Demo.

Need of style based method:

The fake news article can be distinguished from writing style as well since most journalists follow AP style guide. There’s tendency of fake news being written by an individual who is not aware of such standards.

Achieved outcome:

We trained Neural net of word representations of news articles from dataset prepared by George McIntire. The model achieved 91% train accuracy and 90% validation accuracy for 70-30 train-test split. These results are obtained by experimentation with number of layers, epochs to avoid underfitting/overfitting model..

Method:

1. Data Cleaning.
2. Tokenization.
3. Padding.
4. Shuffling and split.
5. Word Embedding using Glove vectors.
6. Modelling using LSTM.
7. Train and Evaluate.
8. Save model and weights for future predictions.

Data and references:

- Dataset has 6335 instances of news articles both from fake and real category.
- Each instance of news is represented as (title,text,label) where title is headline of the news, text is content of the news in more detail and label whether the article is fake or real.
- The fake news portion of this dataset was collected from Kaggle fake news dataset comprising news of 2016 USA election cycle. The real news portion was collected from media organizations such as the New York Times, WSJ, Bloomberg, NPR, and the Guardian for the duration of 2015 to 2016.
- Dataset has almost equal number of instances of fake and real news.

Results:

Confusion Matrix.

Crowdsourcing Method

For detection of fake news due to the difficulty of the classification problem which requires fact-checking skills, crowdsourcing can be used as a service to get **better accuracy,** asking crowd-workers to classify the related articles as fake or satiric stories.

For some news if majority of people are saying it is true and some are saying it is false without any evidence we will mark them and then check their activities more.

Thank You!