CMPT 733

Distributed News Monitor System

1. Motivation and Background

In the modern era of internet, the cost of delivering and displaying information publicly is extremely low, and the efficiency of spreading through the social media is absurdly high. On the other hand, these benefits can also enable extensive spread of low quality news with false information. During the 2016 American presidential election, one of the most escalating fake news was the one that claimed Hillary Clinton ordered the murder of an FBI agent and was spread virally on the social media. There were approximately seven million fake news spreading on Twitter around that time [1]. This gigantic number has drawn our attention to the influence of fake news such that it is detrimental to the integrity of information on social media and is capable of leading to mass panicking and misdirection. Therefore, a mechanism of supervisioning and monitoring needs to be introduced.

Fake news is intentionally and specifically written to mislead readers who lack sufficient deterministic information and could voluntarily believe specicial false news, which makes it difficult to persuade them to overthrow preconceived ideas only through a classification result. Our project not only research efficient machine learning models of news classification, but also explore auxiliary information such as user social engagement that can be transformed to provide readers with an accepted method to make a determination.

Researchers have achieved fake news detection from the news sources themselves by determining if the sources are trusted or biased. [2] This detection perspective has the advantage that it only needs a small amount of news to detect the source, which can be better used to avoid extensive spread. However, the lack of analysis on the news content and social media responses may drag down the accuracy for classifying if the news is indeed fake or not. Thus we will focus on the news content to classify a news with the limited availability and amount of resources (i.e., datasets, published literature).

2. Problem Statement

(1) Model: How to classify news with high performance

We aim to classify the news from the content perspective and achieve a better performance. In terms of news content, different words and sentences have varying importance towards the topic and can affect on the determination of news veracity.

- The challenges in modeling are to select pivotal features and an appropriate word embedding model to map text to vectors, and to figure out an appropriate classification model with proper weight balance on both the word level and the sentence level.

(2) Functions: How to guide people to make a determination

In addition to classifying the news as accurately as possible, we expect to lead people to think over the news and guide them to make their own judgements based on our classification and analysis.

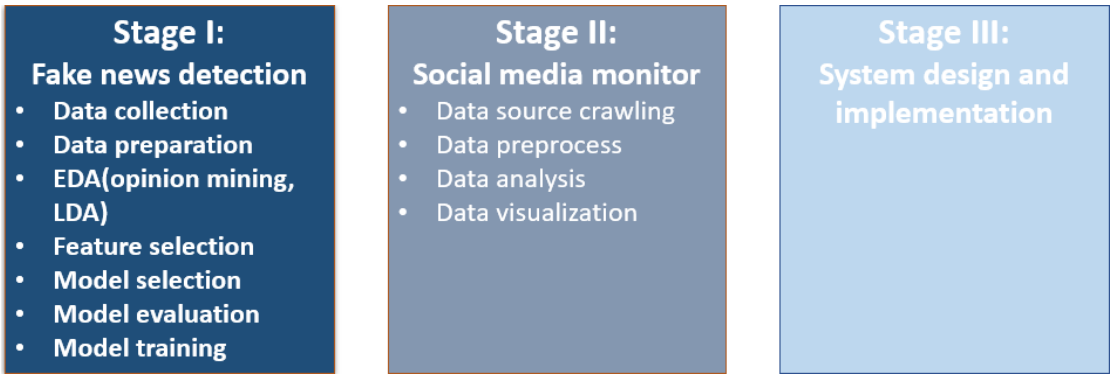
- One challenging part is to dynamically obtain Twitter users' reactions on a news to conduct our real-time analysis.
- Another challenge is to mine valuable information from user social engagement so that it can be used for the public to help make a more confident and assured decision.

(3) System design: How to design a system with scalability and efficiency

Our work involves news classification and social engagement analysis. In the work of social engagement analysis, the comments under all of the popular news need to be crawled in real-time and streaming analytics should be employed to enable applications to integrate certain data into the application flow and to update an external database with processed information.

- The first challenge is the system scalability to handle the growing amount of functions and information because the workload of crawling comments under one news and comments under a hundred news is certainly drastically different in efficiency.
- Exploiting this auxiliary information is also challenging since itself as users' social engagements would produce data that is big, incomplete, unstructured, and noisy, and an efficient streaming data pipeline would improve the overall performance of system.
- Another challenging part of this execution is the integration between the steps in the process in pyspark while all of similar previous implementations of streaming pipeline were written in spark with Scala. So, we have to start from the scratch and research along the way.

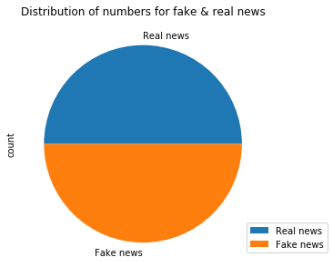
3. Data Science Pipeline and Methodology



Our data science pipeline is composed of three stages: (1) fake news detection(machine modeling for news classification), (2) social media monitor(big data streaming analysis and visualization) and (3) system design and implementation(distributed system for streaming processing) with separate steps within each stages.

Stage I:

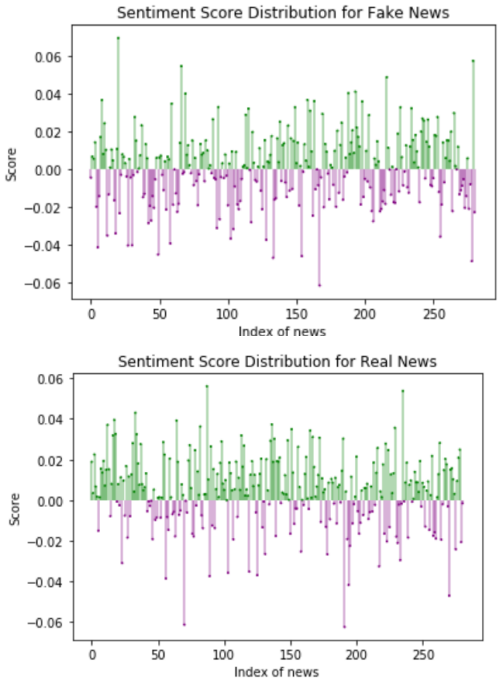
3.1.1 Data collection: News are crawled using newspaper mainly from Snopes website since they are well-labeled by ‘true’, ‘mostly true’, ‘false’, ‘outdated’, ‘miscaptioned’, and ‘mixture’. 562 news labeled either ‘true’ or ‘false’ are randomly chosen and the numbers of fake news and real news are balanced. For each news, news URL and full content are kept for the following steps.



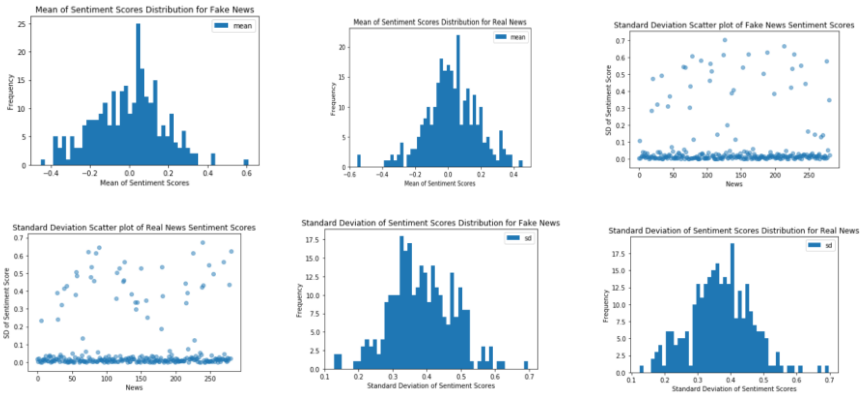
3.1.2 Data preprocess: Based on the crawled news content, we split the news document into sentences and this sentence-level file is kept in preparation for further feature selection. We also tokenized sentences into single words using NLP, dropping stopwords at the same time and finishing morphology, in preparation for the modelling fitting.

3.1.3 Exploratory data analysis and feature selection: Initial brief summation over the dataset is conducted in this phase. We apply a topic model using Latent Dirichlet Allocation(LDA) [3] to get an overview about topics around fake news and real news. They are clustered into ten topics respectively, which can be shown in the following graphs. We observe there exists obvious similarities between certain topics for the fake news, which means fake news are produced mainly on specific aspects such as famous politicians.

Since the characteristics of official news turn to be balanced and objective[4], this implies that the content should be written in a factual, accurate and concise way with few subjective opinions towards the events. Thus, we propose to apply sentimental analysis on the news from both word level and sentence level. In word-level (average all sentiment scores through all words in each news), observing that there are higher amount of score points lying on the positive side for real news than for fake news, we conceive that sentiment scores of news would likely be a feature for classifying news. However, this assumption is proved to be insufficient through further investigation. It's noticed that these values are very small (-0.06 ~ 0.06) because most words in each news are neural that neutralize the final average value. Thus sentence level analysis should be considered.



With regard to the sentiment score sequence of each news (each item in sequence is the aggregated sentiment score for each sentence of news), more analysis are shown in the following. Take standard deviation distribution as an example, from the histogram, the variation for both fake news and real news are similar. Overall, they have extremely similar spread in the same range and both center around 0.4. This basic statistics index shows a sign that sentiment scores or their variance will not have notable effect on classifying a news.



3.1.4 Data modeling:

Based on our initial goal on classifying news from content, we apply a bi-directional RNN with attention mechanism to the news. This mechanism leads the model to give different weights to individual words and further, different sentences.

3.1.4.1 Hierarchical Attention Networks[5]

Hierarchical attention network along with its supported parts: sequence encoder and hierarchical attention should be introduced. The sequence encoder is a stack of GRU cells that each propagates an input within a sequence and pass it into the next hidden states with updated and reseted controls. Hierarchical attention structure can be divided into four parts: a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level encoder. In terms of word sequence encoder and sentence encoder, single words are embedded into vectors and bi-directional GRU takes the word embedding from both forward and backward directions of the sentences to get this two hidden states. By concatenating the hidden states for one word, we are able to get the hidden annotation of a specific word with the summarization of nearby words. The sentence encoder works in the way. The forward and backward bi-directional GRU take in the sentence embedded vector respectively and concatenate to get the hidden annotation emphasised on the current sentence. With regard to the word-level and sentence-level attention, a context vector is randomly chose and used to be compared with when finding attention weights. The hidden annotations from the previous step are taken into a multilayer perceptron and we compare this obtained hidden representations with the previous mentioned context vector.

3.1.4.2. Model fitting and training process

562 fake and real news are randomly separated to training data set(448 news) with test data set(112 news) for 5 times(i.e: 5 training set with 5 test set). Top 16000 words are selected and ranked by its own tf-idf. The maximum number of sentences for a news is fixed to be 100 and we only retain the words appearing within these 16000 words. We then load the pre-trained word2vec model to initial the word embedding matrix. For hyperparameters, we set the embedding dimension to be 300 and the GRU hidden dimension to be 50. In the training process, we use the batch size of 50 and use the gradient descent to train the model.

Stage II:

3.2.1 Data source crawling: News API is adopted to crawl daily news headlines as the preliminary step. Then, the power of Name Entity Recognition will aid us in extracting the most popular daily news keywords which will be triggered once a day. According to these keywords, we are off to the hunt for the Twitters that are related to keywords. However, there are still too many Twitters to crawl, some of them might have a couple likes whereas the other one might has more than a thousand reactions, so standards and criteria must be set. Considering the feasibility and scalability, only the Twitter that receives more than a thousand reactions(i.e. number of comments and likes combined) will be crawled and analyzed. Then, we use Twitter Search API to scrape news URL, Twitter URL, comments, and etc. for further cleaning, preprocessing, and analyzing.

3.2.2 Data Preprocessing: We save crawled data as dataframe for convenience of aggregated computation. From the crawled columns, the comments analysis part would receive these following columns for further analysis: Twitter content, comments content, news keyword, news URL, Twitter URL, retweet count, and likes count. It would trigger the distributed web crawler every minute to scrape the real-time comments off the Twitter that we are interested in. From on the crawled comment content, each comment is tokenized into single words using NLTK, dropping stopwords at the same time of finishing morphology, to prepare for further analysis and visualization.

3.2.3 Data Analysis: For data analysis, the NLTK sentiment analyzer is utilized while taking in the tokenized comments and outputting the polarity sentiment score for each comment. A common consensus is that the more likes a comment receives the more likely the public is to agree with this comment. Therefore, a weighted average feature is introduced to the sentiment analysis such that the sum of the total number of likes across all the comments plus the number of comments is calculated first. Then each comment receives a weight that is equal to its number of likes plus one(one is added to symbolize the viewpoint of the person who posts it) divided by the sum from above as shown in equation(1). After calculating this weight for each individual comment, it is multiplied with each respective aggregated polarity sentiment score. Finally, these weighted scores will be summed to arrive at the final result that is between [-1, 1] to represent the general attitude of the public toward a matter. The closer it is to 1, the more positive the reaction is, whereas the closer it is to -1, the more negative the responses are, and 0 stands for a more neutral outlook. This is constantly performed on the fly, utilizing the Spark streaming function to achieve a real-time output.

3.2.4 Data Visualization: All the columns that are crawled and calculated are all stored in MongoDB for better visualization speed. On the web page the following items will be shown for each individual news: a calculated weighted sentiment score, top ten comments with the most likes, a word cloud, a line chart that shows the number of comments against time for a more clear visualization. As for the word cloud, POS-tagging is used to only count the adjectives and thus displaying them. The scatters of all polarity sentiment scores for each news is also plotted to visualize the attitudes from all readers to the original tweet.

Stage III:

3.3 System Design and Implementation: Since our goal is to build a well-rounded pipeline from start to finish with considerable efficiency, feasibility,and scalability. Spark streaming function is implemented between the components so that the result can be as real time as possible. Kafka is used when triggering the news classifier which it happens once a day, as well as when triggering the distributive web crawler every minute for the comments analysis. After these components have finished the calculation and determination from the model, the result would be passed through the Kafka streaming into MongoDB for our frontend web. Every news has its own unique

identifier, the analysis from these two components would merge into one single document.

Moreover, we choose to implement a distributed web crawler because of scalability and the possibility to exploit more options for further investigations later on. Therefore, a separate distributed web crawler for each component written in PySpark is deployed in the pipeline.

Techniques and Technologies List

Web Scraping: news API, Python newspaper, Twitter Search API, Selenium.

Data preprocessing: NLP(NLTK, corpus, SentimentIntensityAnalyzer, SpaCy, en_core_web_sm), Pandas, pyspark ML Pipeline, pyspark SQL.

Data modeling, analysis: Keras, Tensorflow, gensim, numpy, scipy, sklearn.

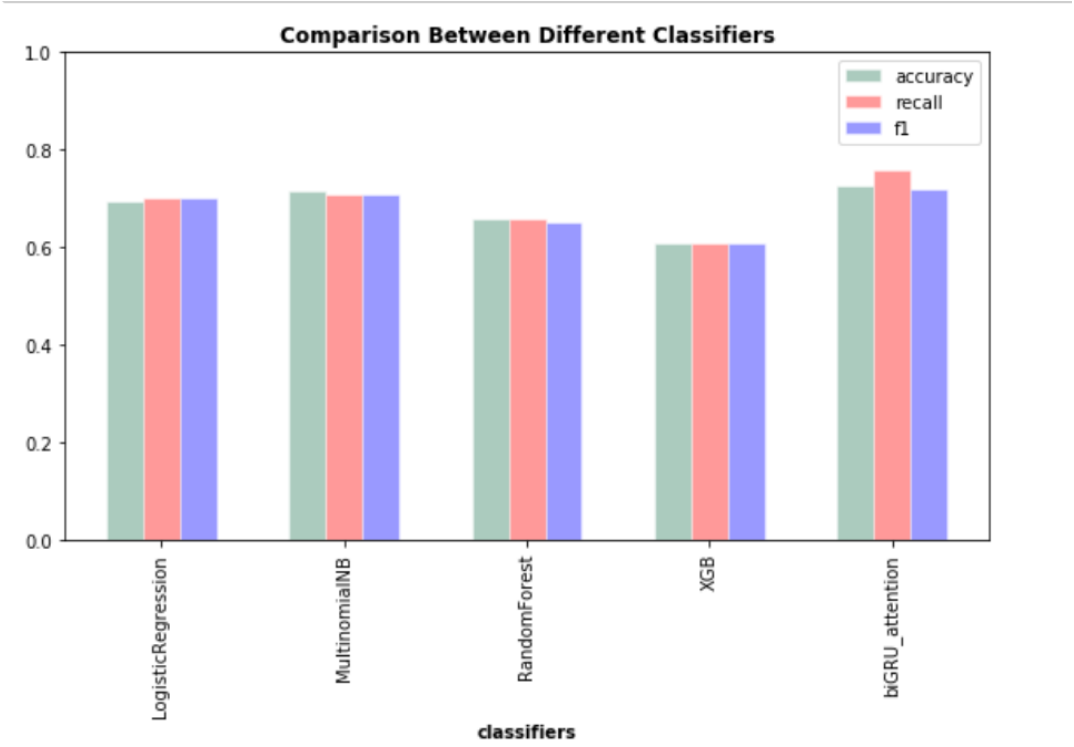
Data visualization and Web UI: seaborn, matplotlib, Node JS, React, RESTful API.

Data storage and stream-process: MongoDB, Apache Kafka.

Evaluation

1. Model

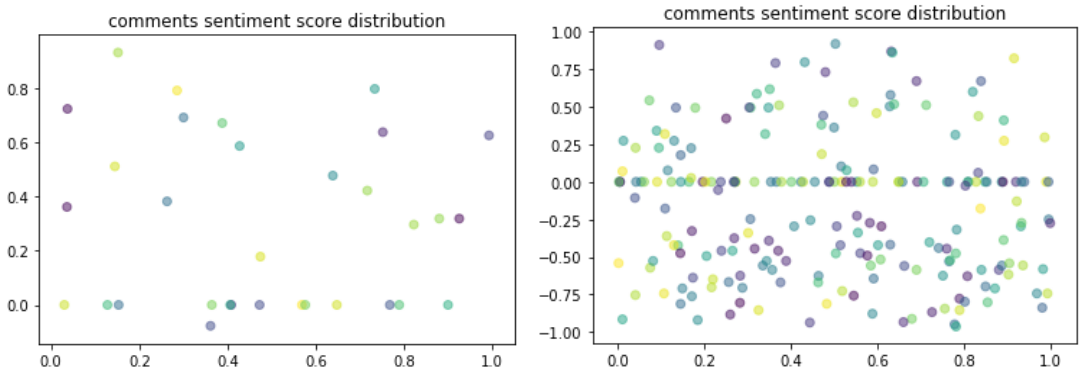
In addition to the bi-directional RNN model described above, we applied four additional model which are logistic regression, multinomial Naive Bayes, random forest and XGB. Accuracy, recall and f1 score are used as the criteria to compare the performance of these classifiers. The bi-directional RNN model has better performance over the other classifiers under all these three statistics.



Classifier	Accuracy	Recall	F1 score
Logistic Regression	0.69	0.70	0.70
Multinomial Naïve Bayes	0.71	0.71	0.71
Random Forest	0.66	0.66	0.65
XGB	0.61	0.61	0.61
biGRU_attention	**0.73**	**0.76**	**0.72**

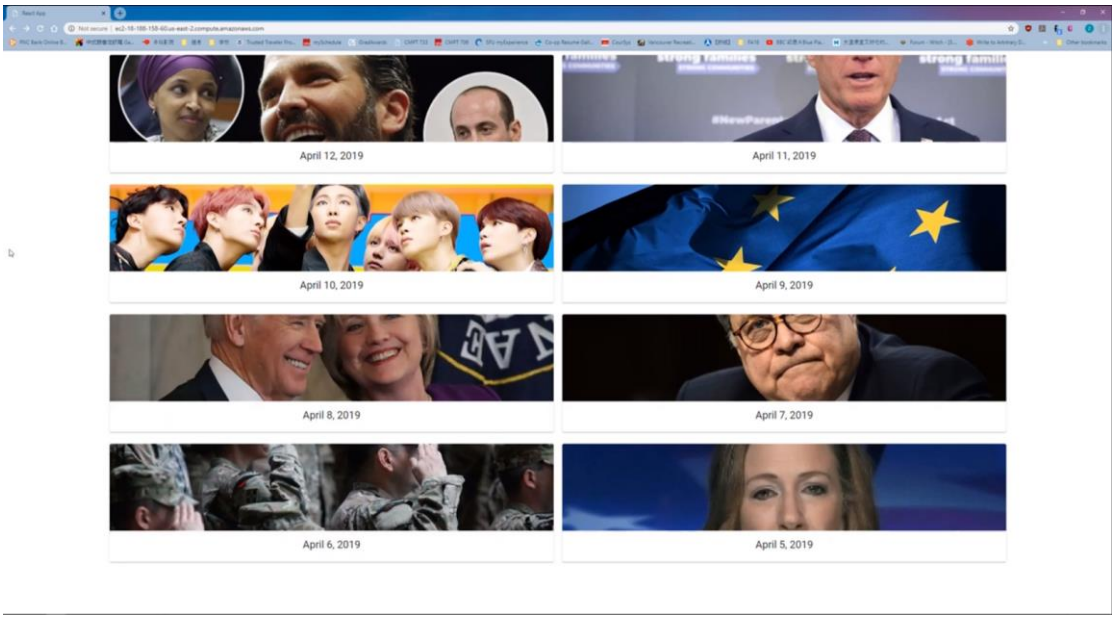
(2) Streaming analysis

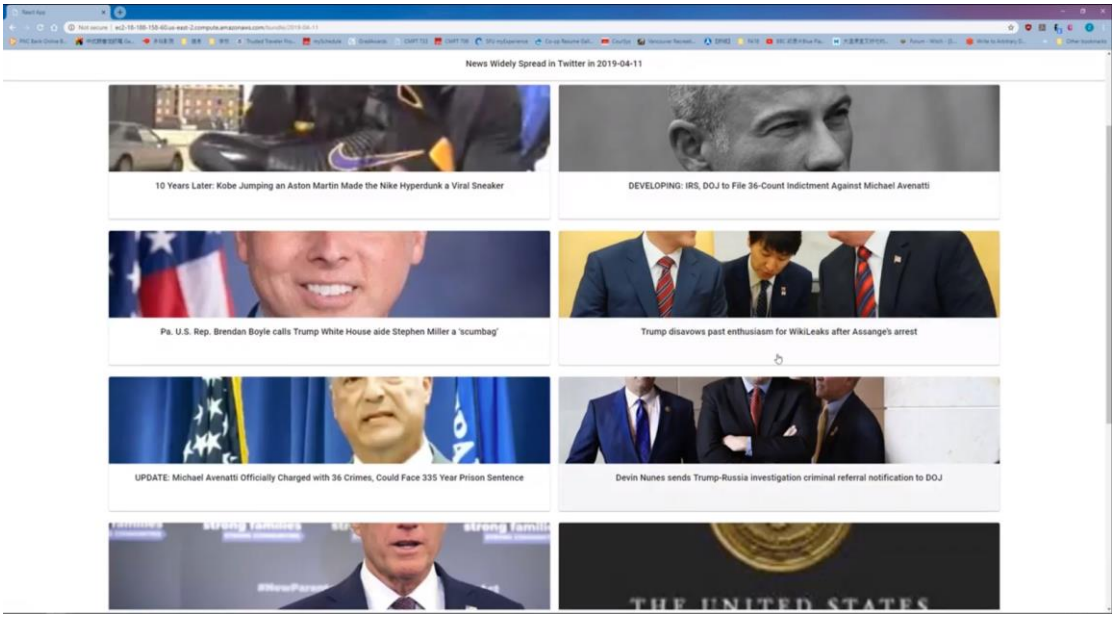
Among these streaming analysis results, the sentiment score distribution of tweets comments shows a meaningful finding. Even though we found that the content of either real news or fake news is not polarized, the polarization/bias among the user's comments plays a key role in misinformation spreading on online social media. Left graph shows the distribution for fake news and readers' comments are biased toward positive, while right graph shows a more diverse discussion among Twitter users. The novelty of our finding consists in taking into account characteristics related to users' attitudes to news on social media, which can be used to assist other reader to making a decision and facilitate the mitigation of misinformation phenomena.

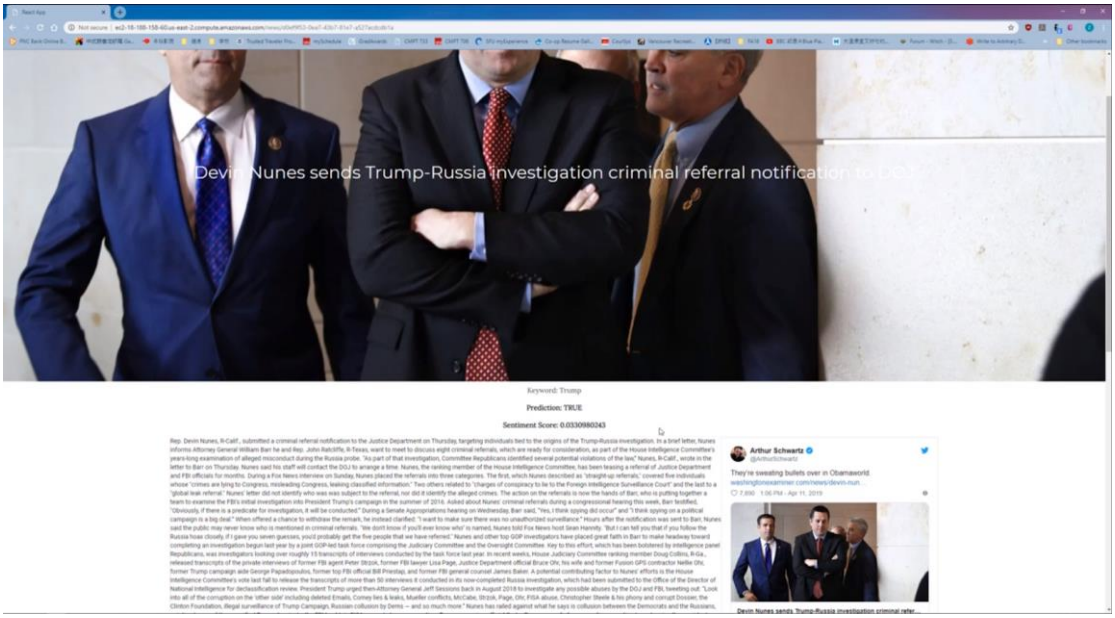


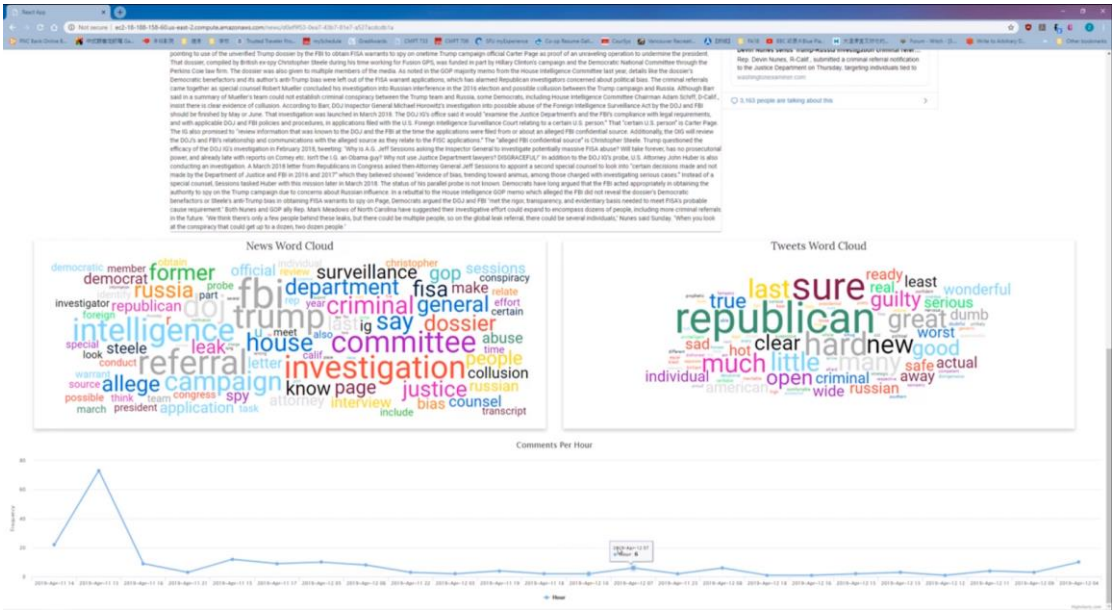
Data Product

This Distributed News Monitor System along with the web visualization is the data product. It is capable of collecting the users' interactions in real-time, deriving to more information from those, automated decision-making on the reliability of the news, and finally displaying all of those in one on the webpage. The whole system along with its webpage is uploaded and hosted on the AWS server. There are three main pages. The first page lists the hottest news with their titles for the recently eight days. In the web page for each separate day, the headline is the hottest news and the other news according to this day's keywords are also shown. Inside each news, the classification label and the weighted sentiment score are shown. Moreover, news content, tweets, the wordcloud and the line chart of comment frequency are presented.




News Widely Spread in Twitter in 2019-04-11
10 Years Later: Kobe Jumping an Aston Martin Made the Nike Hyperdunk a Viral Sneaker
DEVELOPING: IRS, DOJ to File 36-Count Indictment Against Michael Avenatti
Pa. U.S. Rep. Brendan Boyle calls Trump White House aide Stephen Miller a 'scumbag'
Trump disavows past enthusiasm for WikiLeaks after Assange's arrest
UPDATE: Michael Avenatti Officially Charged with 36 Crimes, Could Face 335 Year Prison Sentence
Devin Nunes sends Trump-Russia investigation criminal referral notification to DOJ


Devin Nunes sends Trump-Russia investigation criminal referral notification to DOJ
Keyword: Trump
Prediction: TRUE
Sentiment Score: 0.0330980243
Rep. Devin Nunes, R-Calif., submitted a criminal referral notification to the Justice Department on Thursday, targeting individuals tied to the origins of the Trump-Russia investigation. In a brief letter, Nunes informs Attorney General William Barr he and Rep. John Ratcliffe, R-Texas, want to meet to discuss eight criminal referrals, which are ready for consideration, as part of the House Intelligence Committee's years-long examination of alleged misconduct during the Russia probe. "As part of that investigation, Committee Republicans identified several potential violations of the law," Nunes, R-Calif., wrote in the letter to Barr on Thursday. Nunes said his staff will contact the DOJ to arrange a time. Nunes, the ranking member of the House Intelligence Committee, has been teasing a referral of Justice Department and FBI officials for months. During a Fox News interview on Sunday, Nunes placed the referrals into three categories. The first, which Nunes described as "straight-up referrals," covered five individuals whose "crimes are lying to Congress, misleading Congress, leaking classified information." Two others related to "charges of conspiracy to lie to the Foreign Intelligence Surveillance Court" and the last to a "global leak referral." Nunes' letter did not identify who was was subject to the referral, nor did it identify the alleged crimes. The action on the referrals is now the hands of Barr, who is putting together a team to examine the FBI's initial investigation into President Trump's campaign in the summer of 2016. Asked about Nunes' criminal referrals during a congressional hearing this week, Barr testified, "Obviously, if there is a predicate for investigation, it will be conducted." During a Senate Appropriations hearing on Wednesday, Barr said, "Yes, I think spying did occur" and "I think spying on a political campaign is a big deal." When offered a chance to withdraw the remark, he instead clarified: "I want to make sure there was no unauthorized surveillance." Hours after the notification was sent to Barr, Nunes said the public may never know who is mentioned in criminal referrals. "We don't know if you'll ever know who' is named, Nunes told Fox News host Sean Hannity. "But I can tell you that if you follow the Russia hoax closely, if I gave you seven guesses, you'd probably get the five people that we have referred." Nunes and other top GOP investigators have placed great faith in Barr to make headway toward completing an investigation begun last year by a joint GOP-led task force comprising the Judiciary Committee and the Oversight Committee. Key to this effort, which has been bolstered by intelligence panel Republicans, was investigators looking over roughly 10 transcripts of interviews conducted by the task force last year. In recent weeks, House Judiciary Committee ranking member Doug Collins, R-Ga., released transcripts of the private interviews of former FBI agent Peter Strzok, former FBI lawyer Lisa Page, Justice Department official Bruce Ohr, his wife and former Fusion GPS contractor Nellie Ohr, former Trump campaign aide George Papadopoulos, former top FBI official Bill Priestap, and former FBI general counsel James Baker. A potential contributing factor to Nunes' efforts is the House Intelligence Committee's vote last fall to release the transcripts of more than 50 interviews it conducted in its now-completed Russia investigation, which had been submitted to the Office of the Director of National Intelligence for declassification review. President Trump urged then-Attorney General Jeff Sessions back in August 2018 to investigate any possible abuses by the DOJ and FBI, tweeting out: "Look into all of the corruption on the 'other side' including deleted Emails, Comey lies & leaks, Mueller conflicts, McCabe, Strzok, Page, Ohr, FISA abuse, Christopher Steele & his phony and corrupt Dossier, the Clinton Foundation, illegal surveillance of Trump Campaign, Russian collusion by Dems -- and so much more." Nunes has railed against what he says is collusion between the Democrats and the Russians.
Arthur Schwartz
@ArthurSchwartz
They're sweating bullets over in Obamaworld.
washingtonexaminer.com/news/devin-nun...
7:06 PM - Apr 11, 2019
Devin Nunes sends Trump-Russia investigation criminal refer...



Lessons learnt:

In this section, our findings are discussed during our project from both the perspectives of project content and usage of various techniques.

During stage I, we used Tensorflow and found most of the fake news are written with respect to politics, celebrities. This shows a direction for our further work on which we should paid more attention to and this may be a feature for classifying a news. In addition, the news content is not biased in sentiment analysis and thus the sentiment level of a news has no effect on the classification for a news. During stage II, we found that the popularity for most of the news will last for around 6 hours. This is of importance since it can be a signal for us to give the public a notification if the news has a high probability to be fake. During stage III, we learned how to build the whole big data pipeline deploy a distributed system with streaming process including how to achieve the distributed web crawler with dynamic tool of Selenium for tweet comments, how to write and read a whole block from Kafka topics, how to design the spark SQL processing with streaming dataframe to improve efficiency, how to deploy the distributed crawler for news content and the machine learning model in the streaming platform to write data to Kafka, how to write streaming data to MongoDB, how to design and implement web visualization including learn NodeJS and React.

Summary:

When social media has become the most cost-efficient way of communication among people, it is extremely intriguing to analyze people's reaction to a popular news post while eliminating the false information online. Therefore, we designed the Distributed News Monitor System that concentrates on the news content to alert the public about the fake

news and produce analysis of the public opinions from the Twitter comments on the news. Deep learning model is deployed and able to detect the integrity of the news according to its content and comments with 73% accuracy. Big data streaming analysis is expanded to reach real-time news monitoring and thus guide people to think deeper about the contents of the news. This system encompasses advanced modelling, real-time analytics, and scalability all in one.

Reference:

[1] Alexandre B.& Hernán A. M.(2019)10:7.Influence of fake news in Twitter during the 2016 US presidential election.*Natural Communications.*Retrieved from https://www.nature.com/articles/s41467-018-07761-2.pdf

[2] Adam C.S. (2018). *Detecting fake news at its source.* Retrieved from http://news.mit.edu/2018/mit-csail-machine-learning-system-detects-fake-news-from-source-1004

[3] David M.B., Andrew Y.N. ,&Michael I. J.(2003).Latent Dirichlet Allocation.*Journal of Machine Learning Research. 993-1022*.Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[4] Umar F.(2015). Characteristics of News are Accuracy, Balance, Concise, Clear & Current. Retrieved from http://www.studylecturenotes.com/journalism-mass-communication/characteristics-of-news-are-accuracy-balance-concise-clear-current

[5] Zi Z.Y., Di Y.Y., Chris D., Xiao D.H., Alex S., & Eduard H. (2016). *Hierarchical Attention Networks for Document Classification*