"Alternative facts and fake news are just other names for propaganda" ~ Johnny Corn[3]

Introduction

The rise of fake news is one of the downsides of the internet age and concern over the problem is global. Aggravating the problem, much information remains unknown regarding the vulnerabilities of individuals, institutions and society to manipulations by malicious actors. Resolving the problem using natural language processing is a hot topic of discussion among data scientists. This report discusses key findings and prediction results of claim truth ratings obtained from the 2019 'Leader's prize Competition' dataset.[1] Data is related to political aspect as observed from the claimants' names – who are mostly politicians.

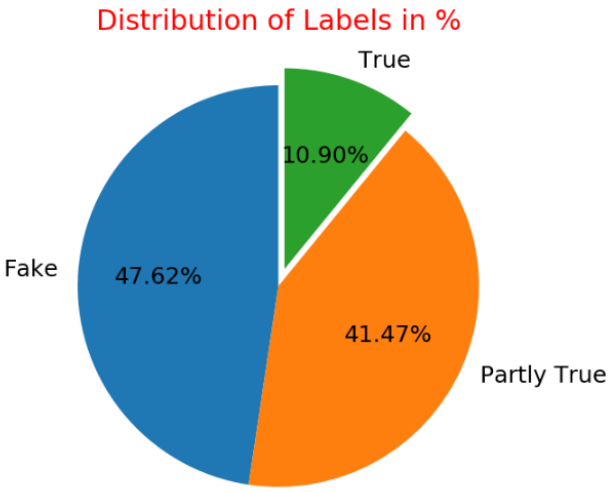

Figure 1 Claim labels and its distribution

The politics based dataset consists of 15,555 claims, its related articles and associated metadata (claimant, date and id).

The claims are classified as 'true', 'partly true' and 'false'. However, as seen in the figure below, the data is imbalanced with only **10.9 % of true claims**. This imbalanced data poses problems while training the data.

What could be the malicious feature in fake claims?

Well, on exploring the claimants, it's observed that approx. 5000 claims have no claimant, with most of the non-claimant claims belonging to fake claims. Further, fake claims have high frequency of its sources from Donald Trump, bloggers, viral images, websites, facebook posts, and other multiple sources from social media. The partly true and true claims are sourced from individuals, with no or very less social media claimants.

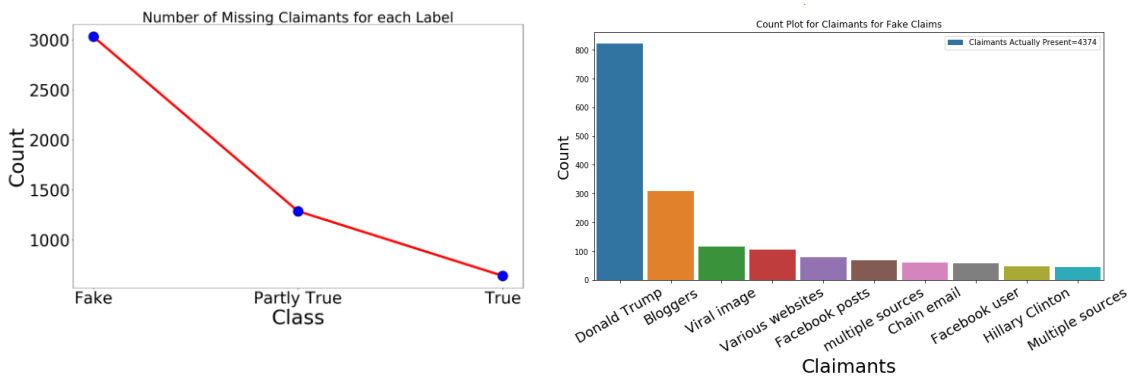

Figure3. Line plot for missing claimants for all classes (Left), Count Plot for Claimants for Fake (Right) Claims

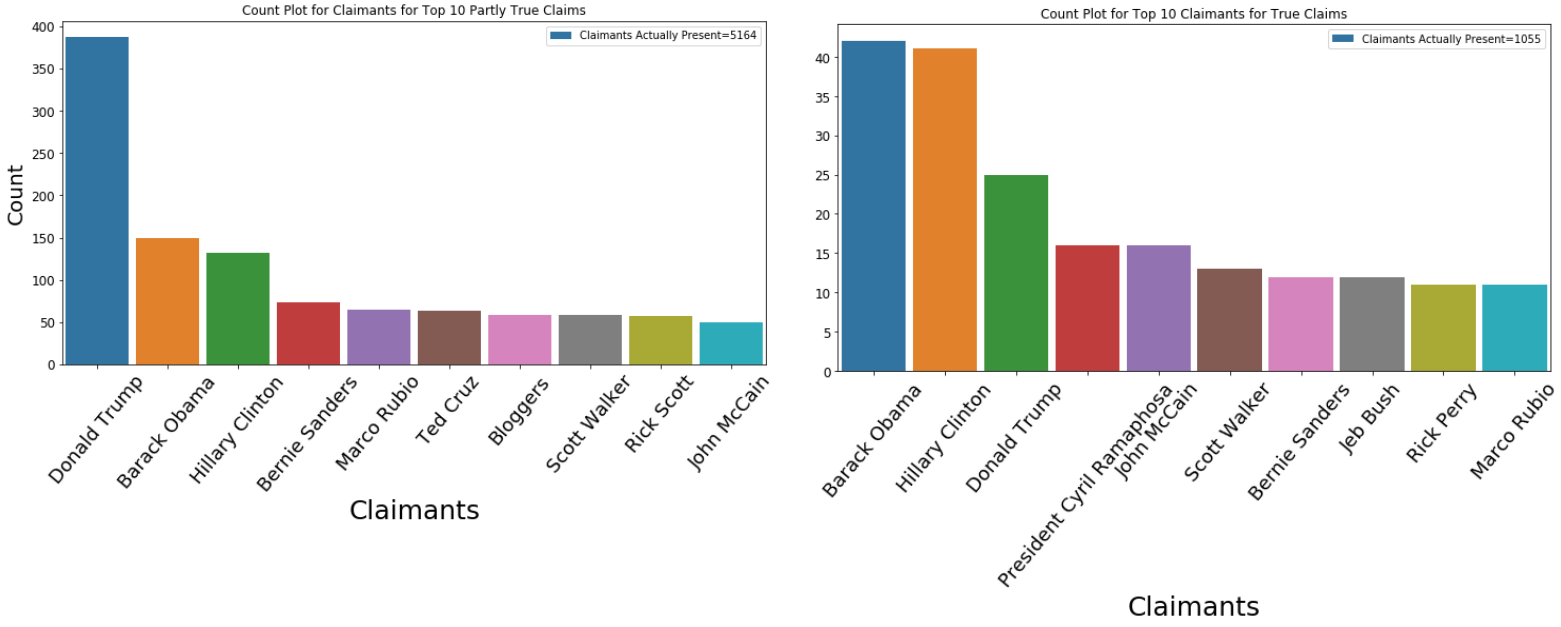

Figure3. Count Plot for Claimants for Partly True (Left) and True (Right) Claims

Observing the words in claim and articles for each class (like state, year, united, republican, Donald Trump, American, president, etc. it can be said that the data is related to political aspects in the United States since Donald Trump became the president. The words of claims have maximum similarity with the words of articles for the true cases, lesser for partly true and least similarity for false.

Claims and Articles Credibility Analysis with Textual Content:

These textual words can provide important signals for distinguishing the true articles from the false ones. Analysis of textual words is shown below for all three classes.[4]

Fake claims and Related Articles (Label=0): The cloud plot for fake claims and articles is dominated by words like will, people, example, claims, including, election etc. These words are not very strong words with positive polarity. These words sound like a politician for ex- Trump is giving a speech and promising people with false promises for their nations. Also word cloud for claims and related articles is quite different. Hence, these words can be related to the fake label for the claims.

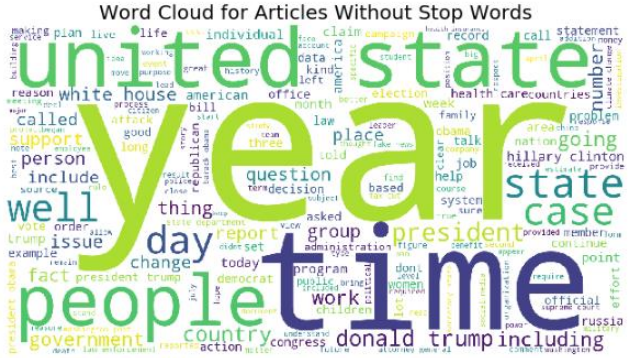

Figure 4. Word Cloud Plot for Fake Claims (Left) and Related Articles (Right) without Stop Words

Partly claims and Related Articles (Label=1): The word cloud plot for claims and articles for class 1 partially matches with some common words like year, people, etc.

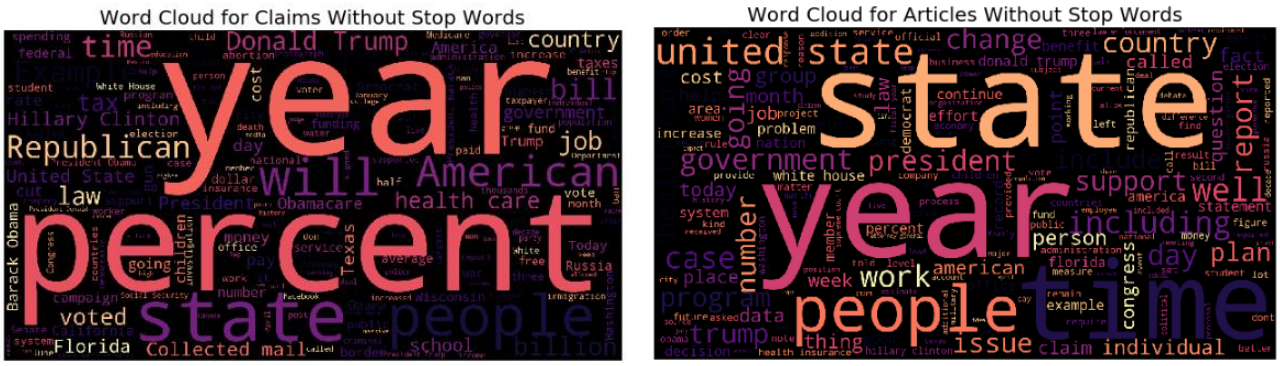

Figure 4. Word Cloud Plot for Partly True Claims (Left) and Related Articles (Right) without Stop Words

True claims and Related Articles (Label=2): Many words like year, state, time, people, Issue appear in both the word clouds. These words can be related to some facts stated by some personalities at certain time which justifies the True label for these claims.

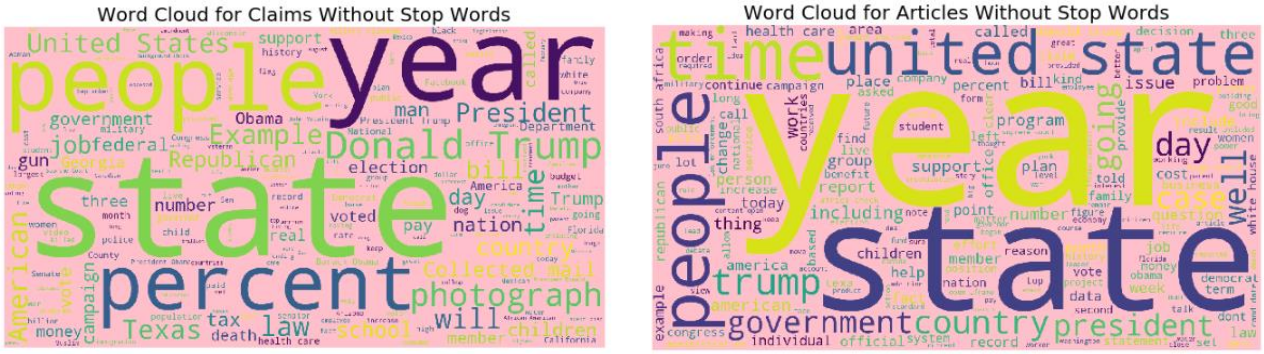

Figure 6. Word Cloud Plot for True Claims (Left) and Related Articles (Right) without Stop Words

Detection of fake news

With the development of distributed storage for holding large amounts of data and distributed computing for processing these vast amounts of data in less time, machine learning and natural language processing (NLP) techniques have become practical and critical in many industries.
We have leveraged algorithms from machine learning, natural language processing and neural networks to combat the challenge of fake claims and articles.
We have preprocessed claims and articles by noise removal, lemmatizing, tokenization and stop word removal. It's essential to convert the text into the features which algorithm can understand, following are the features that we extracted from the data:

1. Vectorizers – Bag of Words, Inverse Document Frequency (TF-IDF) with N-Grams, Hash
2. Word2Vec using cosine-similarity
3. Extracting sentences from articles with good cosine similarity with respective claim
4. Meta- Data Features

Modelling:

We have developed three classification models to predict the truth ratings (labels) that human fact checkers would assign to each claim based on some related articles and the metadata. The models include Passive Aggressive algorithm, LSTM and BERT. The comparison for models with different features on the basis of F1-Score and Train, Test Accuracy can be seen in the plot below:

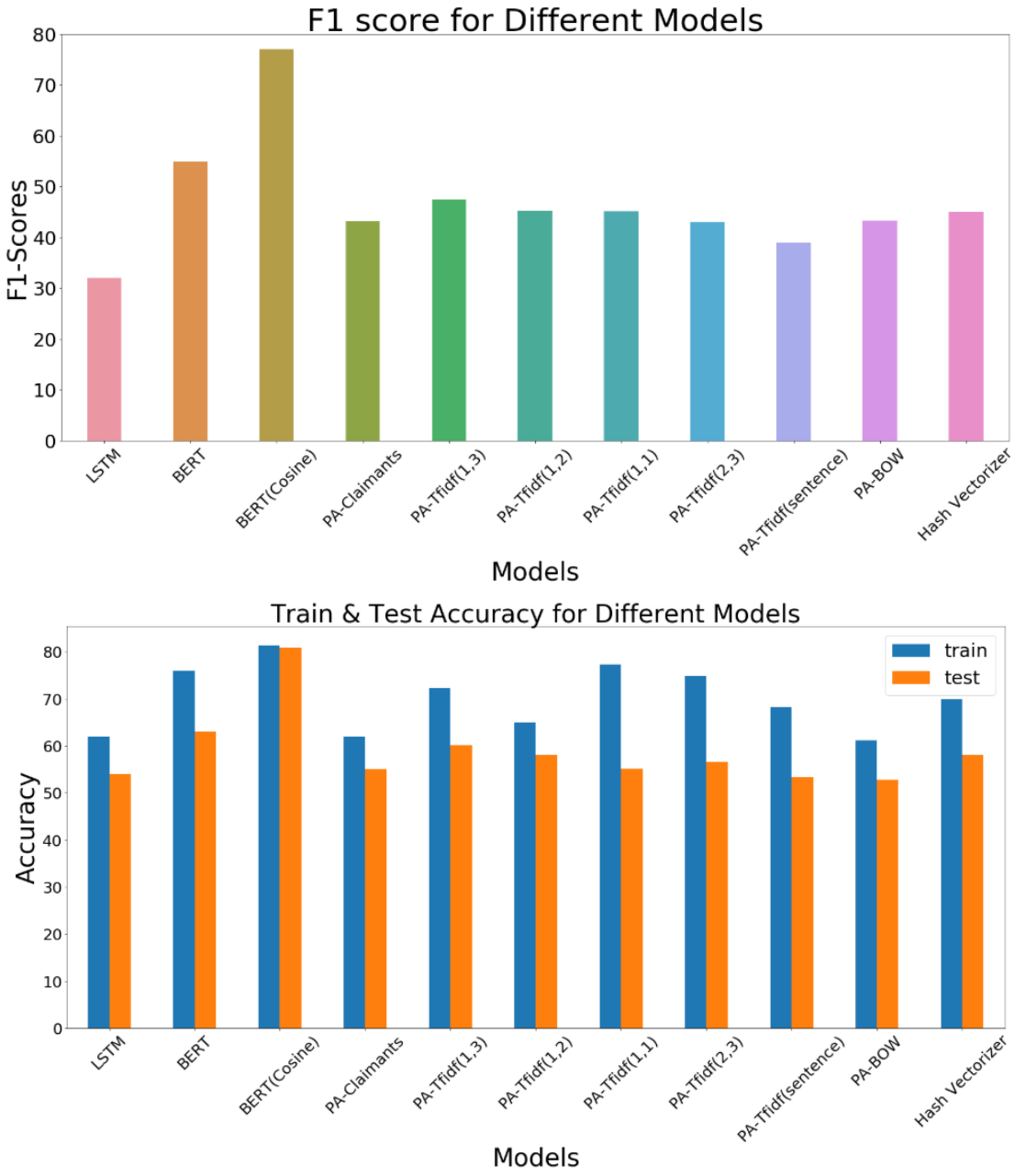

Figure 6. Distribution of F1-Score (Up) and Accuracy (Down) on Train and Test set for Different Models

The illustrations above depict the performance of models with BERT performing the best. Second figure depicts that models are not overfitting.

Discussion & Conclusion

The best model that can be efficiently used for the purpose for fake news detection emerged out to be the BERT model which gives a high **F1-score of 0.77 and accuracy of 81% on test set**. The more detailed analysis based upon bias-variance trade off and fitting is being worked on and will be reported in presentation in detail.

From the analysis of claimants it's certain that most of the fake news originate on social media platforms. The media is so central to our lives that we believe what we see onscreen is real. In fact, it's more real than reality: emotions are heightened, drama sharpened, issues simplified. Therefore, classification of claims and articles is an essential need.[3] The issue can be resolved with NLP using various features like word frequency, similarity factor, frequency of combination of words and classifying using various models in machine learning and neural nets.

Spread of misinformation during presidential elections on WhatsApp is one of the most famous examples of fake news. Of all the forwards, the information was high asymmetrical, favoring far-right winner.[2] As a solution to many more such instances, WhatsApp limited the number of forwards to groups and every forward message is labelled as 'forwarded'.

Similar to the Brazil elections, our claim and articles dataset seems highly biased towards Donald Trump. The first reason being the Donald Trump being the claimant of highest number of fake claims. Secondly, word clouds show dominance of 'Donald' and 'Trump' words in dataset. Leveraging effective models like passive aggressive classifier and BERT, spread of misleading texts can be limited and new or articles should last for shorter duration.

The snap of submission uploaded is given below. Final submission with high score is still to be made. We will be done before presentation.

47	In Search of Truth	0.352226
48	G12	0.350073

References:

1. Leaders Prize: Fact or Fake News? https://leadersprize.truenorthwaterloo.com/en/ [accessed on November 26, 2019)
2. WhatsApp fake news during Brazil election 'favoured Bolsonaro' in The Guardian https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests [accessed on November 28, 2019]
3. Goodreads : https://www.goodreads.com/quotes/tag/fake-news

4. FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network (https://arxiv.org/pdf/1805.08751.pdf) [accessed on November 28, 2019]

.