Pretrained Model to Annotate Unlabeled Twitter Data in PySpark

DS-5640: Big Data Scaling

Final Project Report

Ken Fahmidur Rahman

GitRepo: https://github.com/KenR22/twitter_spark_nlp

Introduction:

Twitter provides extensive social network data that includes not only social connections but also text data. It is a rich source of text data. However, social media tends to have a lot of false and unnecessary information, making the text data very noisy. In this project, Twitter posts from one of the most controversial accounts that belonged to Donald Trump will be analyzed and explored.

Problem:

Tweets by Donald Trump have caused a lot of controversies on social media. These tweets have been extracted and stored on the Kaggle database: All Trump's Twitter insults (2015-2021). The dataset contains the tweet's date, the actual tweet, the target of the tweet, and what type of insult was in the tweet.

The dataset did not contain any variable that seemed like a good target variable to make predictions. Prediction of tweet date or type of insult does not make sense. The dataset includes the target of the tweets, but can we extract the entities mentioned in the tweet?

One of the ways to extract entities and sentiment on data without labels is to use pre-trained models. The idea is that using a model that has already been trained to do a specific task, the model can be used on other datasets and achieve reasonable accuracy.

In this project, spark RDD was be used to use these pre-trained models on Trump's tweets. Several other exploratory data analyses were conducted to get an understanding of the data.

Methodology:

Software:

1. Data analysis was done by loading the dataset from CSV format to Pyspark RDD.
2. The pre-trained models were loaded using the sparknlp module.

Hardware: All the analysis was conducted on Google Colab. The data was directly downloaded using Kaggle API on Google Colab.

Results:

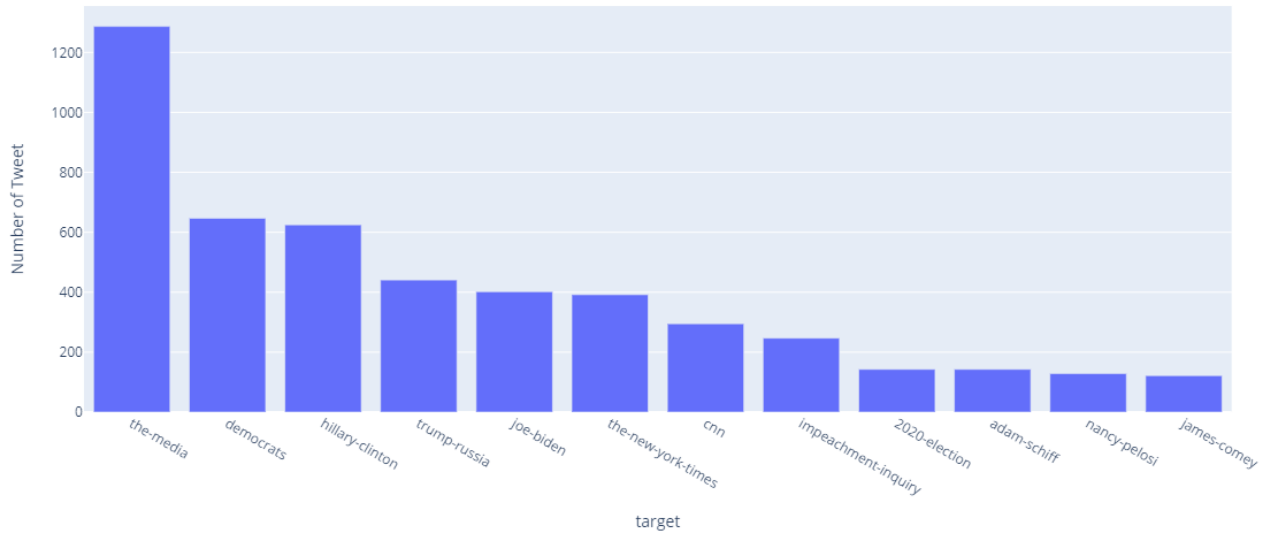
Loading the Data:

Loading the data was a simple process. First, the authentication JSON file was uploaded to google colab manually then moved to the "./Kaggle" directory. This gave google colab access to download the data from Kaggle directly. Then the data was loaded as pyspark RDD format from CSV. The infer schema method was not working on the date column. So, the date column was formatted to DateTime type manually.

Exploratory Data Analysis:

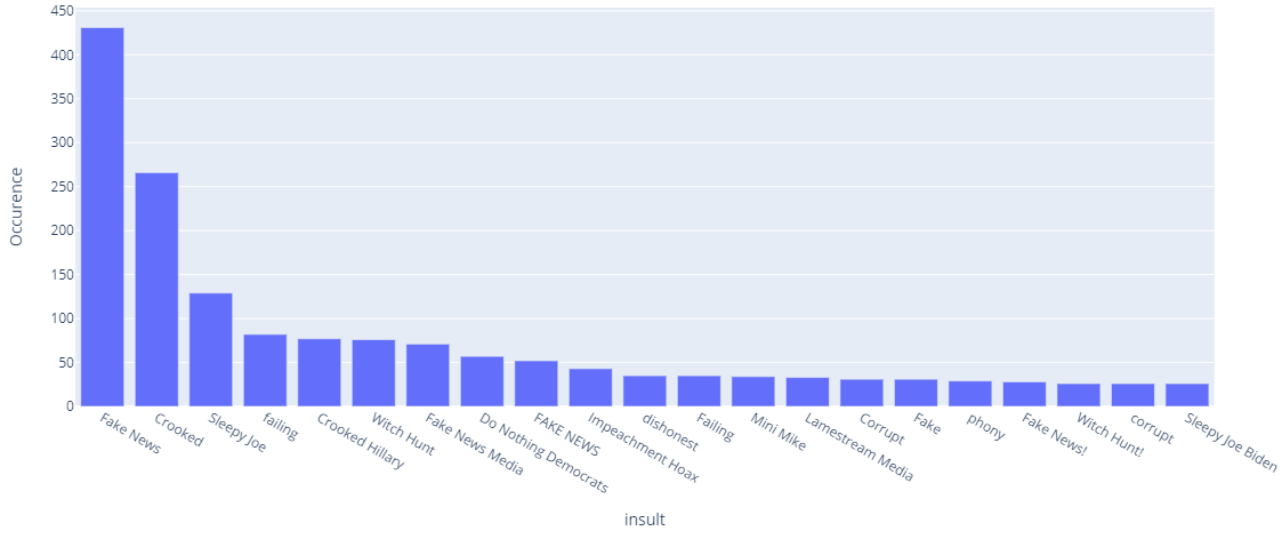
Some primary exploration was conducted on the data before feeding it into the pre-trained models.

1. Who are the targets of Trump's tweets?
   The target column indicated who are the targets of tweets by Trump. This column was grouped, and the number of instances was counted using the "groupby" method from pyspark dataframe.



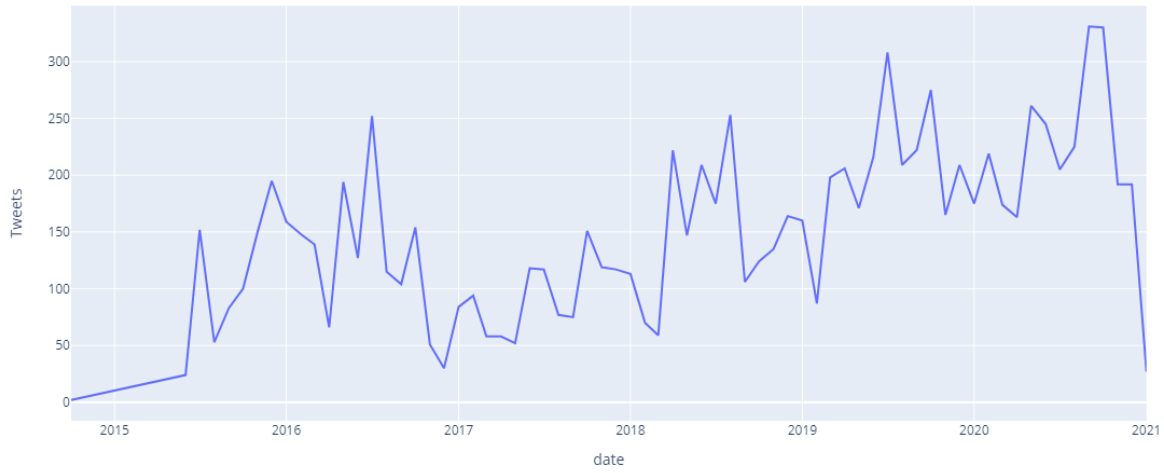
The media is the biggest target of his tweets. His political opponents like Hillary Clinton and Democrats also take up a major count of tweets. Political events like elections and impeachment are in the data too.

2. The data had another column that specified the types of insults that Trump threw at his target. The aggregated count looks like below:



The most counted insult is fake news. This goes with the fact that his biggest targets were the media

3. Finally, let's see how his tweet count changed over time:



The post count is aggregated by month and plotted. It does not show any specific trend over the year. But there are some sudden peaks that are related to specific political events like impeachment and elections. However, an increase of tweets can mean either there was a longer tweet, and they were broken down, or there were more tweets.

Pretrained models:

This dataset did not contain any specific target column. Predicting tweet date does not make much sense. Insult type and target sometimes did not have any information from the text. So, as an alternative to get more information out of the text data, pre-trained models were deployed using the sparkNLP module.

SparkNLP is a pyspark based NLP module. It is developed by John Snow lab and contains most of the functions from the state of the art NLP modules like spaCy and nltk. The main advantage of SparkNLP is that it can be directly applied to a Spark dataframe. SparkNLP also contains different pre-trained model weights like glove, word2vec, and BERT. Note that the training of these models might take weeks in a good machine. So, being able to use these weights without training is a huge advantage.

1. Sentiment Analysis: To test the sparkNLP pipeline first a pre-trained model for sentiment analysis was loaded. The loaded model is titled "analyze_sentimentdl_use_twitter" for the English language. The mode pipeline accepts a text column named "text" only. So, the column tweet was renamed to text. A sample of the output from the pipeline is shown below:

date	target	insult	text	document	sentence_embeddings	sentiment
2014-10-09	thomas-frieden	fool	"Can you believe ...	[[document, 0, 14...	[[sentence_embedd...	[[category, 0, 14...
2014-10-09	thomas-frieden	DOPE	"Can you believe ...	[[document, 0, 14...	[[sentence_embedd...	[[category, 0, 14...
2015-06-16	politicians	all talk and no a...	Big time in U.S. ...	[[document, 0, 12...	[[sentence_embedd...	[[category, 0, 12...
2015-06-24	ben-cardin	It's politicians ...	Politician @Senat...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-24	neil-young	total hypocrite	For the nonbeliev...	[[document, 0, 12...	[[sentence_embedd...	[[category, 0, 12...
2015-06-24	rockin-in-the-fre...	didn't love it	.@Neilyoung's son...	[[document, 0, 12...	[[sentence_embedd...	[[category, 0, 12...
2015-06-25	willie-geist	uncomfortable loo...	Uncomfortable loo...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	jeb-bush	will NEVER Make A...	Just out, the new...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	molly-sims	a disaster	The ratings for T...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	nicole-wallace	a disaster	The ratings for T...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	the-view	dead T.V.	The ratings for T...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	the-view	put it to sleep	The ratings for T...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	nicole-wallace	doesn't have a clue	.@WhoopiGoldberg ...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	the-view	close to death	.@WhoopiGoldberg ...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	lawrence-o-donnell	dopey political p...	I hear that dopey...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...
2015-06-25	lawrence-o-donnell	one of the dumber...	I hear that dopey...	[[document, 0, 13...	[[sentence_embedd...	[[category, 0, 13...

Three new columns are added. Document and sentence embedding contain latent space information about the text itself. These values can be passed through another classifier to train different problems, i.e., whether a tweet is insulted.

Our primary interest is in the sentiment column, which contains the probability of a specific tweet is positive, negative, or neutral. Ideally, all the tweets should be negative as they are all insults. Let's see how many tweets are positive and how many are negative:

output	Number of Tweet
[negative]	8472
[positive]	1476
[neutral]	412
[]	132

Most of the tweets are negative, but many tweets are predicted as positive, surprisingly. A sample of those positive tweets are shown below:

positive	Yet more evidence of a media-rigged election: https://t.co/rVh4ocgx3r
positive	We did it! Thank you to all of my great supporters, we just officially won the election (despite all of the distorted and inaccurate media).

Some positive words may be affecting the sentiments. For example, the second tweet contains the word sentiment and an indirect dig towards media. The gratitude part played more roles in the prediction. As a remedy, there could be a threshold of prediction of certainty. If a prediction is uncertain then that needs to be manually labeled.

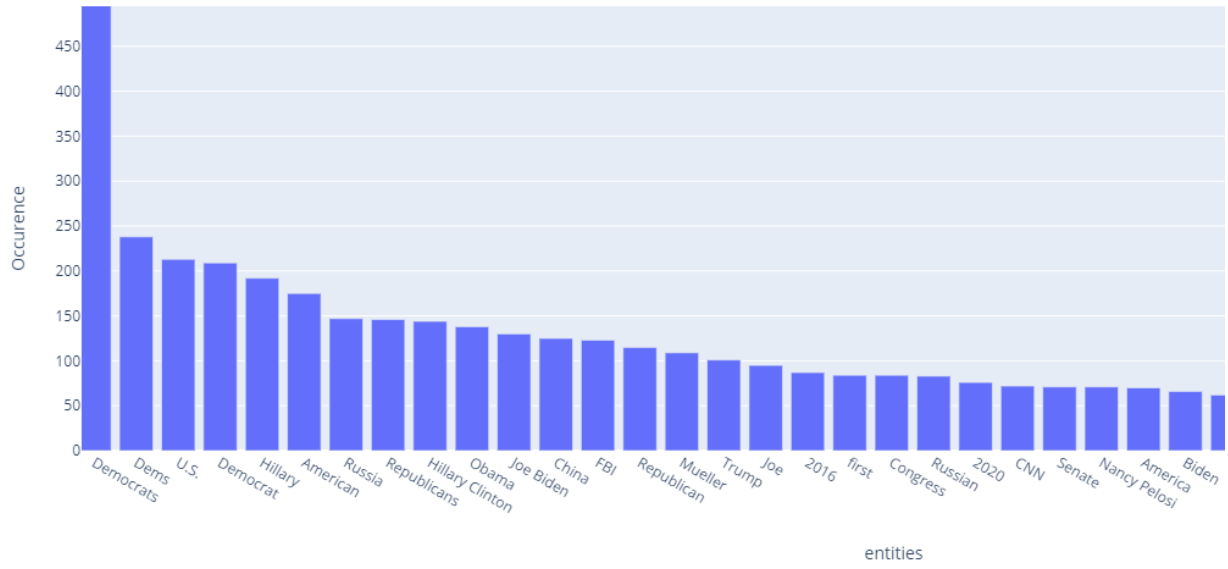
2. Named Entity Recognition: It does not make much sense to predict whether a tweet was positive or not in data that contains insult only. The second pre-trained model was deployed to extracted all the named entities from the text. This will help us understand what Trump was talking about in the tweets other than the targets of the insults.
   The dataframe from the pre-trained BERT model to find named entities are shown below:

entities	text
[]	Wow, sleepy eyes @chucktodd is at it again. He is do totally biased. The things I am saying are correct. - far better vision than the others
[]	"""In politics
[Petraeus]	The system is rigged. General Petraeus got in trouble for far less. Very very unfair! As usual, bad judgment.
[Hillary Clinton]	Crooked Hillary Clinton was not at all loyal to the person in her rigged system that pushed her over the top, DWS. Too bad Bernie flamed out
[Hillary, African-American]	"How quickly people forget that Crooked Hillary called African-American youth ""SUPER PREDATORS"" - Has she apologized?"
[F-35, Billions of dollars, January 20th.]	The F-35 program and cost is out of control. Billions of dollars can and will be saved on military (and other) purchases after January 20th.

tokens
[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O]
[O, O]
[O, O, O, O, O, B-PERSON, O, O, O, O, O, O, O, O, O, O, O, O]
[O, B-PERSON, I-PERSON, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O]
[O, O, O, O, O, O, B-PERSON, O, B-NORP, O, O, O, O, O, O, O, O]
[O, B-ORG, O, O, O, O, O, O, O, B-MONEY, I-MONEY, I-MONEY, O, O, O, O, O, O, O, O, O, O, B-DATE, I-DATE]

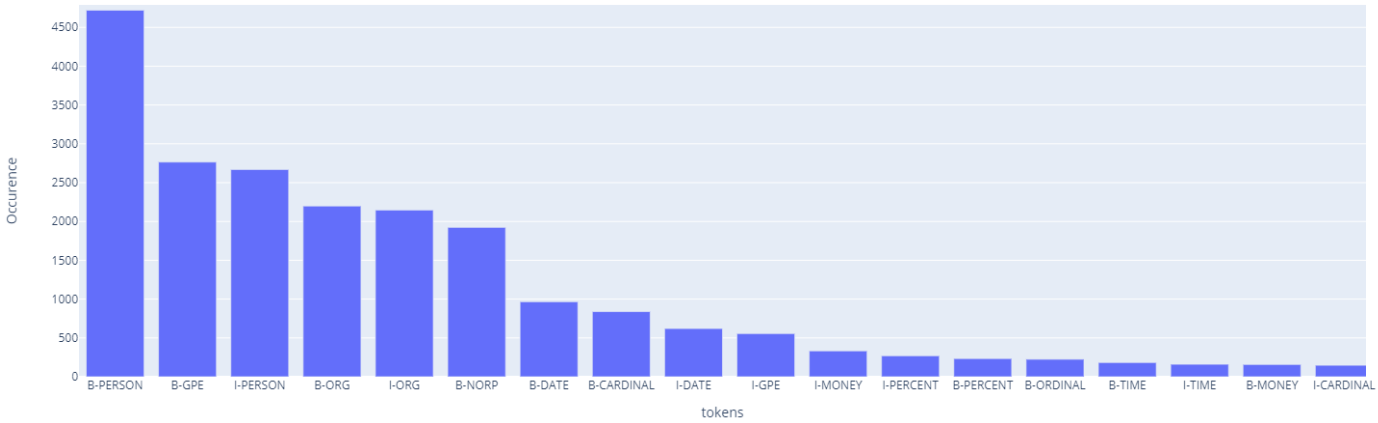
The entities column gives the name of the entities, and the tokens column shows the types of entities occurring in one tweet.

Let's see what the entities most occurring in his tweets are:



Unlike the columns of the target the most occurring entity is democrats. This might be because he mentioned democrats when tweeting against Hilary Clinton and Joe Biden. Interestingly 2016 and 2020 came up as years. This makes sense because these two years were the ones when Trump was involved in the presidential elections.

How about the types of entities most occurring in his tweets?



The plot above shows his insults mentioned more persons than organizations. B-Person represents a person entity. B-GPE is a geopolitical entity. This may be due to the fact that he mentions the U.S. a lot in his tweets.

Conclusion:

In this project, unlabeled text data was explored using pyspark dataset. Since there was no clear target column for prediction, pre-trained models were deployed for further analysis of the text data. It seems news media are the most targeted entities from the target data. However, from named entity recognition, it was found that he mentioned people the most.