**SPRINGBOARD - Capstone project I**

**General**

After some consideration because of my new experience with Data Science, I have decided to focus on projects where the NLP algorithms and techniques are used. When I decided to choose for projects related to NLP, I was merely interested in the semantics of a language in general and what the effects might be in specific situations regarding exploratory outcomes from a business perspective. What I mean by that is, for example, what can be measured when using specific text in tweets and what can I do with the data once we have tokenized text into analyses after using methods and algorithms. That is a very general statement but for someone who is new to Data Science and statistics, I thought that would be a good start.

With NLP we can do a lot nowadays, I have seen different scenario’s where recommendations start to play a big role and which can be very beneficial to users of a certain platform. Netflix, Amazon and bigger companies starting to use this technique for recommendations for some time already. All in all, what I have seen in the Data Science world that mostly the use of this ‘science’ is to optimize and make life better, be it in finance, health, movie industry. Generally speaking, I think that is what makes Data Science so beautiful and useful.

**Project General**

For some time, I have seen and witnessed a plethora of words used on Twitter by the current President Mr. Trump, and often have I thought if his way of communication would be effective to the world. America should be great again, and I was willing to find out if his way of use of Twitter was the right way to make America Great Again. With that said, the project is subjective and therefore my comments and findings might be subjective as well.

In this Capstone I would like to use NLP techniques to find out the sentiment of president Trump’s tweets ( @realDonaldTrump) and see *if the used words are in general positive or negative and what the effect can be in regards to retweets and in regards to actualities in the news*. *Through the use of words can we actually predict what the effect is on retweets? Do people like to retweet his tweets when using his own typical vocabulary? But also, would it be interesting if we can reflect the tweets onto the news regarding stock prices or other impacts in the news, and see what the effect of his words might be.* In other words, what is the influence of this ‘negative’ or ‘positive’ tweet of the president. The results may be answer a few of the above questions but I am merely interested **if we can predict retweets based on his own words**.

Moreover, after analyzing the data and using NLP techniques, I should be able to find tags in tweets which inherently are related to sentiments in general. By using the pre-trained sentiment classifier Flair, I might be able to take sequences of characters and words into account and relate to a positive or negative outcome of words. I have chosen for the ‘Flair’ classifier since it can recognize negated tweets but also intensifiers like the president is using many times (eg. “While the Do Nothing Democrats FAIL the American People, and continue the Impeachment Scam, my Administration will continue to deliver REAL RESULTS, as seen over the past month, below!”).  The model which I am going to train will be related to a prediction for retweets and the correlation between the retweets and the effects in occurrences in the news.

**The Dataset**

For this prediction I have found a dataset from [‘The Dataverse Project’](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FKJEBIL), an open source research data repository from The Institute for Quantitative Social Science (IQSS) which collaborates with the Harvard University Library and Harvard University Information Technology organization to make the installation of the Harvard Dataverse openly available to researchers and data collectors worldwide from all disciplines. Besides, Twitter has an open-source policy about their tweets and by which interested can use Twitter’s API’s and also find good documentation on the attributes used in a tweet.

Looking at the dataset[[1]](#footnote-2), we can see typical attributes which makes a tweet unique[[2]](#footnote-3):

|  |  |
| --- | --- |
| Attribute | Semantics |
| id | Id of the tweet |
| created\_at | Time the tweet was created |
| coordinates | Geographic location of the tweet reported by the user |
| user | The user who has posted the tweet |
| text | Actual UTF-8 text which might be truncated and be available in full text under |
| retweet\_count | Number of times the tweet (with this id) has been retweeted |
| retweeted | Indicates whether the tweet has been retweeted by the authenticating user |
| retweeted\_status | Indicates whether the tweet has been retweeted but contains a representation of the *original* Tweet that was retweeted |
| favorite\_count | Indicates approximately how many times the tweet has been liked |
| truncated | Indicates whether the value of the text attribute was truncated, because of the limitation of 140 characters. |

Although the dataset is very large to make predictions out of it, the current attributes have been slightly different. The field ‘geo’ for example has been deprecated and can now be found in the field ‘coordinates’. For my model I will mainly only use the attributes ‘text’ and ‘retweet\_count’ because the ‘text’ field can give us the data for analyzing the text and turn this into a matrix used for the prediction of the retweets and the ‘retweet\_count’ might be of later use for tweaking the model. Roughly said, for the prediction of the retweets I will transform ‘text’ into a matrix so it can be used for my model to predict the retweeted count.

## **Dataset Metrics**

Total size of data uncompressed:115901693 bytes  
Number of objects (submissions): 40,241  
Start Date: Mon May 04 18:54:25 +0000 2009  
End Date: Thu Jul 11 15:52:19 +0000 2019  
Format: ndjson (new line delimited JSON)

## **Overview**

This dataset contains all known publicly available tweets for Donald J. Trump's (@realdonaldtrump) Twitter account. The shape of the dataset consists of 40241 rows and 34 columns. Because the dataset consists of all the tweets wherein ‘@realdonaldtrump’ has been mentioned

## **Methodology**

This data was compiled from multiple sources including several online Github accounts that contained the status ids for previous tweets made by Donald Trump. All ‘ids’ were compiled into a single list and then those ids were requested from Twitter's "statuses lookup" endpoint. This dataset will also include the tweet information for any retweeted tweets under the "retweeted\_status" key for each JSON object. The user object has been left in each tweet (both the main tweet and retweeted / quoted tweets if they exist).

**Data Wrangling Process**

Generally speaking, I would have the data in 2 formats:

* Text of the tweets formed into a Corpus
* DTM-matrix for tracking the term frequency for each term by each document (a tweet)

**But first…**

After importing the data, I have thought about the results which I wanted to gain. Ultimately it would be great if I could place the data into chunks and standard format so it would make more sense of how the words are related to each other and for further analyses, if I could make any relation yet between the data-sentiment and retweets.

Along the way I have looked into the created DataFrame and searched for unusual text and null values. Generally, there were 450 tweets which did not get retweeted, but overall there was an average of 4956 times that a tweet of the President was retweeted!! The given dataset was very clean and there were no abnormalities in the text column nor in the retweeted column.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The dataset was substantially loaded with time and date stamps and I took the advantage of that by converting this into a histogram to see how the twitter behavior of Mr. Trump developed over the years.

Since NLP is not yet in our curriculum, I looked in documents, the internet and books about how to work with text. There is an amount of libraries which can support my project but there is one which I should start with and that is the CountVectorizer from sci-kitlearn for feature-extraction. To eventually use this method I should clean the data to transform the DataFrame into a Vectorizer with numerical values which can then be absorbed by the used prediction training model.

The cleaning of the data has been done in 3 steps:

* Make text in lowercase, remove text in square brackets, remove punctuation and remove words containing numbers
* Take out punctuation and non-sensical text
* Take out the ‘http’ text which would not add value to the CountVectorizer

As Machine Learning models take numerical values as input, the DTM-matrix comes in very handy and should therefore be very carefully analyzed. Good data makes good models.

During the cleaning of the data, I have noticed some foreign-text/signs which could not be interpreted by the CountVectorizer.

A picture containing screenshot

Description automatically generated

I have decided to start with the dataset again and make a loop to extract the correct tweets of @realDonaldTrump

1. Baumgartner, Jason, 2019, "Twitter Tweets for Donald J. Trump (@realdonaldtrump)", <https://doi.org/10.7910/DVN/KJEBIL>, Harvard Dataverse, V1 [↑](#footnote-ref-2)
2. See also <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object> [↑](#footnote-ref-3)