Retweet Prediction

David McMonagle

June 2021

1. Introduction

The ongoing pandemic caused by the Coronavirus Disease (COVID-19) has massively impacted almost every community and society across the globe, and therefore it has changed billions of peoples behaviour. These changes can be seen in everyday life affecting habits and social interaction in and around the workplace as well as at home. Though, more importantly for this research, the pandemic has massively impact the online world, this can be seen in the change of reading habits on Wikipedia and Reddit [1,2]. Naturally, a world pandemic is massively discussed on social media, with Facebook, Twitter and YouTube being the platforms of choice.

This research has the goal of understanding how information spreads on one specific social media platform of choice, in our case that is Twitter. Twitter is one of the biggest social media platforms in the world with 192 million active daily users [3]. Users follow and share with each other information, pictures and videos though this study will be specifically focusing on text tweets or more importantly on retweets which is when a user reposts or forwards a message posted by another user [4].

Retweets are a fast and rapid way of spreading information as well as disinformation, very famously Donald Trump showed the true power of twitter by mastering it in his drive to be president [5]. As such, understanding retweet behaviour is useful and has many practical applications, e.g. political audience design [6,7], fake news spreading and tracking [8,9], health promotion [10], mass emergency management [11], etc.

These are the steps that have been taken to try and attempt to solve this problem:

- Loading Data
- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering
- Modelling & Evaluation

2. Related Work

This problem is one that has been tackled by many researchers before. A very large variety of techniques have been used, though it is hard to identify the most common. The choice of model very much depends on the data that an individual has and how he would like to use it, though it has to be noted that. Joukov Costa de Oliveira uses a lot of different supervised learning models such as Decision Trees, Logistic Regression and SVM [12]. There are many more papers on this field of work, with some individuals using only numerical metrics and others TF-IDF methods as well as Word2Vec to analyse how the language used impacts the retweetability of a tweet.

3. Methods

3.1 Supervised vs Unsupervised

As we know, Machine Learning can be categorised into two areas, Supervised Learning and Unsupervised Learning. Firstly, Supervised Learning is used when the target of said information is known, meaning it can be easy to predict the data for these users based on the information compiled about them previously. Unsupervised Learning however, is applied when these specific targets are indeed not known and we would like to pursue the collection of this data in order to compile said datasets.

Given that the whole idea is to build a model which can predict the popularity of a tweet prior to its composition using said datasets, the use of Supervised Learning would be most effective in this area. In this field of study particularly, Supervised Learning is the most widely used approach for this exact reason. Alongside this, Supervised Learning also contains two key elements, input values and output values. These both permit the algorithm to take this information and build a model that can predict a result.

Furthermore, the algorithms themselves can also be categorized into two elements, classification and regression. Classification, the clue being in the name, refers directly to a specific category and regression is a continual value. In this instance we are tackling a regression problem, that is why Linear Regression, Decision Tree regression and Gradient Boost Tree Regression are the three models that have been chosen for this study.

3.2 Data Loading

This dataset was very challenging to work with as it has over 8 million data point. Therefore it meant the data was hard to work with and took up huge computational power. To load the data a spark session was created and then the data was loaded from three different files, containing column names, variable data and the label or in other words the amount of retweets.

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("Retweet").getOrCreate()
df= spark.read.csv("../Downloads/Assignment-2/train.data", sep= '\t')
label= spark.read.csv("../Downloads/Assignment-2/train.solution", sep= '\t')
cols = spark.read.csv('./Assignment-2/feature.name', sep= '\t', header= True)
df = df.toDF(*cols.columns)
label = label.withColumnRenamed("_c0","label")
```

3.3 Data Cleaning

First of all it was noticed that all of the variables were classed as strings therefore the datatypes where changed accordingly using the function seen below:

```
#Changing data type
def convertColumn(df, names, newType):
  for name in names:
     df = df.withColumn(name, df[name].cast(newType))
  return df
```

```
columns = ['#Followers', '#Friends', '#Favorites']
df = convertColumn(df, columns, 'double')
label = convertColumn(label, ['label'], 'double')
```

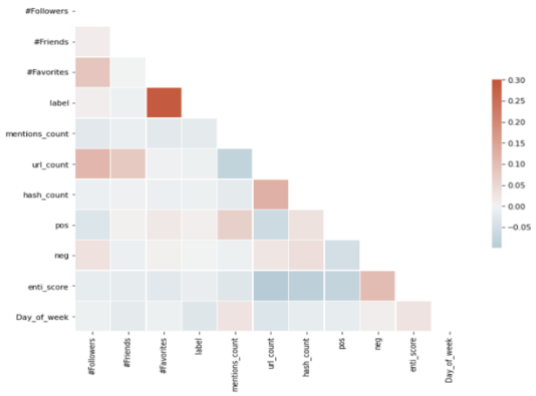
The next step in the process was to combine the data together coherently using the 'monotonically_increasing_id' function as well as '.join'.

```
from pyspark.sql.functions import monotonically_increasing_id

#Indexing both sets of data
df = df.withColumn("id", monotonically_increasing_id())
label = label.withColumn("id", monotonically_increasing_id())

#Joining data
df = df.join(label, "id", "inner").drop("id")
```

3.4 Exploratory Data Analysis

While conducting the EDA, it can be noticed that a lot of the data is skewed and there is very little correlation between the dependent variable and independent variables. This can be seen in the pairplots and correlation matrices within the notebook.

Furthermore from the summary statistics gathered it can be noticed that the variables are very much skewed to the left. Creating the correlation matrix and pairplots was very difficult as Pyspark kept presenting error after error, so a small sample of the data (0.2%) was converted into a pandas dataframe so that these plots could actually be made.



3.5 Feature Engineering

- **Hashtags Mentions and URLs**
  - Count of each of these was calculated using the formula below:

```
#Counting mentions and changing data type
df = df.withColumn("mentions_count",F.when(F.col("Mentions")!='null;',F.size(F.split("Mentions"," "))).otherwise(0))
df = df.drop("Mentions")
df = df.withColumn("mentions_count", df["mentions_count"].cast(IntegerType()))
```

- **Timestamp**
  - Day of the week was extracted using a reformatting function and a UDF
- **Sentiment**
  - Split between positive and negative
- **Entity Score**
  - Defined function to split entities with UDF to create score
- **Standard Scaler**
  - Features were scaled using standard scaler
- **Final Features**

4. Experimentation

- At first unscaled data was used but the results were insignificant
- Used the scaled data and in some cases the results were worse than with the unscaled data
- Changed the elastic net and removed it in some cases but insignificantly changed the results
- Attempted to use cross validator but kept getting error messages
- Attempted to make it into a classification problem with SVM and Random Forrest but frustratingly had the same problems as with the cross validator.
- Removed the null values since they had been kept with the hope of improving the results, it improved the results a lot but with very insignificant values
- Increased the size of the sample which had very little impact on the output
- Removed all labels equal to 0, this was because the regressions were constantly assigning retweets to non-retweeted tweets. The results were even worse.
- Attempted using Word2Vec unsuccessfully, getting many errors.

5. Results

5.1 Metrics for evaluation

- RMSE (Root Mean Square Error): standard deviation of residuals

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- R2 (R squared): proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

5.2 Linear Regression with Null values

- Performed terribly as the R2 value is of **0.007**, though on the test data it performed a tiny bit better with a value of **0.016**
- RMSE: **538.233**

5.3 Linear Regression without Null values

- Performed better than the previous linear regression, though it's R2 value was of **0.01** and **0.023** in the test phase.
- RMSE: **637.009**

5.4 Decision Tree Regression with Null values

- Performed terribly as the R2 value is of **0.00085**
- RMSE: **542.413**

5.5 Decision Tree Regression without Null values

- Performed better than previous one but still terribly as the R2 value is of **0.0025**
- RMSE: **643.658**

5.6 Gradient Boost Tree Regression without Null values

- Performed terribly as the R2 value is of **-0.0098**
- RMSE: **545.323**

5.7 Gradient Boost Tree Regression with Null values

- Performed better than the previous one but still terribly as the R2 value is of **0.0024**
- RMSE: **643.689**

6. Discussion

This research was very hard to conduct as pyspark kept crashing and/or presenting errors. Though if this research was conducted further with better equipment and pyspark knowledge, the results would be a lot better.

It can be noted that the results would have probably been better if Word2Vec or TF-IDF had successfully been implemented, since one could think that the word choice of a tweet probably vastly influences how many times a tweet will get retweeted.

Furthermore, if this problem was tackled as a classification problem the results obtained would hopefully be different. Though to attack this issue with purely classification issues probably means a change in the hypothesis and problem statement.

7. Conclusion

All three of the regressions used were unsuccessful at predicting the rate of retweets. When looking at the correlation coefficients it can be seen that favourite has a larger influence on retweets than any other value but it is questionable how much of an influence it really has since the regression were not good at predicting retweets.

Unfortunately this work is not going to be able to be applied in the real world, but with more pyspark knowledge and a deeper dive into the feature extraction this work could definitely produce better results.

8. References

[1] Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A. and Perra, N., 2020. Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia. *arXiv preprint arXiv:2006.06446*.

[2] Ribeiro, M.H., Gligorić, K., Peyrard, M., Lemmerich, F., Strohmaier, M. and West, R., 2020. Sudden Attention Shifts on Wikipedia Following COVID-19 Mobility Restrictions. *arXiv preprint arXiv:2005.08505*.

[3] Ying Lin, 2021, January. 10 twitter statistics every marketer should know. https://www.oberlo.com/blog/twitter-statistics

[4] https://www.merriam-webster.com/dictionary/retweet

[5] Micheal Barbaro, 2015, October Pithy, Mean and Powerful: How Donald Trump Mastered Twitter for 2016. https://www.nytimes.com/2015/10/06/us/politics/donald-trump-twitter-use-campaign-2016.html

[6] Stieglitz, S. and Dang-Xuan, L., 2012, January. Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences* (pp. 3500-3509). IEEE.

[7] Kim, E., Sung, Y. and Kang, H., 2014. Brand followers' retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth. *Computers in Human Behavior*, *37*, pp.18-25.

[8] Lumezanu, C., Feamster, N. and Klein, H., 2012, May. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[9] Vosoughi, S., Roy, D. and Aral, S., 2018. The spread of true and false news online. *Science*, *359*(6380), pp.1146-1151.

[10] Chung, J.E., 2017. Retweeting in health promotion: Analysis of tweets about Breast Cancer Awareness Month. *Computers in Human Behavior*, *74*, pp.112-119.

[11] Kogan, M., Palen, L. and Anderson, K.M., 2015, February. Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 981-993).

[12] Nelson Joukov Costa de Oliveira, 2018, January. RETWEET PREDICTIVE MODEL IN TWITTER.