How many retweets will you get?

Louis Fenouil
louis.fenouil@epfl.ch
Paul Margain
paul.margain@epfl.ch
Hugo Thiallier
hugo.thiallier@epfl.ch

Abstract

Retweets are, with the number of follower, the most famous indicator of popularity of tweets and more globally users. An interesting analysis that could lead to the development of usable model, is to train a machine learning algorithm in order to get a prediction of tweets popularity (and more precisely the retweet count of a given tweet). This report tries to address a confident model to perform such prediction, only by using a pre-defined dataset that has not been chosen in a way that it could surely provides could results.

1 Introduction

Popularity on social networks is less an objective than a life goal for numerous users, and Twitter is no exception as the race for retweets and followers is constant. If popularity surely goes beyond simply these retweet and follower counts, they are great indicators of scope and interactions of a given tweet. More, they are particularly full of sense for companies looking for a bit of advertisement since they could use 'famous' users to broadcast their ads and touch a wider market. For such companies the goal is thus to be sure that a given user will be a good broadcaster and that he will maintain a sufficient audience throughout his future tweets. Nobody can predict the future but it could be valuable to get an hint on what will be the mean retweet count (major indicator of interaction and tweets popularity) for a given user, in order to be sure to choose the most fitted one.

In this project we aimed at performing retweet count prediction for a given user, depending on his follower count, his friends count, the presence of an hashtag in his tweets or of an URL. The main challenge will be to perform such prediction by using only the data set provide for previous milestones, to see if it is possible to enhance the overall analysis that they made.

2 Related work

Works on Twitter data are not a lacking thing in data analysis and tweets popularity prediction is no exception. A good example of such predictive attempt mostly use the same features as the ones at our disposal but goes further by taking into account information related to the content of a tweet such as its length and its overall 'mood' (by sentiment scores). This latter analysis gave some strong retweet predictions both with and without tweet content related data. Most papers highlight the influence of follower count on retweet count, therefore it is consistent to think that our data set could enables to have a interesting predictions despite is relative size.

3 Dataset description

Datasets used are the same provided and analyzed by the paper mentioned before (Liang, Hai and Fu, King-wa., 2015) and more precisely the two main datasets taken into account are the following:

EgoAlterProfiles.txt is mainly about the number of friends (users who mention each others usually) and followers for a given twitter account of the dataset.

EgoTimelines.txt A list of every tweet published by the twitter accounts of the dataset, with hashtag, date, number of RT, mentions, urls.

EgoTimelines data set will be the most used one as it provides tweets information, and so just by combining them with follower and friends information from the other data set we can perform prediction more easily.

4 Methods

4.1 Data pre-processing

The first thing we did was to keep only the tweets made by an ego being in the EgoAlterProfiles data set since they are the only users that have known follower and friend counts. Moreover we had to remove each lines corresponding to a reply, a mention or a retweet in order to only have the tweets apart from interactions.

An important process to do was to select only the tweets published in 2014 since we only got access to retweet count for this year, even if we had to do the assumptions that this count was constant across ego for this year.

To have a first glance at weither or not the data set could possibly provides good prediction we plotted the retweet count function of follower count and friend count because they are supposed to be impacting parameters, hence we expected to have more or less an increase of retweet with more follower or friends but it turned to be false. Therefore we also plotted influence of an hashtag and a URL on this retweet count. If an hashtag increase the retweet count, it is not really true for the URL.

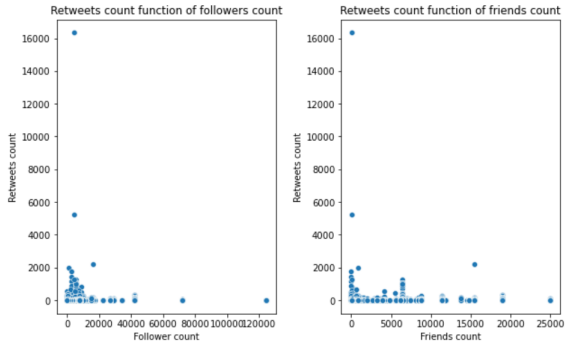

Figure 1: Retweet count depending on follower and friend counts

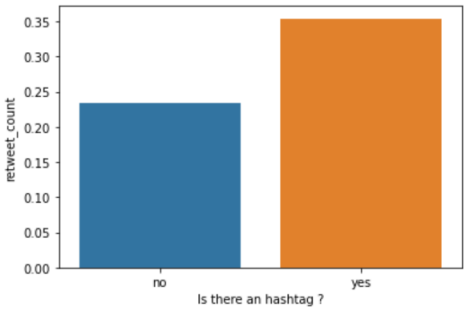

Figure 2: Retweet count depending on presence of an hashtag.

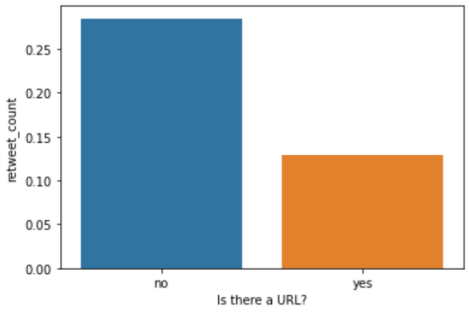

Figure 3: Retweet count depending on presence of a URL

Facing the possible issue of having data not really interpretable it seemed interesting to try building cluster to perform prediction on specific tweets, in order to get more accuracy.

4.2 Clustering models

Because of the disparate reparation of our data we were looking for a method to gather the tweet into groups with similar retweets number. This group will have pattern in their features such as followers count, friends counts etc. If we achieve to create these groups we can solve our retweet number prediction problem in a simple way. Indeed, to predict the retweet number of a given tweet, we just have to associate him to the best group with similar pattern in the features. Then, because of the similarity in the retweet count inside a given group, we can predict the retweet number easily by taking the mean retweet number inside the group.

This method is about clustering, as we studied during the ADA course. Thus we used the K-Means algorithm, which data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. The K-means algorithm needs to know the number of cluster a-priori. To determine it we are using the elbow method: Find the ”elbow” in the curve of the Sum of Squared Errors.

4.3 Predictive models

A the first sight, the more followers (or friends) a user has, the more retweet he will get. So we started our data analysis with a simple linear regression in order to look at this hypothesis. Using the tools we learned in the ADA course, we first made a fit, then evaluate it with a R^2 score, with regards to tuned hyper parameters. Then we

took more sophisticated models such as Gradient Boosting or Logistic Regression in order to try to get better fit.

5 Results and Findings

Using the elbow method, we find a optimal values of 5 clusters for our data. We wanted to have a low

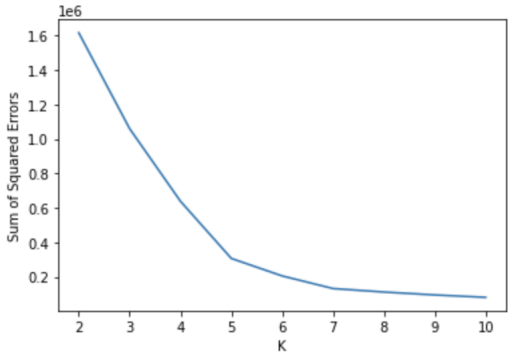

Figure 4: Elbow test for clustering

standard deviation into the cluster in order to be able to predict the retweet count. Unfortunately, as you can see on the graph below, our standard deviation were too high to allow us to have a good similarity in the retweet count into the cluster.

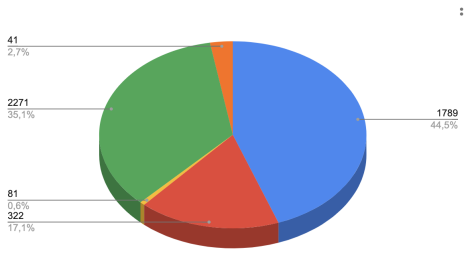

Figure 5: Standard deviation in the clusters

When training linear regression models we got pretty bad R^2 values for all tests even with more restrained clusters (often around 0.05) and expand features (polynomial basis). This was more or less expected regarding the dependency of retweet count to follower, hashtags or URL. However when using Gradient Boosting regressor the results were always better and we managed to get 0.78 for R^2 with large features (URL, follower count, friends count, hour of the tweet, day...). In fact this result is still kind of bad, even if it means that our prediction is getting closer to the targets, because the prediction obtained from a testing set gave some inconsistent results.

Beyond that it seemed that using the hour at which a tweet is posted into the training features provides much better prediction, so its influence on retweet count is not to be overlooked.

6 Conclusion

To conclude we figured out that using a predefined dataset, especially the one we tried to analyze, is not really the most efficient way of performing the retweet count prediction. Indeed, because of the lack of information regarding the follower count and the numerous outliers detected, the model was not that good despite a lot of processing. As highlighted in other papers (Liang, Hai and Fu, King-wa., 2015) the influence of tweets 'mood' and is context are important parameters to take into account when analyzing retweet and more globally popularity on twitter. This project showed the importance of choosing the right dataset and that sometimes bringing additional analysis to a paper cannot be conclusive at all.

References

[Liang, Hai and Fu, King-wa.2015] *Testing propositions derived from Twitter studies: Generalization and replication in computational social science.* August 19, 2015

[Nelson Joukov Costa de Oliveira2018] *Retweet Predictive Model in Twitter* Master's Degree in Informatics Engineering Dissertation January, 2018

[P. Pirolli B. Suh, L. Hong and E. H. Chi.2010] *Want to be retweeted? large scale analytics on factors impacting retweet in twitter network* In Proceedings of the 2nd IEEE International Conference on Social Computing, SOCIALCOM, pages 177–184, 2010.