TEXT MINING TWITTER: COLOMBIA 2018 ELECTIONS

JUAN SEBASTIN GARCA RODRIGUEZ
TRAVIS DUNLOP

Abstract

Between May and June of 2018 the people of Colombia will vote for their next president. As with any modern election, people are using Twitter, the social media platform, to support candidates they like, discredit the others, and debate who should win. Twitter provides a massive open forum to create dialogue across the country. Some groups take advantage of this by creating *social bots*, which automatically post political tweets to in an attempt to sway votors [7]. In our project, we use the text mining skills gained in this course to assess the influence of these bots. We hope to answer two questions: what percentage of users tweeting about the election are bots and what is the sentiment of the tweets?

INTRODUCTION

In order to answer these two questions we split the project into two sections: bot detection and sentiment analysis. For bot detection we leverage an existing algorithm and try both supervised and semi-supervised learning to extrapolate it's findings. With sentiment analysis, we hand-labeled a portion of the tweets and use those to help classify the rest into positive, negative, or neutral.

DATA SET

Since February we collected tweets that were related to the political situation and in specific about the Colombian elections for Senate and President. We collected 3.1 million of tweets from 3th of March to the 7th of May. We got information from 326,966 users and we have identified the following candidates: Gustavo Petro, Humberto de la Calle, Sergio Fajardo, Ivan Duque and German Vargas Lleras. Also, we collected tweets that were talking about politicians who are indirectly involved in the elections such as Alvaro Uribe and Claudia Lopez. We have this data set stored in a SQL data base and we used Python to process the data.

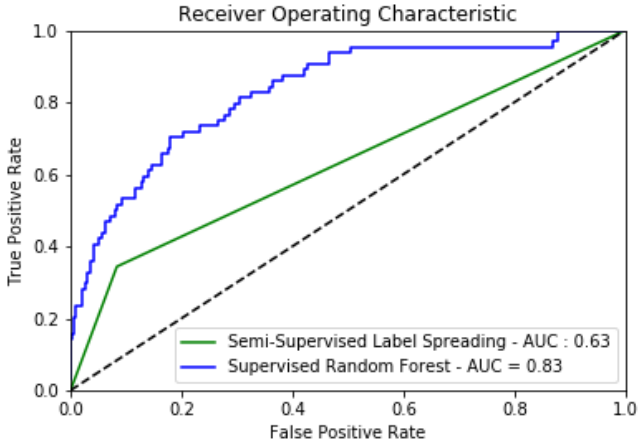
Dectecting social bots is a complicated phenomenon, because they usually aim to not be found. This results in a game of cat-and-mouse - researchers attempting more sophisticated classification strategies, and bot-makers emulating increasingly human-like behavior. One technique developed to identify bots is to use a *honeypot*. In this case, the honeypot is a set of researcher-created Twitter users who tweet mostly nonsense. The users that interact with them are likely to be bots – exploiting their desire to engage with many users. Of course, some real users just happen to message or follow one of these bots. In order to parse out which are the real, the researchers use unsupervised learning to cluster the data into groups. They find that some of the clusters seem more human-like than others. This particular strategy was developed and implemented by Lee et. all [3]. Once this group of bots are identified, features are extracted to compare them with other users. The researchers use things like frequency of tweets, number of followers, time between posting, ratio of tweets to retweets, even the length of the username as features in the model. Another group from the Indiana University Network Science Institute has made this model available as a API [2]. Unfortunately, the API has limits on access and so, we label some users and then build a model to extrapolate our results to label other users.

The features for the model is built by accessing Twitter's API for user data [1]. We include as features:

- account age
- tweets count
- friends count
- user description
- favorites count
- username length
- location listed
- protected status
- default image

To process the user description, we were heavily inspired by the paper by Hannes Mueller predicting conflicts from newspaper text [4]. We first take the raw user description and create a bag-of-words model, converting each description into a vector where the entries are counts of each token. With those vectors, we perform Latent Dirichlet Analysis utilizing the gensim package [6]. We do this as method of feature engineering and dimension reduction.

Since we only have labels for a portion of the users, we wanted to try a semi-supervised learning technique to leverage any underlying structure uncovered from the unlabeled data. We used the Label Spreading method from the sklearn package [5]. Unfortunately, it didn't work very well, managing an area-under-curve metric of 0.63. So instead, we tried a Random Forest classifier that worked much better - area-under-curve of 0.83. Below are the ROC curves for the two techniques.



In the end we had labels from 4,247 users and we had features from 8,721. Of the whole dataset, we find that 3% of them are bots.

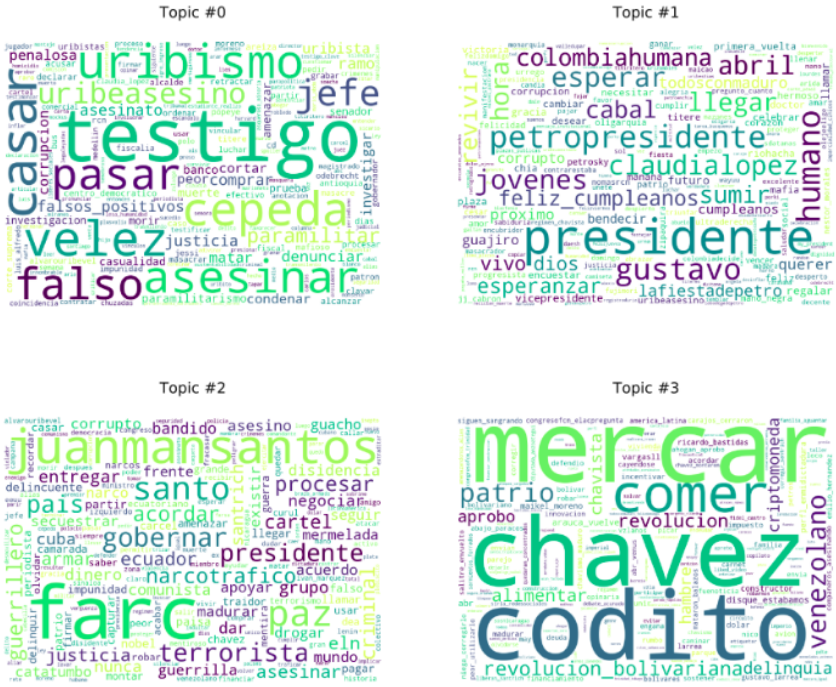
SENTIMENT ANALYSIS

In political times people tend to express in social media more fiercely their feelings about the current situation This time Colombians are divided between people who support the peace deal and people who support a reformulation of it. Also, there is a lot of uncertainty about what's next for the country since the peace deal is being implemented and Colombia is still living with crime and war. To approach such negative feelings we implement topic modeling to extract weekly topics in our data set and in parallel we train a model with a semi-supervised technique for learning the sentiment in those tweets. Respectively, we used a Latent Dirichlet Allocation(LDA) implementation from the package gensim and LabelSpreading from sklearn.

For learning the weekly topics we processed the tweets from each user as an entire document and then we gathered this as a weekly collection of documents. Since we noticed that the resulting topics from the model were more diffused when we provided each tweet as a document. For the topic modeling, we processed only the raw data for the week between April 16 to April 22. Additionally, during the training process, we noticed that some words (such as, "Gustavo petrol","Colombia") were highly frequent in each document, such that those were present in all the resulting topics. Then we decided to add them as stopwords. Also, for the sentiment analysis, we used a randomized subset of the data for labeling and we manually labeled 1800 tweets. With this data, we ran a supervised classification algorithm and a semisupervised classification algorithm.

Results

The topics modeling was trained with several topics in order to find the best fit for the tweets in this week for model selection we used the coherence score and perplexity, which are metrics implemented for evaluating these models. The best model that ge got has 10 topics with a coherence score of 0.46 and perplexity around 8.7. In the following wordclouds we can see the 4 most important topics for the week:



In the first topic, we see words related to Alvaro Uribe and recent news about the "Falsos Positivos" scandal. During Uribe's term as president of Colombia ten thousand young men were dressed up as FARC members and murdered by the Colombian military. During the week we analyzed tweets, a key witness tying Uribe to this scandal was found murdered.

The second topic, show us a topic related to tweets that were supporting Gustavo Petro campaign "Colombia Humana", where the words directly say "petropresidente". Interestingly the word Claudia Lopez appears probable because Petro's followers want that he merge forces Petro's campaign with Sergio Fajardo and Claudia Lopez.

The third topic introduces us to the current president and the topic is about the peace deal. People tweets probably talk about the fact the current president is giving to FARC seats in the Senate and other political rights as an official party, so people do not still process this as a good fact, also they might think that FARC is still a group of narcos and terrorists. Finally, the fourth topic is about Venezuela and the cryptocurrency Petro which is something that we were expecting since one of the keywords in our twitter scrapper is taking "Petro" as a keyword.

CONCLUSION

In the end, we have shown some evidence that

REFERENCES

[1] Twitter API . `https://developer.twitter.com/en/docs`. Accessed: 2018-07-19.

[2] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. *CoRR*, abs/1602.00975, 2016.

[3] Kyumin Lee, Brian Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*, 2011.

[4] HANNES MUELLER and CHRISTOPHER RAUH. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2):358375, 2018.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[7] Jon Swaine. Twitter admits far more russian bots posted on election than it had disclosed. *The Guardian*, Jan 2018.