Tweeting about #AOC: a text and semantic analysis perspective.

Computational Social Science Course – MSc Data Science
Antonia Donvito
antonia.donvito@studenti.unitn.it

1. Introduction

On 3 November 2020, towards the end of a year rich in critical episodes, the world will be assisting at yet another very crucial event: the American presidential elections. In 2016, when Trump was voted president of the United States for the first time, the role of social networks as a key tool of political communication for both candidates and voters already emerged. This time, however, due to the covid19 pandemic, the battle between the republican candidate Donald Trump and the democrat Joe Biden is going almost entirely online, to the point that even the traditional face-to-face presidential debates will be conducted virtually.

While the US, once again, will not have a female president, a record number of women are running for Congress on the same day as the presidential election[1]. Among them, Representative Alexandria Ocasio-Cortez has been in the spotlight for many reasons. Born in the Bronx to a working-class family and of Puerto Rican descent, in 2018, at the age of 29, AOC became the youngest woman ever elected to Congress. During the current election, before his drop-out, she endorsed Bernie Sanders, and she is currently supporting Biden's campaign. Before getting to politics, she worked as a waitress and a bartender to support her family. With her agenda focused on social, economic and environmental justice, she is now considered the rising star of the Democratic Party, and through her digital communication strategy and her growing presence on social media, she is amplifying her political messages also online (Lewinstein, 2019).

2. Research Questions

The aim of this exploratory research is to investigate what people write about Alexandria Ocasio-Cortez on social media and, more specifically, on Twitter. A growing number of studies in computational social sciences has focused on the use of Twitter in politics (Jungherr, 2014) with different scopes: predicting political alignment (Conover et al., 2011); looking for features in political speech, as in Trump's tweeting style (Clarke & Grieve , 2019); mining public opinion on specific subjects, such as economic issues (Karami, Bennett, & He, 2018). Overall, these works have highlighted the need to integrate traditional opinion polls and surveys with social media data. The reliability of Twitter data to make predictions, compared to polling aggregates, was extensively validated by a study

[1] https://www.npr.org/2020/06/16/878852938/record-number-of-women-run-for-congress-in-2020

on tweets about Hilary Clinton and Donald Trump (Bovet, Morone, & Maks, 2018). Similar findings were reported in the UK, with a study on tweets from Members of Parliament (Thapen & Ghanem, 2013).

Making predictions on Alexandria Ocasio-Cortez's election at the Congress is beyond the scope of this study; nonetheless, I will provide an exploratory analysis of the contents of tweets that mention her to illustrate what topics Twitter users link to her and whom and what they connect her with. To find such connections, I will first look for the most common words. Then I will focus on the most popular bigrams and trigrams, which have been proved as an effective preliminary step to identify themes and trends in tweets (Conway M., 2009; Yoon, Elhadad, & Bakken, 2013). Finally, I will conduct a topic modelling analysis, which is a method to identify the main theme covered by a large collection of documents (tweets in this context). To do so, I will rely on the most popular algorithm for topic modelling, Latent Dirichlet Allocation (LDA), which is a Bayesian Hierarchy model, in which a set of text data is modeled as a mix of various topics. This method has been largely tested on tweets (Negara et al., 2019; Montenegro et al., 2018).

To sum up, this paper intends to answer the following questions: R1) What are the most common words of the tweets related to Alexandria Ocasio-Cortez? R2) What words are linked together in terms of co-occurrences? R3) What are the main topics of the tweets?

3. Methodology

To start with, data was collected using Tweepy[2], an open source Python library for accessing Twitter API. Tweets in English containing at least one of the hashtags or keywords *#AOC*, *#alexandriaocasiocortez*, *AOC*, *Alexandria Ocasio-Cortez*, were collected over three different days randomly chosen, between 24 and 30 August. Data was then filtered by ID, the identification number that is uniquely assigned to each tweet. At the end of the process, the dataset resulted to be made up of 24253 tweets, including 12037 retweets, that in turn can be reconducted to 2304 distinct tweets. In total, the tweets with unique content are 14479, shared by 19667 different users. Together with the content of the tweets and their IDs, other features were collected, such as the datetime, the retweet count (how many times a tweet was retweeted), the hashtags being used and, in case of retweets, the extended original text (as they are truncated in Tweepy).

The study consists of three main parts. First, a simple exploratory analysis was carried out in order to extract preliminary insights on the tweets. The analysis was run on unique tweets, rather than on the whole dataset, in order to avoid redundancy caused by frequently retweeted contents. Therefore, the tweets considered at this stage were 14479. While hashtags extraction did not require

[2] Documentation at http://docs.tweepy.org/en/latest/

any specific preprocessing, text cleaning operations were performed to find the most common words. Numbers and symbols were removed, all letters lowered, mentions ("@" followed by a username) and hashtags discarded, stopwords eliminated; finally, all the remaining words were lemmatized to reduce the inflected forms of words to a single item.

Second, a simple ngram investigation was carried out to capture relations between words. To do so, the most common bigrams and trigrams were identified by means of the CollectionFinder functions included in the nltk package[3]. What is it meant by "common" in this context? The metric that was chosen to measure the ngrams relevance is not the raw frequency (i.e. how many times two words make a couple, or three words make a trio), but the likelihood ratio. Likelihood ration takes into account both the frequency of the words taken together and the frequency of each word of the bigram or the trigram throughout the corpus as well. This allows to get a better understanding of what information is relevant in this specific dataset.

The last part of the analysis is centered on topic modelling through Latent Dirichlet Allocation (LDA), implemented by the LdaMulticore function provided by the gensim library[4]. To run the algorithm, some preparatory steps were to be taken. In addition to the text cleaning that was performed previously (such as the stopwords removal), all word functions but nouns and adjectives were discarded. This approach, which is known as the noun only approach, has been proven to increase the topic coherence and the speed of the model (Martin & Johnson, 2015). To assign the function to each word, the pos_tag function included in the nltk package was employed. To find the best number of topics, the model was run with different topic sizes, measuring the coherence score for each size, until the best score was found.

4. Results

4.1 Most frequent words

To start with, the most retweeted tweet, with a retweet count of 85306, is:

> *RT @cspan: Rep @AOC: "I do not need Rep. Yoho to apologize to me. Clearly he does not want to. Clearly when given the opportunity he will not & I will not stay up late at night waiting for an apology from a man who has no remorse over calling women & using abusive language towards women."* https://t.co/XKymFh3Oyf[5]

Taking aside #AOC, which appears more than 5000 times because of the data collection criteria, the most frequent hashtags are *#DNC*, *#aoc*, *#DemConvention*, *#BernieSanders* and *#Biden*, each of which with hundreds of appearances. As the histogram below shows (Fig 1), the most frequent hashtags are

[3] Documentation at https://www.nltk.org/howto/collocations.html
[4] Documentation at https://radimrehurek.com/gensim/models/ldamodel.html
[5] https://twitter.com/cspan/status/1286315637275598849

politics-related, and some of them specifically refer to the election candidates. The only exception is *#BLM*, the acronym for Black Lives Matter movement, which is used almost as many times as *#MAGA*, Trump's slogan "Make American Great Again".

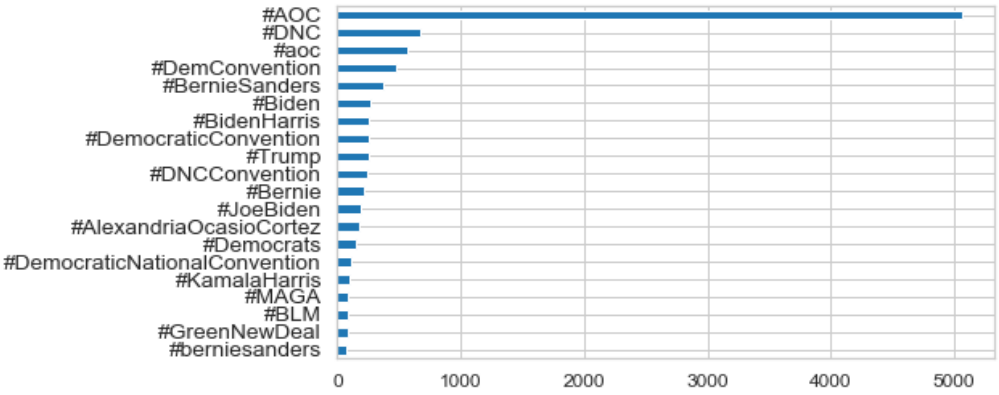

Figure 1*: first 20 most frequent hashtags*

Taking a step forward, after data pre-processing and lemmatization, I looked for the most frequent words. The wordcloud above (Figure 2) shows the 150 most frequent lemmas. Besides the names of AOC and of the presidential candidates that, it is possible to discern words such as *woman*, *president*, *black*, *support* and *American*, which are consistent with the what was discussed above. Interestingly, the word *bartender* was used 82 times; *waitress* 9 times; *Puerto Rico* 101 times; *b*tch* 455 times[6].

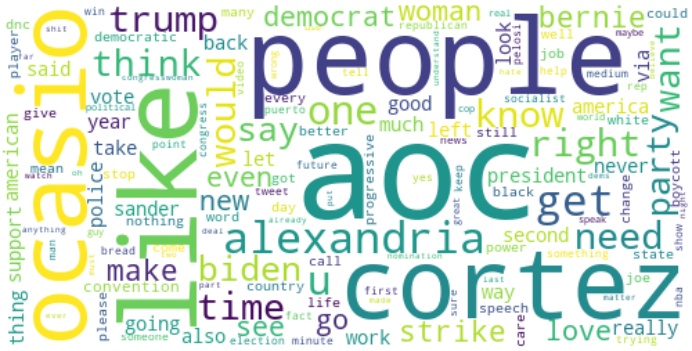

Figure 2*: wordcloud of the most frequent lemmas (hashtags excluded)*

To better capture the semantic relations between words, a bigrams analysis was performed as well. The "BigramCollocationFinder" included in the nltk package was used to find the most common bigrams, i.e. the words that frequently appear together. Similarly, the "TrigramCollectionFinder" was adopted to look for the most recent trigrams. As mentioned in the methodology section, likelihood ration was chosen to measure the popularity of ngrams, which implies that both the frequency of the

[6] See https://www.bbc.com/news/world-us-canada-53521143

ngrams and the frequency of the words taken alone were considered. The results are summarized by the histograms below (Figure 3).

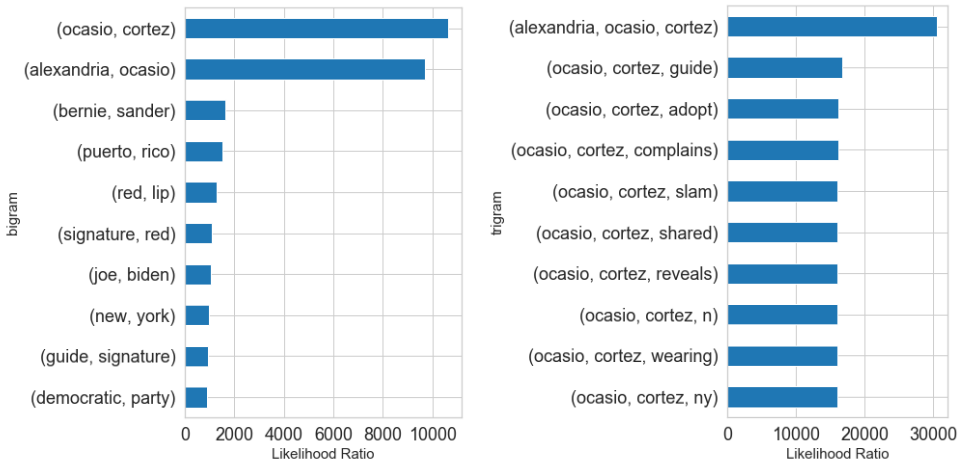

Figure 3: 10 most common bigrams (right chart) and trigrams (left chart) based on likelihood ratio

To better visualize such relations, I transferred the bigrams and their likelihood ratio value to an undirected graph having words as nodes. To make the graph readable, only the first 40 bigrams were considered to build the graph (Figure 4). It is possible to recognize, in the central part of the figure, a small cluster of words connected to AOC, such as *congresswoman* and *rep* (short for Representative), but also *red* and *lip*, which allude to her signature makeup. It is also easy to see that the adjective *new* often precedes *york* but it is also frequent in the expression *green new deal* (which indeed is a central theme of AOC's agenda).

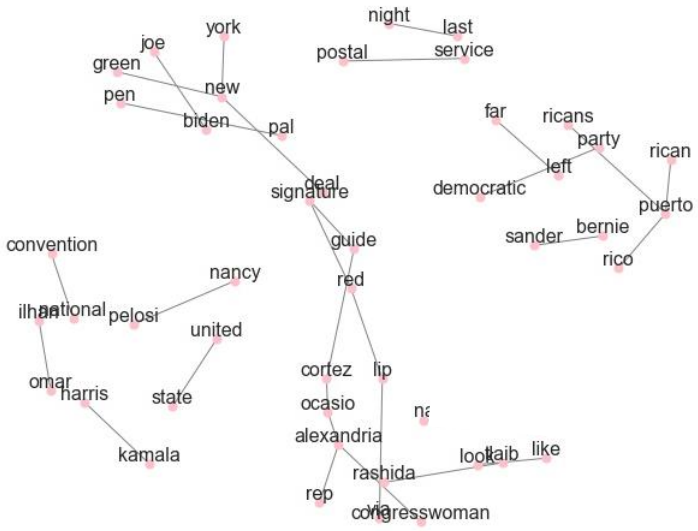

Figure 4: *semantic network of the most frequent bigrams*

4.2 Topic modelling

The second part of the analysis is centered on topic modelling through LDA, using the LdaMulticore function provided by the gensim library. After performing the preparatory steps and the model tuning to find the best topic size (Figure 5), four main topics were identified and can be visualized at this link[7].

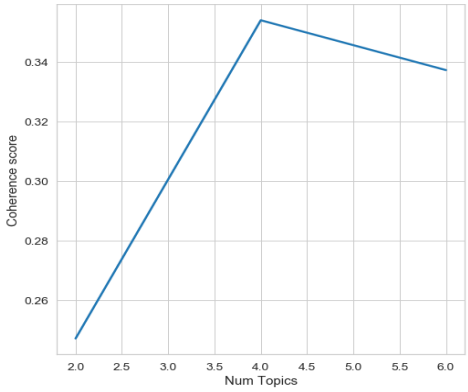

Figure 5: LDA model tuning of topic size

The ten most relevant words for each cluster are the following:

- Topic 1: *dnc*, *trump*, *biden*, *bernie*, *democrat*, *berniesanders*, *party*, *joebiden*, *demconvention*, *sander*;
- Topic 2: *strike, boycott, nba, word, player, word, point, medium, cnn, blm;*
- Topic 3: *need*, *right*, *police*, *black*, *white*, *time*, *berniceking*[8]*, cop, stop, care;*
- Topic 4: cortez, ocasio, alexandria, love, woman, congress, congresswoman, puerto, video, rep.

By reading into these results, it could be argued that the first topic groups together tweets that refer to the American elections or, more broadly, to politics. The model was used to classify the tweets of the dataset, by computing the classification probability of each topic for every tweet. The tweet that is assigned to this cluster with the highest probability says:

> *Trump attacks the whole Democratic Party from Biden to Bernie, from Pelosi to AOC. Biden goes after Trump alone.Leaving the whole GOP untouched. We have the weakest opposition possible. Biden wants to be friends with his old Republican friends. He doesn't want to fight for us.*[9]

The most relevant words to the second topic are most specific as they refer to the NBA world: more narrowly, by looking at the tweets of this cluster and at the events of the last days, this cluster builds on the NBA players' strike against racial discrimination. The tweet that is assigned to this cluster with the highest probability is a reply to an Alexandria Ocasio-Cortez's tweet[10] on those events and says:

[7] See the *topic_modelling.html* file in the main folder
[8] @BerniceKing: CEO of Martin Luther King Jr. Center for Nonviolent Social Change
[9] https://twitter.com/BadJohnBrown/status/1298650987402416128
[10] https://twitter.com/AOC/status/1298838471906193408

@AOC My guess is that the term strike is a less accepted term in the American society, due to the lack of power workers and worker unions have. Boycott is a widely known and accepted term, even though the interpretation by some is unique (buy & destroy to 'boycott' it)[11].

The third cluster, as the previous one, seems to be centered on racial discriminations too, but more specifically on the events in which police was involved. The tweet that is assigned to this cluster with the highest probability, and which is a reply to AOC once again, is:

@BerniceKing @AOC Have you though maybe just maybe the police know these people?Nah.Y'all to dumb. That black man had a knife. That black man had warrants.that black man was under arrest.that black man resisted arrestNone of that justifies being shot seven times. Why not support this #ZykierYoung

The last cluster looks as the most focused on Alexandria Ocasio-Cortez, with words that relates, for instance, to her political role and her origins (Puerto Rico). It has to be recalled that, although her name results to be the most significant word of this topic, it appears in some forms across all the clusters, as the tweets were filtered out based on such keyword in the very first place. The tweet that is assigned to this cluster with the highest probability is:

Alexandria Ocasio-Cortez slams 'proud' Latina Kimberly Guilfoyle for saying her Puerto Rican mother is an immigrant https://t.co/mcu4IOZ6zZ
Puerto Rico is part of the United States, but then Trump didn't even know that, and today the @GOP is whatever Trump is (on that day).[12]

Looking at the whole dataset, the topics are not equally distributed. As the graph below shows (Figure 6), Topic 1 is the most common, which suggests, as predictable, that most of the tweets are politics related and mention the main protagonists of the presidential elections. A large part of tweets is classified as Topic 3, the cluster that is likely to be related to racial discrimination events; this comes as no surprise, as racial discrimination has been massively brought up in the public and political debate. Interestingly, a minor part of the documents is classified as Topic 4, which indicates that not all the tweets that mention AOC actually talks specifically about her.

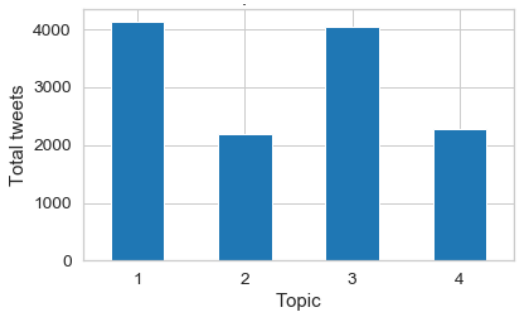

Figure 6: topics distribution over the dataset

[11] https://twitter.com/MasterOtenko/status/1298842513017581569
[12] https://twitter.com/SwampGreen/status/1298616629933334529

5. Discussion

Overall, the results are in reasonable agreement with the expectations and consistent with the main topic. The first and the second part of the analysis suggests that when people tweet about Alexandria Ocasio-Cortez, they mention her in the American election context together with the two candidates Trump and Biden as well as the ex-candidate Sanders. Other frequent mentions regard her mum's origin, Puerto Rico, being a woman (also in offensive terms) and her make-up. However, as the topic modelling has highlighted, one of the most pervasive topics is certainly the events around black discrimination, which is not only exemplified by the citation of George Floyd, BLM and NBA protests, but especially by the two clusters that were identified by the LDA model.

On this regard, it is extremely important to stress how two factors might have contributed to these specific outcomes, in order to provide guidelines on the methodology for future works on the subject. First, tweets shall be collected on a wider time window with a random approach, in order to get a more diversified dataset. For instance, if we look closely at the data, we may conclude that the persistence of make-up related tweets was due to a YouTube video that AOC shared in the data collection week. On the other hand, I would not expect much different results in terms of BLM mentions on a more spread period of time, as the racial crisis that America has been going through is growing day by day both in the streets and on social media. Second, the size of the dataset could be further increased to better generalize the results of the analysis. On this matter, the amount of data needed in social science research varies significantly across literature, and the size considered in this context is in line with many studies.

6. Conclusion

Alexandria Ocasio-Cortez is considered as a rising star of the Democratic Party of the Unites States and there are high expectations on the changes that she may bring to the US politics in the near future. While her growing social media presence has been a matter of research, there are not studies about what people tweet about her. This study has to be seen as a first exploratory analysis on what people write about when they mention her on Twitter, and it could be considered as a starting point for further research. Sentiment analysis, hate speech recognition and in-depth semantic network analysis (that were beyond the scope of this paper due to space limitations) would move the focus from content to perception, which would reveal how people are oriented towards her.

References

1. Bovet, A., Morone, F., & Maks, H. A. (2018). Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports - Nature, 8*(8673). doi:10.1038/s41598-018-26951-y

2. Clarke, I., & Grieve , J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE, 14*(9). doi:https://doi.org/10.1371/journal.pone.0222062

3. Conover, M. D., Gonçalves, B., Ratkiewicz, J., & Flammini, A. (2011). Predicting the Political Alignment of Twitter Users. *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT)*, (pp. 192-199). Boston.

4. Conway M., D. S. (2009). Classifying disease outbreak reports using n-grams and semantic features. *nt J Med Inform, 78*(12), 47–58.

5. Jungherr, A. (2014). Twitter in Politics: A Comprehensive Literature Review. *SSRN Electronic Journal*.

6. Karami, A., Bennett, L. S., & He, X. (2018). Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election. *International Journal of Strategic Decision Sciences (IJSDS), 9*(1).

7. Lewinstein, J. (2019). *Alexandria Ocasio-Cortez: A Case Study of Social Media as an Agenda Setting Tool in the U.S. House of Representatives.* Scripps College: Scripps Senior Theses.

8. Martin, F., & Johnson, M. (2015). More Efficient Topic Modelling Through a Noun Only Approach. *Proceedings of the Australasian Language Technology Association Workshop 2015.* Parramatta, Australia.

9. Montenegro C., L. C. (2018). Using Latent Dirichlet Allocation for Topic Modeling and Document Clustering of Dumaguete City Twitter Dataset. *Proceedings of the 2018 International Conference on Computing and Data Engineering (ICCDE 2018)*, (pp. 1-5). New York, US.

10. Negara E. S., T. D. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *International Conference on Electrical Engineering and Computer Science (ICECOS)*, (pp. 386-390). Batam Island, Indonesia.

11. Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the International Conference on Language Resources* (pp. 1320-1326). Malta: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

12. Thapen, N. A., & Ghanem, M. M. (2013). Towards Passive Political Opinion Polling using Twitter. *BCS SGAI SMA 2013: the BCS SGAI workshop on social media analysis.* Cambridge, UK.

13. Yoon, S., Elhadad, N., & Bakken, S. (2013). A Practical Approach for Content Mining of Tweets. *Am J Prev Med, 45*(1), 122-129.