In God We Trust

Luca Paoletti

Febraury 2023

Contents

1 Introduction 1

2 Research question and methodology 2

2.1 Research question 2

2.2 Methodology 2

3 Experimental results 3

3.1 Pre-Processing 3

3.2 Topic Modeling 6

3.3 Sentiment Analysis 8

4 Concluding remarks 10

1 Introduction

Religious 'nones', those who answer 'none of the above' on religious identification surveys, are the fastest-growing cohort of young adults in the United States, Western Europe and parts of Latin America.
They are also among the most stigmatized groups in religiously conservative parts of the developing world. Yet, many of them are deeply committed to values like tolerance, service and economic justice that are vital to healthy, stable societies.
Their sheer numbers in the West have already begun to reshape conversations about ethics and belief. And the deep penetration of social media in the global South, particularly Millennials, will likely mean that individualism, the touchstone of both globalization and religious disaffiliation, will shape religious culture even in parts of the world where affiliation trends remains high.

2 Research question and methodology

2.1 Research question

The project aims at extracting, pre-processign and analyzing tweets about non-religious groups. The analysis is about implementing *Sentiment Analysis (Gensim / Sklearn)* and *Topic Modelling (LDA)*. In addition, more descriptive stats and results will be presented (distribution of tweets, most used words and others). With these results I will try to classify different organizations and twitter profiles according to their style of communication and their main topics and language indicators. Moreover, after having detected the main topics in the non-religious groups profiles, I will extract sentiment and opinion from each of these clusters.

2.2 Methodology

Data has been collected thanks to the Twitter API, a powerful tool that lets people scrape tweet directly on their local storage. I searched tweets through the username of some non-religious groups starting from the 1st of March 2022 to the 31st of May 2022. The result has been a dataset of 7.353 tweets from several users. In order to run the project I decided to use only the text of the Tweet and the name of the page where the Tweet was published. Despite that, for further discussion it could be interesting to use also other features such as the number of likes, the number of retweets, the datetime and many others available from the API.

Once the data has been collected, the project has been divided in:

- Pre-Processing
- Topic Modeling (both with Gensim and ScikitLearn)
- Sentiment (Opinion) Analysis (of the whole corpus and of the single Topics)

In the following rows, I will present a bit of state of art used during the project in order to present metrics and algorithms used in a more technical way.
During pre-processing I have introduced only techniques of basic Natural Language Processing (NLP), for example: *tokenization*, the task of chopping a defined document unit up into pieces, called tokens, removing at the same time certain characters such as punctuation and stopwords; *stemming*, refers to a crude heuristic process that chops off the ends of words, often including the removal of devotional affixes; *lemmatization*, refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming only at removing inflectional endings only in order to obtain the base or dictionary form of a word, which is known as lemma.

Topic Modeling is based on an unsupervised problem; the documents have not been classified yet and consequently we are dealing with the absence of the response variable. I used a Latent Dirichlet Allocation (LDA) algorithm that can be understood as a generative process where documents are defined by a probability distribution over a set of topics T and a probability distribution of discrete words, in turn, establishes each topic. Then, there are 2 parts in LDA:

- the words that belong to a document, that we already know
- the words that belong to a topic or probability of words belonging to a topic, that we need to calculate.

To get this latter probability:

- go through each document and randomly assign each word in the document to one of k topics (k is chosen beforehand)

- for each document d, go trough each word w and compute:

  1. $P(t|d)$: the proportion of words in document d that are assigned to topic t. I tries to capture how many words belong to the topic t for a given document d, excluding the current word. If a lot of words from d belongs to t, it is more probable that words w belongs to t.
  2. $P(w|t)$: the proportion of assignments to topic t over all documents that come from this word w. Trying to capture how many documents are in topic t because of word w.
     LDA represents documents as a mixture of topics. Similarly, a topic is a mixture of words. If a word has high probability of being in a topic, all the documents having w will be more strongly associated with t as well.
     Similarly, if w is not very probable to be in t, the documents which contain the w will be having very low probability of being in t, because Tthe rest of the words in d will belong to some other topic and hence d will have a higher probability for those topic. So even if w gets added to t, it won't be bringing many such documents to t.

- Update the probability for the word w belonging to topic t, as

$$P(w \text{ with topoic } t) = P(t|d) \times P(w|t)$$

This algorithm is used both with Gensim and Sklearn in order to classify documents and words in the Topic Modeling part.

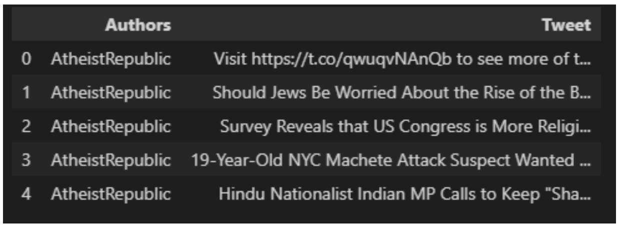
Finally, in the Sentiment (opinion) analysis I have introduced the Vader algorithm. Sentiment analysis is a text analysis method that detects polarity (e.g., a *positive* or *negative* opinion) within the text. It aims to measure the attitude, sentiments, evaluations and emotions of a speaker / writer based on the computational treatment of a subjectivity in a text.
VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive / negative) and intensity (strength) of emotions. It relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.
It does not try to determine if a sentence is objective or subjective, if it is a fact or an opinion. Rather, it only cares if the text expresses a *positive, negative* or *neutral* opinion.

3 Experimental results

3.1 Pre-Processing

Data collected from twitter come from a list of 276 non-religious users (ThinkingAtheist, ProgAtheistsInc, HumanismSpeaks, . . .). I decided to work with only the text of the tweet and not with other features such as datetime, likes and reposts. In addition, tweets are the ones identified with the English language and without retweet (often retweet can lead to change the focus of the analysis).
As we can see from the image above, the raw data are not 'clean' and need to be preprocessed before applying algorithms and other ML techniques.
First of all, on Twitter it is possible to find tweets equal to each other because of retweets and reply to tweet; as a consequence, I have checked for duplicate rows and in case I have removed them.
The results of this first part is the following:

- Total number of words: 348412 words
- Average number of words per tweet: 47.38 words



	Authors	Tweet
0	AtheistRepublic	Visit https://t.co/qwuqvNAnQb to see more of t...
1	AtheistRepublic	Should Jews Be Worried About the Rise of the B...
2	AtheistRepublic	Survey Reveals that US Congress is More Religi...
3	AtheistRepublic	19-Year-Old NYC Machete Attack Suspect Wanted ...
4	AtheistRepublic	Hindu Nationalist Indian MP Calls to Keep "Sha...

Figure 1: Data Overview

- Total length of dataset is: 3540327 characters
- Average Length of a tweet is: 481.0 characters

Another edit that I can apply is to the '@' symbol indicating a username. I have removed all the usernames and the only 'symbol' that I have decided to keep is the hashtag. Hashtags could be really significant when we are dealing with social network's messages, since they can define an opinion or a sentiment about a topic. In particular, in the analyzed pages, the use of hashtags has been really useful as a first overview of the topic covered by the project. Keywords or key-meaning are often used in order to underline some ideas, idles or opinions.

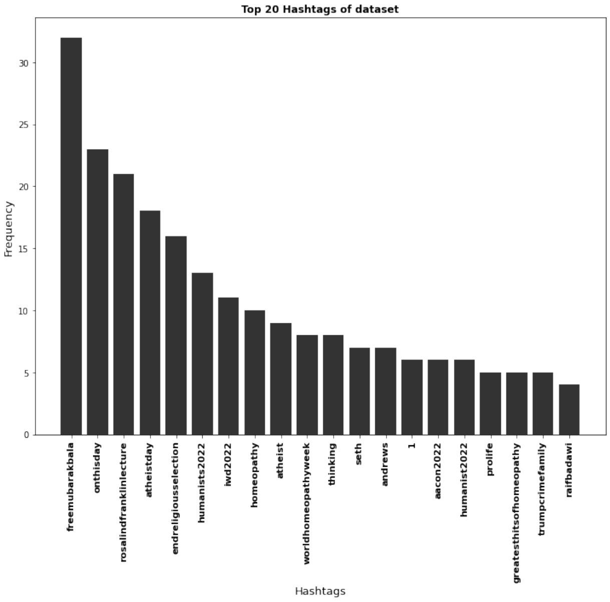


Figure 2: Top 20 Hashtags

Above I have plotted the most relevant and used hashtags in the data. I would say that most of them are significant in the period analyzed (freemubarakbala, iwd2022, humanists2022...), on the other hand I found also more generic hashtags such as atheistday, endreligiousselection or thinking. As far as I understand from this preliminary analysis, I would not say that these hasthtags can help with the sentiment analysis, but for sure they can let me understand which were the hot topics in that particular moment.
Finally I have removed links, punctuations, numbers and special characters that are not significant to the analysis. I have added to the stopwords some more tokens that I assume as not significant, such as https, twitter and pic.

A preliminary analysis of the words that populate the dataset is done with the next image, a wordcloud of the most used tokens among all the tweets.

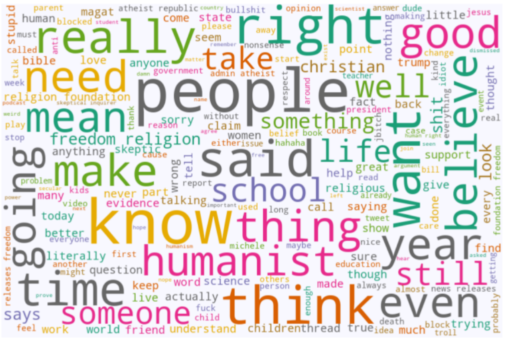


Figure 3: WordCloud

Of course the pic above is just a preliminary view of the texts, it is confused and not significant but it can certainly show the meaning of the pre-processing part of the project; only useful words are presented in the cloud.

While the wordcloud is a good way to analyze the content of the dataset, a more statistical plot can be useful too. In particular it is interesting how the pre-processing impacts on the raw data.

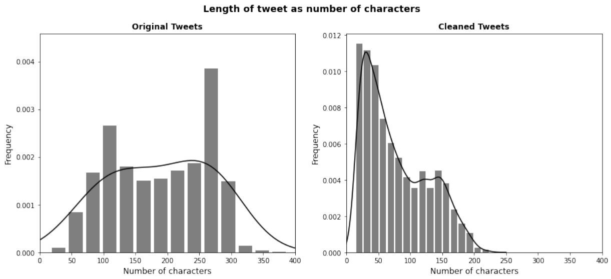


Figure 4: Original vs Cleaned Characters

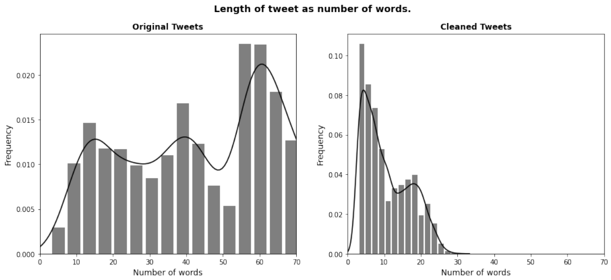


Figure 5: Original vs Cleaned Words

During the pre-processing part I have removed words with less than 3 characters and tweets with less than 3 words. The results are the ones above; many tweets for example had many stopwords, or links or usernames as it is shown in the 'cleaned' histograms. This phase is critical as it reduces the dimensionality and results in significantly valuable tokens for the models, which are explained in the next sections.
Now that the texts are completely clean and with only potentially significant tokens, we can proceed with the next phase, Topic Modeling.

3.2 Topic Modeling

The objective of this part is to cluster the documents in the most efficient way (tweets). In order to do that I used both the packages Gensim and ScikitLearn as to have comparable results.
The algorithm used is the LDA, already discussed in the Methodology section, that works slightly differently in the two packages, despite that, it can be useful to compare. In this paper I will focus on the Gensim's LDA results but I will also put some pictures of the Scikitlearn's results.

First I have taken a step backwards to take the pre-processed data to start the training of the model. In particular, I have considered important not only to use the 1-grams tokens obtained from the pre-processing, but I also got the bi-grams tokens to feed the model. These are some possible combinations of the tokens seen as a couple (bigrams).

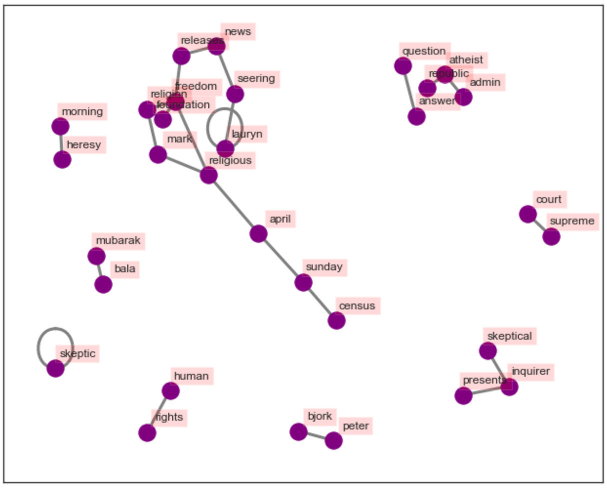


Figure 6: Bigrams

These are the most frequent bigrams among the tweets. It seems that the most used bigrams are mostly relevant in the humanist field, such as human-right, religion-freedom or morning heresy. At the same time, it is interesting to observe how the 'non-religious' topic is related to the government one as clearly underlined in the bigram supreme-court (also human-rights could lead to a political tasks). Of course, bigrams cannot find every type of topic / argument regarding tweets, but they are useful to understand which are the most relevant keywords when we are dealing with atheist or agnostic groups.

Lemmatization and Stemming are also applied to the tokens and now I am ready to proceed with the Topic Modeling algorithm. The main hyperparameter in the LDA model is the number of Topics. To find the optimal number of topics I have decided to run the model many times, always with a different number of topics, to get two evaluation metrics each time, in order to find the parameters that best fit the model. In particular, the evaluation metrics are:

- Coherence: scores a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference
- Perplexity: captures how surprised a model is from new data it has not seen before, and it is measured as the normalized log-likelihood of a held-out test set. However, recent studies have shown that predictive likelihood (perplexity) and human judgment are often not correlated, and even slightly anti-correlated.

The graphs below show how coherence changes with the increasing of the number of topics (and perplexity decreases at the same time).

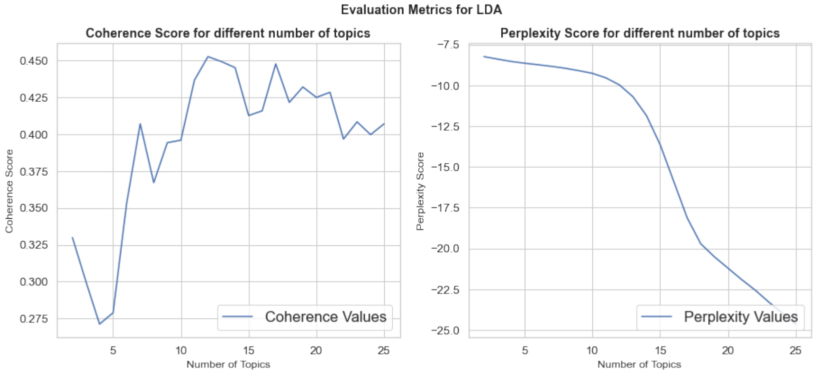


Figure 7: Evaluation Metrics for LDA

As the perplexity score seems to be the less reliable, I have picked 12 as number of topics, like the Coherence plot suggests.
So far, I have just defined the parameter of the model and built it. I can now deep dive a bit more into the newly created topic; it is interesting to analyze which are the keywords in each topic, how the topics are distributed among the documents and also which are the 'top' tweets of each topic.
Moreover, I have noticed how the topics extracted seem to be relevant in a specific timeline but they are not constant in every period (the majority of them); consequently, I have decided to use only the first month as subset of the dataset to make topic modeling. The results are more clear and it is easier to split among the different themes treated in each topic.
Since it seems that the topics are many and the number of documents is not, I have decided to explore the 4 topics with more tweets inside and compare them with the results of Sklearn LDA topic modeling.

Dominant Topic	**Topic Keywords (Gensim)**	**Topic Keywords (Sklearn)**
Topic 0	religi, right, year, belief, woman, report, object, constitut, human, work	people, life, help, make, think, want, need, know, skeptic
Topic 1	say, state, educ, support, want, govern, divest, come, case, show	know, think, say, time, actual, make, come, lose, trump, respect
Topic 2	religion, peopl, time, student, today, public, take, church, mani, thing	say, know, make, good, really, think, want, tell, claim, believe
Topic 3	secular, well, life, tell, book, chang, discuss, group, oblig, moral	religious, school, right, religion, humanist, child, state, support, foundation, woman

It is difficult to find similarities between topics of the two packages, despite the fact that both of them use LDA as topic modeling algorithm. As far as Gensim is concerned, I have reported only the first four topics (the most relevant ones) and it seems that the main ideas in the four groups are laws and order (Topic 0), government's role in education (Topic 1) religion in the school (Topic 2) and finally discussion makes life better.
On the other side, Sklearn seems to find more than 3 similar Topics (0, 1, 2) that are the ones underlined by Gensim, and finally a connection between atheism and school / child (Topic 3).
They are of course non comparable, but it is clear which are the main issues of discussion in these religious-none groups: humanism and activism.

As far as Sklearn outputs, only 4 topics have been defined as best model, I will use these topics to make sentiment analysis and opinion mining; I will also present a more general sentiment analysis on the whole dataset without splitting it into topics.

3.3 Sentiment Analysis

First, following the chapter above, I have presented a sentiment analysis of each of the Sklearn Topic. The VADER algorithm's results are quite simple to analyze due to the few 'sentiment' available. In fact, a document is classified as Positive, Neutral or Negative. The figures below, give the per-topic classification result of tweets to the sentiment class.

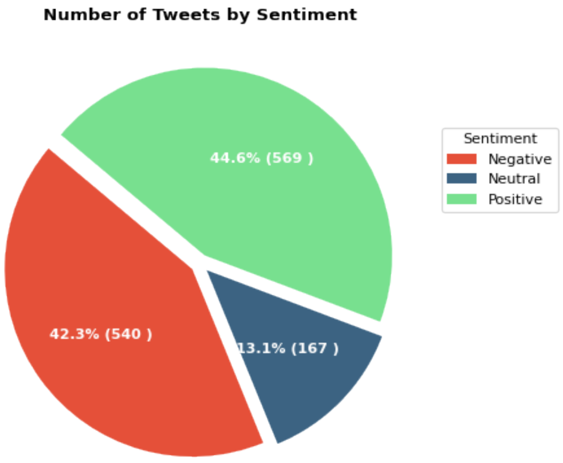


Figure 8: Topic 0

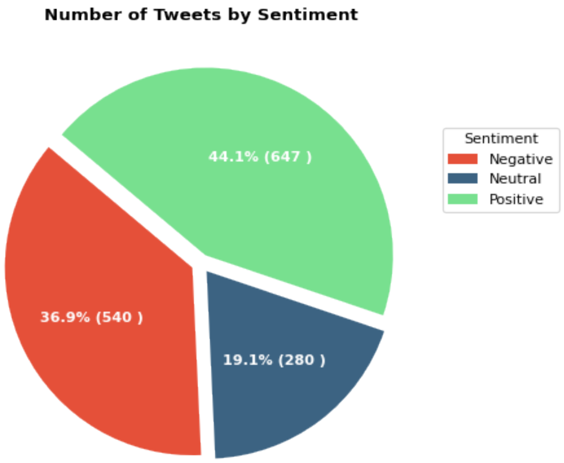


Figure 9: Topic 1

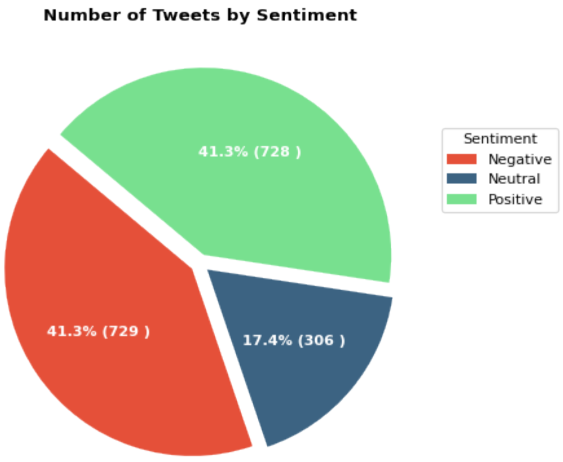


Figure 10: Topic 2

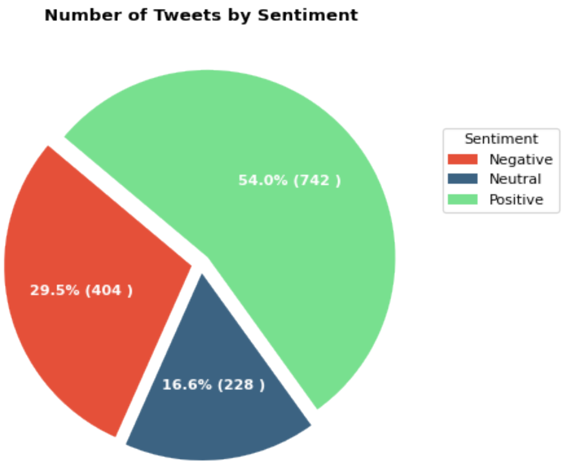


Figure 11: Topic 3

The overall sentiment seems to be positive, but, thanks to the analysis on the overall dataset (all the available tweets) I can have a look at the positive and negative tweets using word-cloud to illustrate the dominant words for each sentiment.

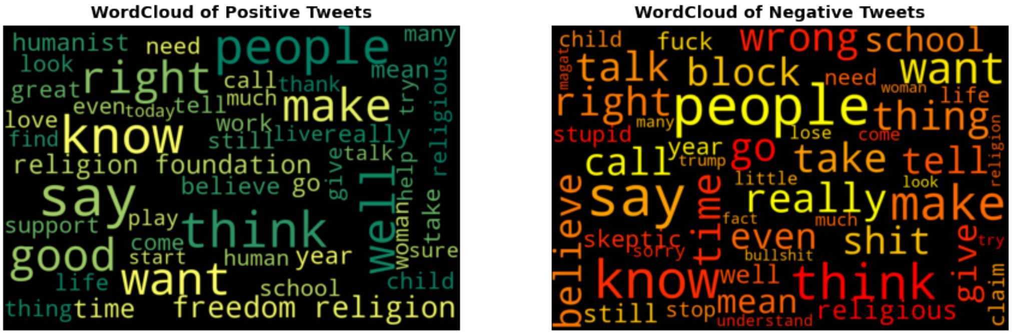


Figure 12: Wordcloud of Positive and Negative Tweets

Looking at these two wordclouds of words from positive and negative tweets, it seems that, even the worst tweet, the most aggressive one, seems to be aggressive because of badwords or very negative words (kill, sorry, little). In fact, there are many words in common between the two wordclouds, which means that the treated topics are the same, but with different sentiment and opinion. Among the positive tweets the most activist and humanist side of the argument is underlined, on the other hand, opinions and insults seems to stand out.

The last point of this research is the compound score's distribution. When we are dealing with sentiment analysis, the single sentiment that is given to a single document (tweet) is quite impossible to be totally positive or totally negative; in fact, a more weighted score is given: compound score. It corresponds to the sum of the valence score of each word in the lexicon and determines the degree of the sentiment rather than the actual value as opposed to the previous ones. Its value is between -1 (most extreme negative sentiment) and $+1$ (most extreme positive sentiment). Below the plot of the compound score divided into negative and positive tweets. What can I draw from this plot? Basically, the tweets labelled as negative and positive find difficulties to be totally positive or totally negative. In fact, the distribution of the compound score is more relevant between mid values of both the x-axis, and the extreme are always low. This is a confirmation of what I said in the positive and negative tweet's wordcloud: there are few terms that let the tweet be classified as positive or negative, most of the tokens are neutral or simply appear in both labels.

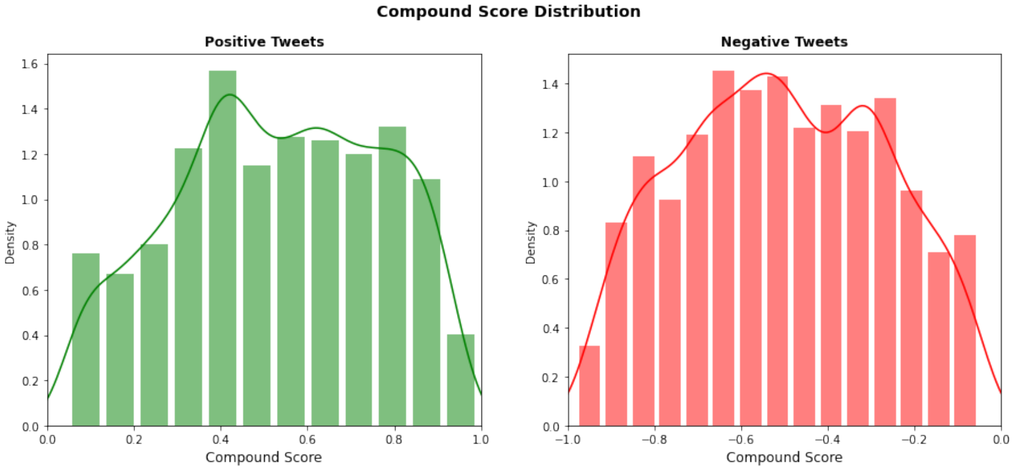


Figure 13: Compound Score

4 Concluding remarks

The aim of the project was to analyze non-religious groups on twitter. I started with a more statistics phase, based on number of tweets, tweets' mean length and all the other pre-processing metrics. Even at this point, it was possible to see the most used words and a bit of sentiment coming from the tweets.
Deep diving into the core part of the project, the topic modeling, I realized how general some of these twitter discussions are when it comes to religious-none users. Moreover, it seems that, despite the clear opinion about the religion that these people have, they deal with religion every day; words like 'religious', 'wedding', 'Christianity' are always among the most used ones.
As far as the non-religious theme is concerned, I understand how the main topics are about humanity, freedom, school, education and of course atheism, but always to be compared with the religious side of the medal. It is also worth noticing that in the last phase of the research that is the sentiment analysis, most of the tweets result to be classified as positive, maybe because the terms used in that field are commonly used in a positive way, or maybe simply because of humanistic side of this groups. The most negative tweets came from negative news on the papers (rapists, killers or wars) and consequently the sentiment of the tweet is quite clear. Unfortunately, Twitter is a powerful tool when we search for people's opinion and ideas, but at the same time, everyone can type on twitter any sort of comment including those that might not be strictly relevant with the topic of this discussion and this could negatively influence the analysis.

In conclusion, I would like to suggest a possible further discussion about this topic, consisting in searching and analyzing tweets during the main religious events (Easter, Christmas, Ramadan, ...) and comparing non-religious group's tweets with religious accounts. This might be a good way to find out what people really differently think about a common event.