Sentiment Analysis and Topic Modeling for Sarcasm Analytics on Reddit

Data Analytics

Cezar Angelo Sas - 781563

Contents

Chapter 1

Introduction

Sarcasm is a form of communication act in which the speakers implicitly communicate their message. The intrinsically ambiguous nature of sarcasm makes it laborious even for humans to determine whether a comment is sarcastic or not. Recognition of sarcasm can help many sentiment analysis NLP applications, such as review summarization, dialogue systems and review ranking systems.

To have an insight on how sarcasm is expressed on social media, I examined a dataset of self-annotated data from the social news website and forum Reddit. Posts on Reddit are organized by subject into user-created boards called "subreddits", which cover a variety of topics. The idea of Reddit involves evaluating posts and using upvotes or downvotes to measure their importance on the site.

The main purpose of the analysis consists in answering to these questions:

- Are any linguistic feature that can help detecting sarcasm?
- Is there a relation between sentiment and sarcasm?
- Which are the topics that people tend to react more sarcastically?

In the next Chapter, I will present the dataset and the analysis performed, answering to each research question individually.

Chapter 2

Analytics

2.1 Dataset

The dataset consists in comments from the social news aggregator Reddit. The comments span across different subreddits (subsidiary threads of Reddit). The dataset contains six columns as shown in Table 2.1:

Column	Description
label	Is the class to be predicted. The data were annotated by the person who posted the comment using "\sarcasm", a common way used on reddit to show sarcasm
comment	Contains the text on which the analysis will be performed
author	The author of the comment
subreddit	The category in which the comment belongs
date	The posting date
parent	The original post at which people comment

Table 2.1: Dataset columns description

The dataset consists in 1 million comments balanced with respect to the label. The records are divided in circa 15.000 subreddits, with the top 10 subreddits having 250.000 comments, this correspond to 25% of the total dataset. In Figure 2.1 we can see the posts distribution in the subreddits.

In Figure 2.2 we can see how sarcastic comments are more common on certain subreddit.

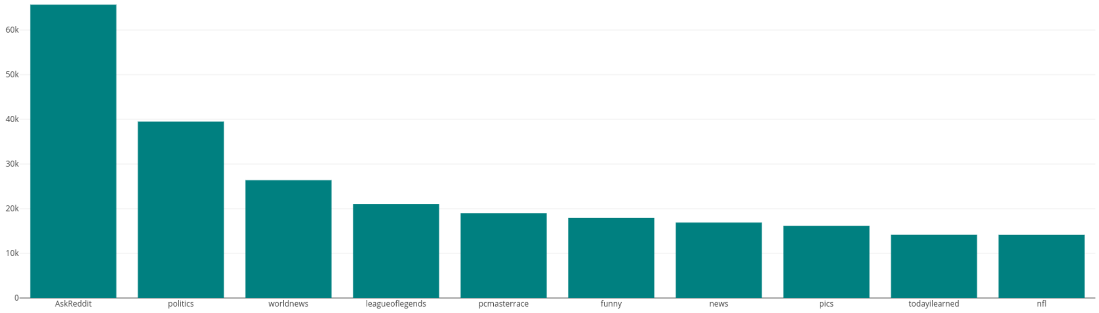


Figure 2.1: Top 10 subreddits with the number of comments

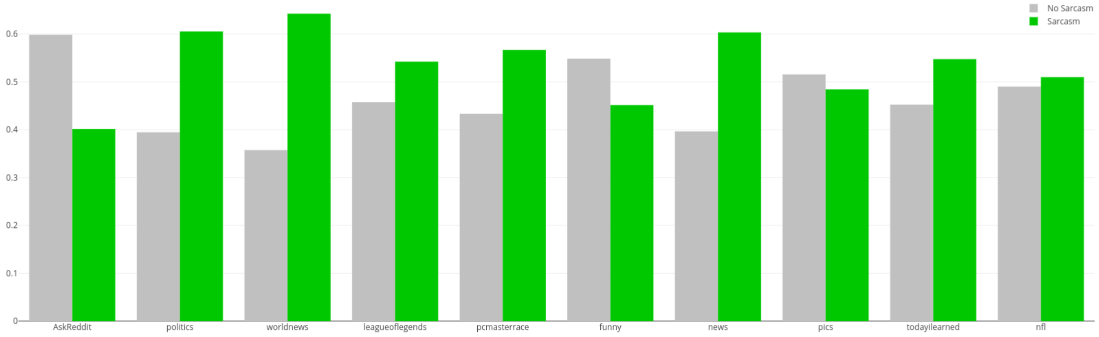


Figure 2.2: Subreddit posts division

2.2 Linguistic Analysis

I performed some linguistic analysis comparing the two classes and looked for differences that can give information on how a sarcastic text differs from a non sarcastic one in terms of formatting and writing style. This will help us to answer the first research question.

2.2.1 Formatting Stats

I performed analysis on the text formatting considering four different behaviors: the number of words in a comment, the number of characters, the percentage of capital chars in the comment, and the number of enlonged words. In Figure 2.3 we can see the average values on sarcastic comments and non sarcastic. The differences are subtle, so the formatting doesn't help separating the classes.

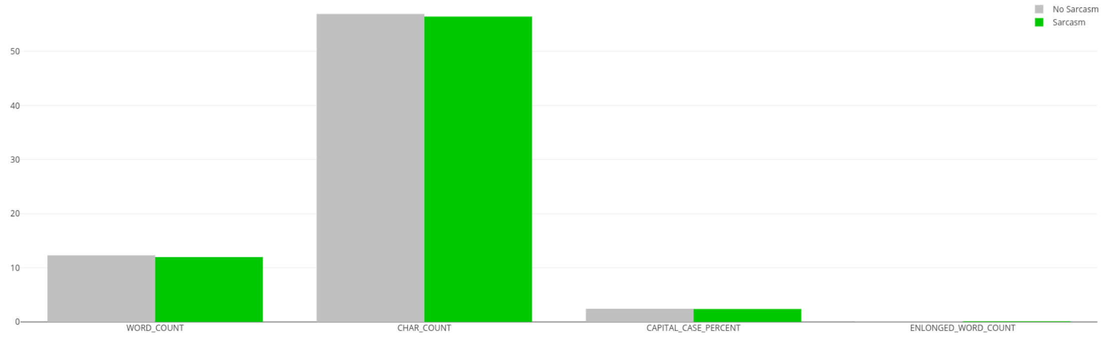


Figure 2.3: Text formatting statistics

2.2.2 Part-of-Speech Analysis

The second type of analysis consist in examining the distribution of part-of-speech tags. In Figure 2.4 we can see some distinction in how the POS tags appear in texts that are sarcastic, and in the one are not. In particular in non sarcastic comments we find more punctuation (PUNCT), nouns (NOUN), particles (PART), determiner (DET), and symbols (SYM). Instead verbs (VERB), adjectives (ADJ) are more frequent in sarcastic texts.

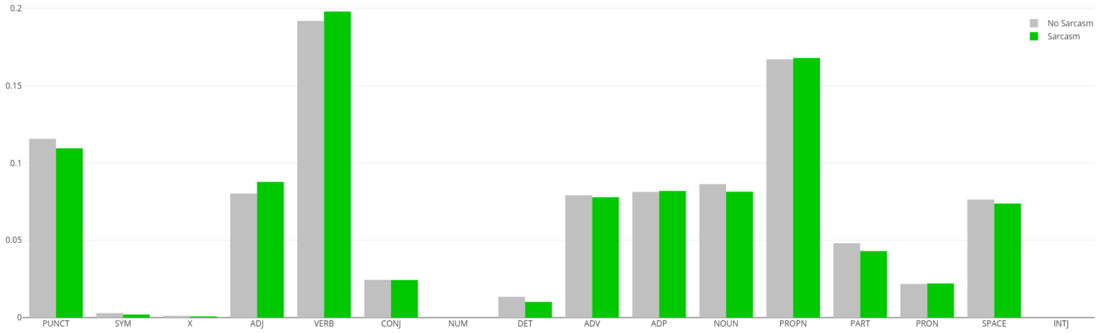


Figure 2.4: Part-of-Speech tags distribution

2.2.3 Words Usage

Now that we have a more clear view of the part-of-speech, we can ask which words are used in sarcastic comments and which in non sarcastic ones. Wordcloud in Figure 2.5 shows that the most frequent words “time”, “make”, and “yeah” are present in both types of comments, but with different frequency, for example the word “yeah” is present 3 times more in sarcastic comments, while the other words have a more similar frequency. Another important difference in words usage consists in the vocabulary size, non sarcastic text is more diverse as it has 120.000 words, in contrast the vocabulary size of sarcastic comments is 100.000 words.

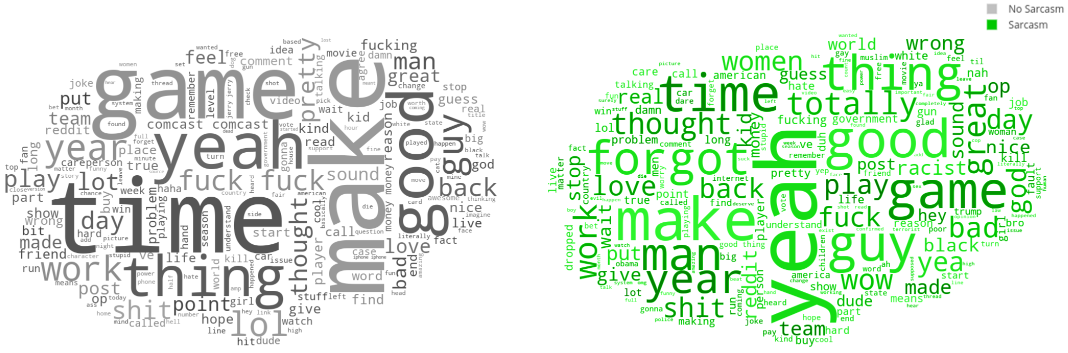


Figure 2.5: Wordclouds for non sarcastic and sarcastic comments

We can now answer the first question: are any linguistic feature that can help detecting sarcasm? The answer is yes, the linguistic analysis presented some characteristic that could help distinguish sarcastic comments from non sarcastic ones.

2.3 Sentiment Analysis

I also performed an analysis on the relation between sarcasm and sentiment. The comment’s sentiment was extracted using VADER Sentiment Analysis tool [1]. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool optimized for social media. The sentiment analysis results is a score in range $[-1, 1]$, and for a $score \leq -0.05$ we have negative sentiment

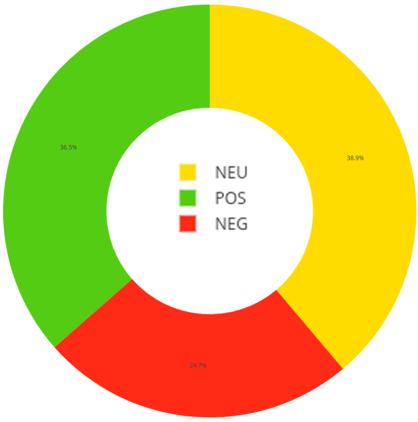


Figure 2.6: Sentiment percentage

(NEG), with $-0.05 < score < 0.05$ the sentiment is neutral (NEU), and for $score \geq 0.05$ the sentiment is positive (POS).

The sentiment of the 1.000.000 comments in the dataset are 36.5% POS, 39% NEU, and the remaining 25.5% in NEG (Figure 2.6). If we take in consideration the sarcasm (Figure 2.7) we can notice that comments with positive, and negative sentiment are more likely to be sarcastic. This can induce us to think that sarcasm is more common when the sentiment score is high. In fact, if we consider the average intensity, computed as the mean of all the absolute sentiment scores, we notice that sarcastic comments have, as expected, a higher intensity score (Figure 2.8).

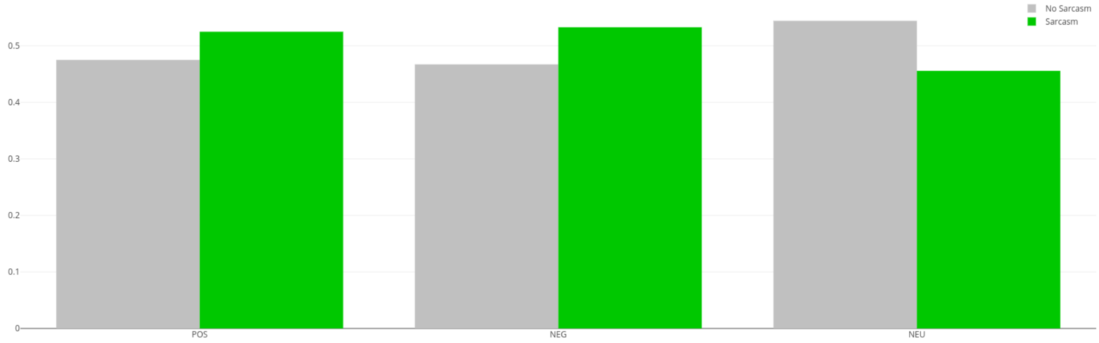


Figure 2.7: Sentiment count

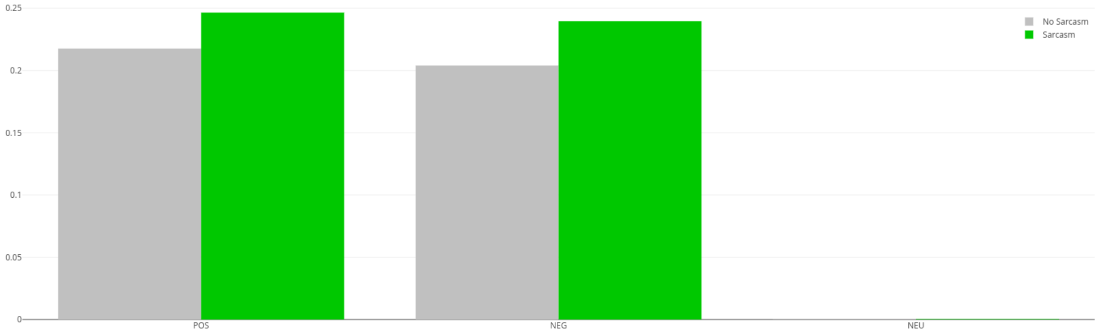


Figure 2.8: Average sentiment intensity

This brought me to the analyze words frequency by sentiment and see the most frequent words that are used in sarcastic and non sarcastic comments that are positive, negative or neutral. In Figures 2.9 and 2.10 we can notice that there is not much difference is the words used, except for the frequency as mentioned in section 2.2.3

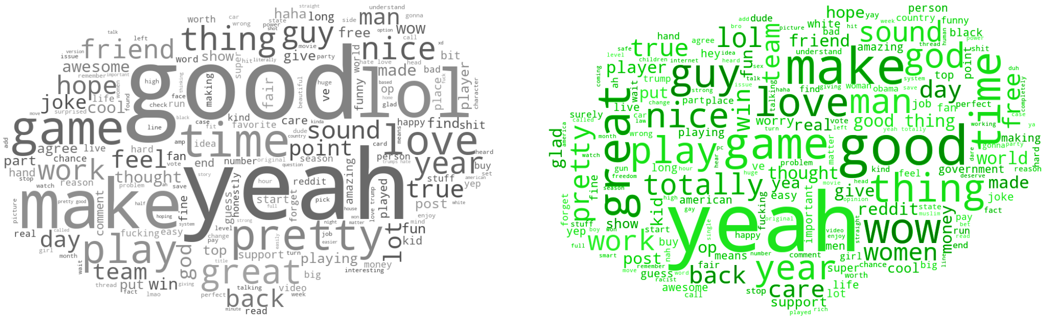


Figure 2.9: Wordclouds for positive comments

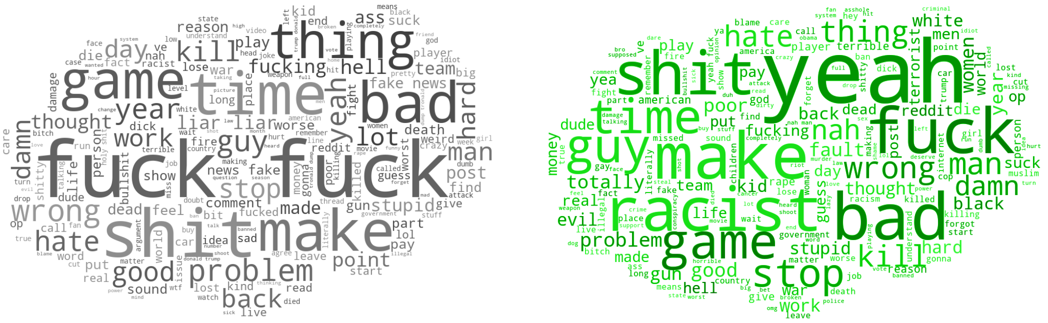


Figure 2.10: Wordclouds for negative comments

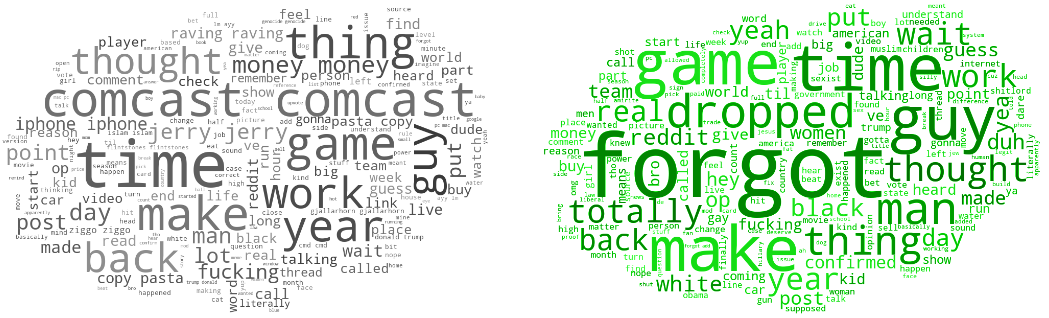


Figure 2.11: Wordclouds for neutral comments

As for the second question; is there a relation between sentiment and sarcasm? Yes, the analysis have shown that comments with a higher sentiment score are more likely to be sarcastic.

2.4 Topic Modeling

Given the high number of subreddits, I investigated the possibility of reducing the total number of the topics in the collection by performing a topic modeling, and group the post by they new topic. This step is required for answering our third question. For the topic modeling I used the Latent Dirichlet Allocation (LDA), a generative model for automatically discover the topics contained in a collection of documents.

2.4.1 Text Preprocessing

Before we can train the LDA model, we need to preprocess the text. The preprocessing phase consists in the removal of: emails, URLs, words containing numbers, stopwords, and non alphabet characters; the final preprocessing step consists in lemming the words.

2.4.2 LDA Model Training

LDA requires the number of topic as an input, in our case the value is unknown so I performed an research using different values for the number of topics, and evaluated the results using the topic coherence function called C_V *Coherence* [2].

I first performed topic modeling using the comments text, the number of topics tested are in the range $(5, 175)$, but as we can see in Figure 2.12 the optimal value is still a high number of topics, around 120. To reduce this value, I also performed this analysis on the comment's parent text. This, as we can see in Figure 2.13, not only increased the coherence up to 54, but also significantly reduced the number of topics from 120 to 10-20.

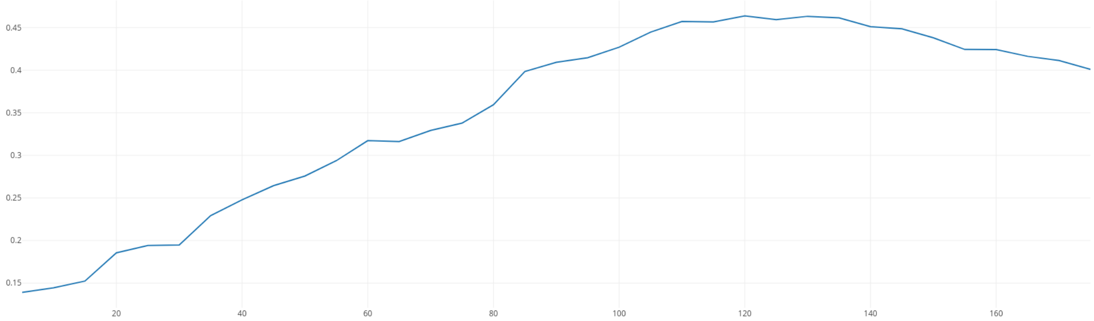


Figure 2.12: LDA C_V Coherence trained with comments

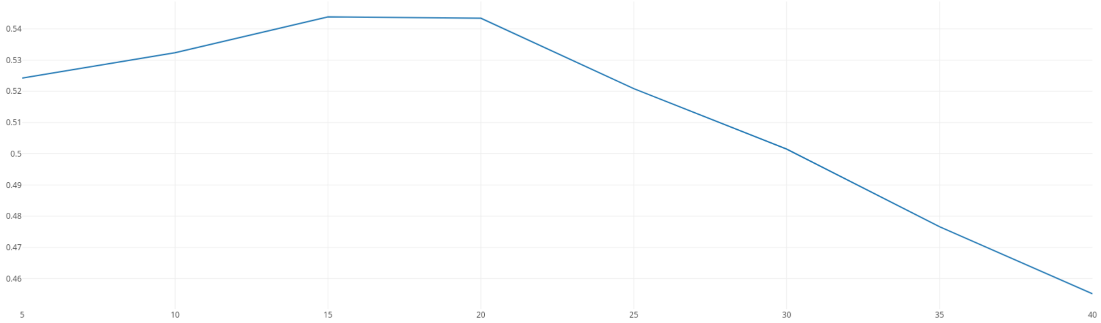


Figure 2.13: LDA C_V Coherence trained with parents

Given the probabilistic nature of LDA, and the close coherence scores for the number of topics in range $(10, 30)$, I executed multiple runs of LDA, to obtain a interval for the coherence with different number of topics. In Figure 2.14, we can notice the highest score for coherence is obtained with 10 topics (and the model $C_V Coherence$ of 0.63). This value will be used for the next experiments.

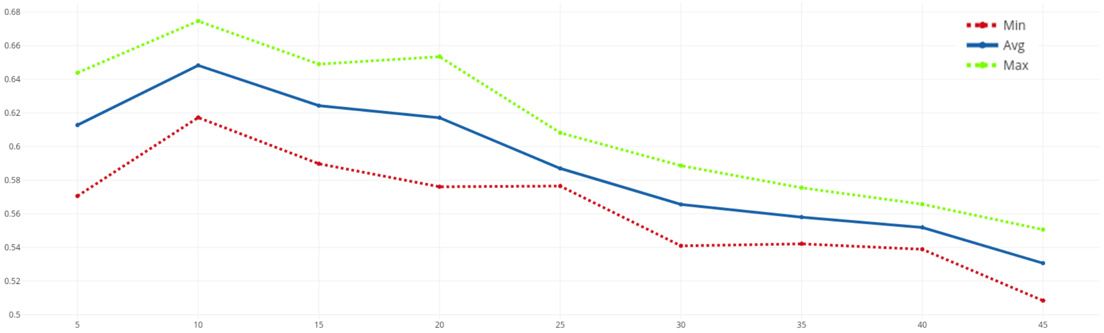
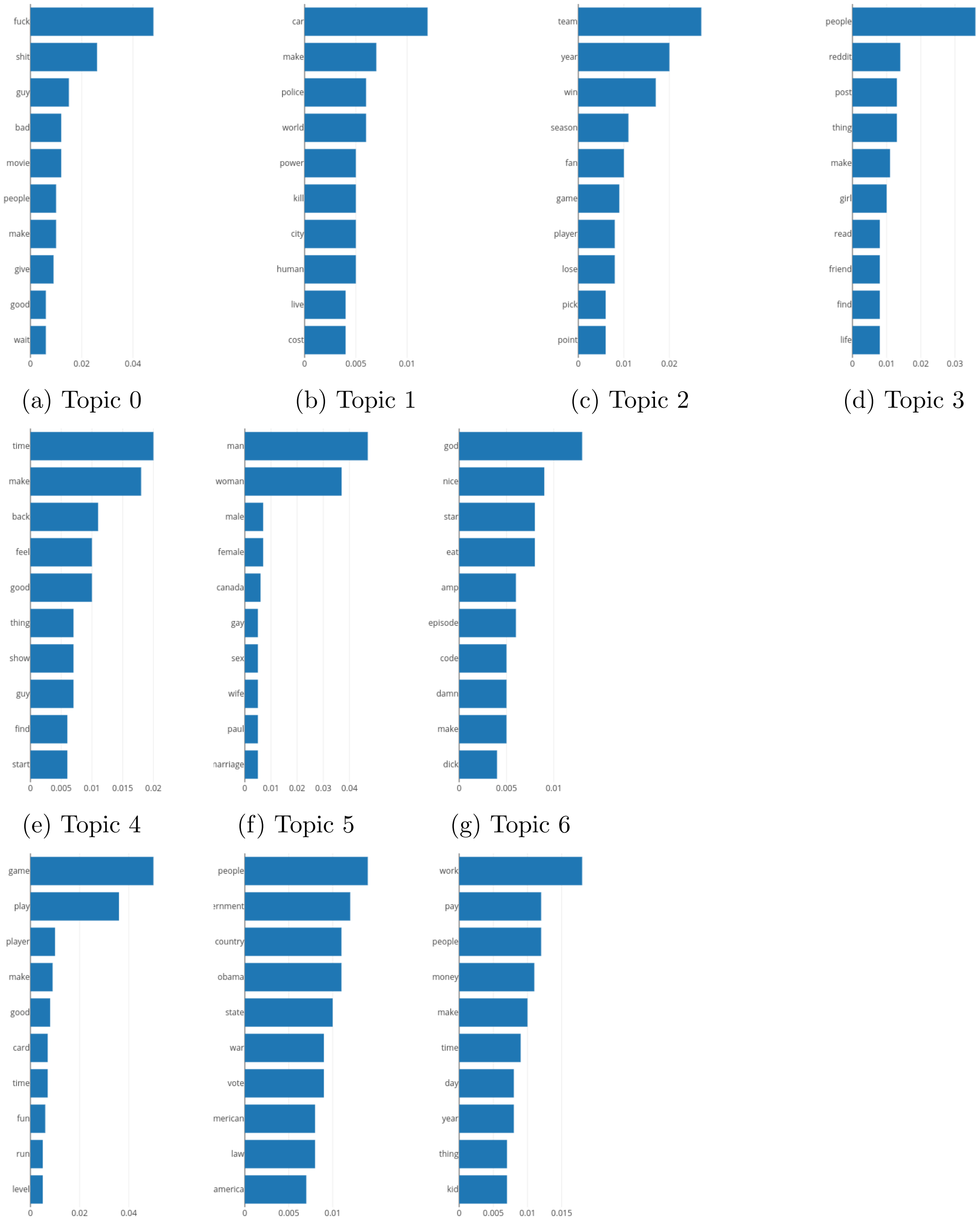


Figure 2.14: LDA C_V Coherence trained with parents

The extracted topics were labeled by hand, taking in consideration topic's most probable words (Figure 2.15), and the subreddits with the highest number of post of the specific topic.



(a) Topic 0 (b) Topic 1 (c) Topic 2 (d) Topic 3

(e) Topic 4 (f) Topic 5 (g) Topic 6

(h) Topic 7 (i) Topic 8 (j) Topic 9

Figure 2.15: Words distribution for each topic

In Table 2.2 are shown the extracted topics.

0	Movies	5	Human Rights
1	Crime/News	6	Mixed
2	Sport	7	PC/Gaming
3	Lifestyle	8	Politics
4	Entertainment	9	Technology

Table 2.2: Extracted topics

2.4.3 Topic Modeling on Reddit

Now that the model is trained, we label the documents and examine the distribution of comments with the new grouping. We can notice in Figure 2.16 that most of comments refer to the "Lifestyle" topic, followed by "Entertainment". Despite the position of the subreddit "Politics" (Figure 2.1), the topic "Politics" is only in sixth position.

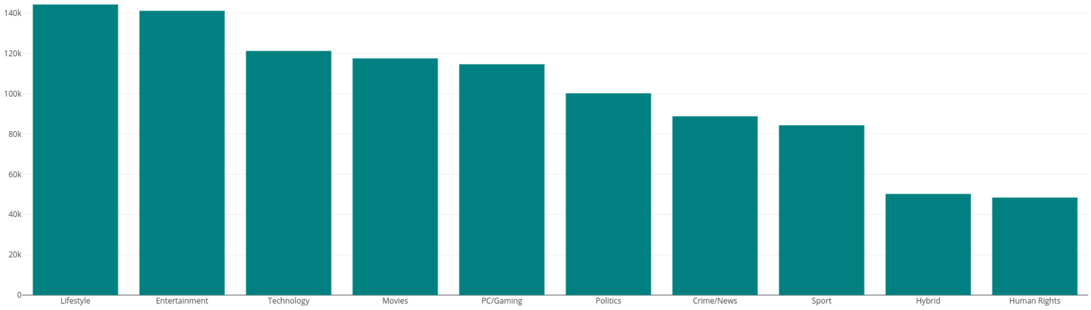


Figure 2.16: Number of comments for each topic

For what concerns the division in sarcastic and non sarcastic comments on the topics, we can see in Figure 2.17 that the topic "Politics" has the highest presence of sarcastic topics.

The next step consists in grouping the posts by subreddit and also using the extracted topic. This gives an insight of how the subreddits are closely-knit. In Figure 2.18 we can have a more interpretable read of the LDA performance. If we take a look at some topics, we will find them in large number in a subreddit that is labeled with e term similar or relevant to the topic. For example the topic

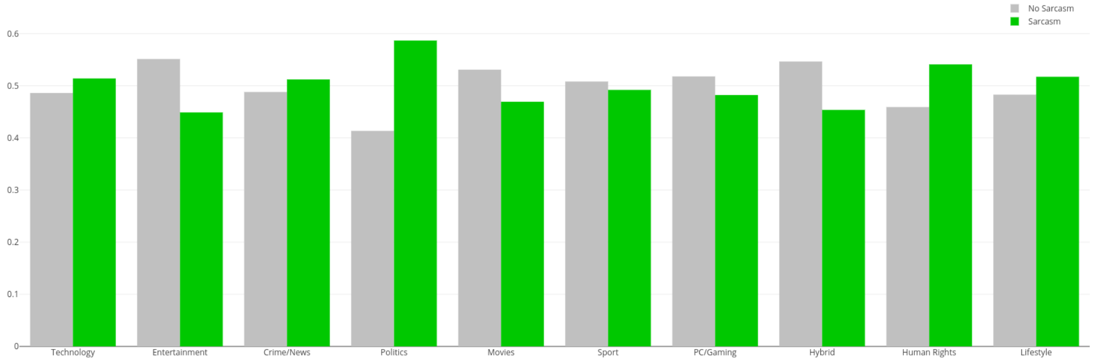


Figure 2.17: Number of sarcastic and non sarcastic comments for each topic

“Politics” is present in both “Politics” and “WorldNews” subreddits. Another example is the “Sports” topic is very common in the subreddit “NFL”, and this can be noticed also for other topics.

An interessant aspect emerged from grouping by topic and subbredit is that some subreddits are more sarcastic on the same topic. As shown in Figure 2.19, we can notice for “Politics” topic that the sarcasm percentage is near 50% in the subreddit “AskReddit”, but in “Politics” subreddit it’s around 61%, while in “WorldNews” it arrives at 67%.

For the last question, which are the topics that people tend to react more sarcastically? I found that some topics are more sarcastic than others, for example “Politics”, but also that the sarcasm depends on the community, in our case subreddits, in which the comment is posted.

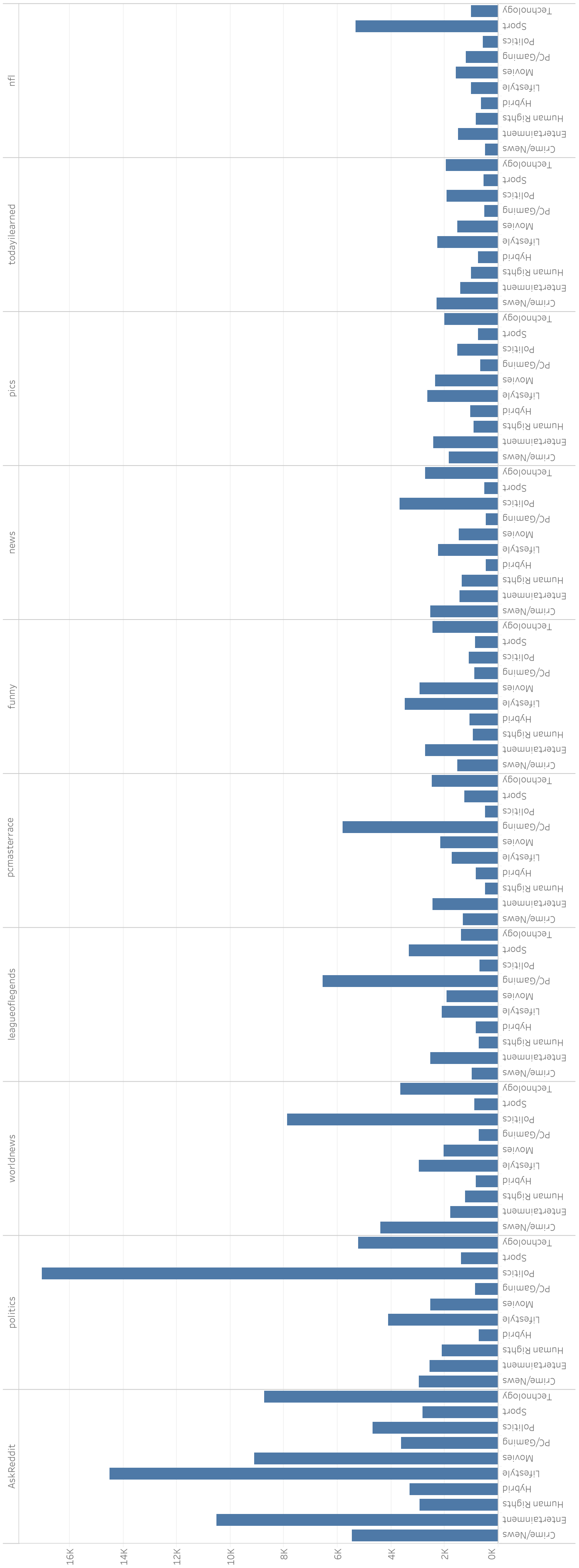


Figure 2.18: Topics for each subreddit

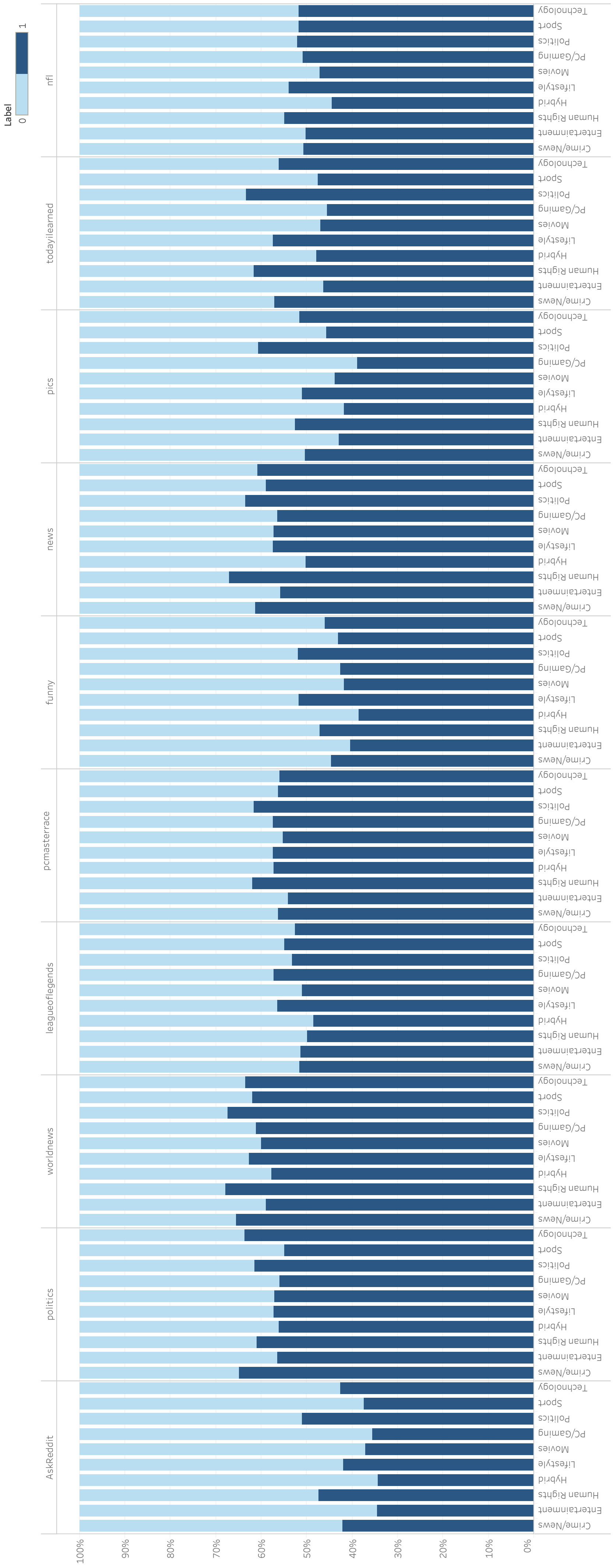


Figure 2.19: Topics for each subreddit with labels percentage

Bibliography

[1] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” 2014.

[2] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, ACM, 2015.