MY459 - Special Topics in Quantitative Analysis:

Quantitative Text Analysis

May 2022

Analysis of sarcasm on an internet commentary website

Candidate Number: 30499

Abstract

The project tries to answer what are the features that identify a sarcastic comment, based on the bag-of-words approach. The corpus in comments from a commentary website that are labelled on sarcasm and is balanced. The methodology followed starts with descriptive statistics on the document variables and some linguistic features of the corpus in relation to sarcasm. Sentiment and morality dictionaries were applied to examine the relationship of sarcasm with them. Naïve Bayes and Lasso Regression Supervised Learning models were used to classify the documents and identify good predictors of sarcasm. Dissimilarity metrics were used to examine the effect of a different text structure to sarcasm between the original comment and its answer. Finally, structural topic models and Latent Dirichlet Allocation were used, to uncover possible themes on the comments and explore their relationship with sarcastic ones. All these methods enlightened different characteristics of sarcasm.

1. Introduction

The corpus that is analysed is a Self-Annotated Reddit Corpus Dataset created by Khodak et al. in 2017. It contains labelled sarcastic and non-sarcastic comments from the Internet commentary website Reddit. Out of the 1.3 million records, a random balanced sample of 4,000 was chosen for this project. Each comment is treated as one document of the corpus. There are 9 more variables in the dataset, which are used as document variables, like the vote of the users for that comment and the 'parent-comment' to which the 'comment' was a response to.

The project tries to answer what are the features that identify a sarcastic comment, based on the bag-of-words approach. The methodology followed starts with **descriptive statistics** on the document variables and some linguistic features of the corpus in relation to the sarcastic labels. It is seen that Readability and Lexical Diversity are higher for sarcastic comments and that users tend to assign worst scores to sarcastic comments. However, length of a comment or the number of punctuations don't predict sarcasm.

Preparation of the document-feature-matrices was conducted to contain the features in the format that would be more meaningful for the analysis to follow.

A sentiment and a morality **dictionary** were applied to the corpus, to examine the relationship of sarcasm with sentiment and morality. The dictionaries were also applied on the parent-comments, so the effect of sentiment and morality of a parent-comment were examined on the sarcasm of a response. A relationship between sarcasm and negative sentiment or vicious morals, especially unfairness, is seen.

The **Supervised Learning** models, Naïve Bayes and Lasso Regression, were used to classify the documents into sarcastic and non-sarcastic and to estimate good predictors of sarcasm. Naïve Bayes model on monograms performed better than the others.

To examine whether comments that are different from their parent-comments in terms of their ‘bag-of-words’ features are more likely to be sarcastic, a **distance metric** was calculated between the document-feature-matrix of parent-comments and comments. The correlation found was not significant, although higher distant comments had slightly more sarcastic responses.

Finally, **topic modelling** was used, to uncover possible themes on the comments and explore their relationship with sarcasm. The optimal number of topics was estimated to 6 and then 3 models were fitted. A structural topic model with one topic prevalence covariate, a structural topic model with two topic prevalence covariates and one content and a Latent Dirichlet Allocation.

2. Motivation.

Sarcasm is defined as “a form of verbal irony that is intended to express contempt” (Joshi et. Al., 2017). In natural language understanding systems, like dialogue systems, detecting sarcasm is a significant aspect and sometimes the thorn of model building. Sarcasm detection is challenging because its occurrence is rare (Khodak et. Al., 2017), and hence different techniques are needed to balance the data, which contain some pitfalls. Also, sarcasm is even difficult for humans to detect and hence difficult for a third person to label a corpus (Wallace et al., 2014). The dataset in this project has been labeled from the authors of the comments themselves, so it’s considered a great opportunity for exploring sarcasm.

In the literature, sarcasm has been analyzed using Bag of Words methods, bigrams, naïve methods, deep neural networks (Persson et. al., 2018) and more.

This project tries to uncover some of the characteristics that sarcastic text entails, as it is a vital component of text analysis.

3.Description of corpus.

Each record in the column 'comment' will be treated as one document of the corpus. There are 9 more variables in the dataset, which are used as document variables and are explained in the table below.

Table 1: Data Dictionary of the corpus

Column	Description
Label	Binary sarcasm label
Author	Author name of the comment
Subreddit	The subreddit category under which the comment was made
Score	The comment score as voted on by users
Ups	The number of thumbs-up received by users
Downs	The number of thumbs-down received by users
Date	Date of the comment
Created_utc	Time of the comment
Comment	The comment that, which is used here as the document
Parent_comment	The original comment for which the 'comment' was a response to

The distribution of the numeric document variables is shown for both categories and with a coloured line indicating the median. There are a few outliers in both categories.

Figure 1: Distribution of the document variable 'Score' for the sarcastic and non-sarcastic comments.

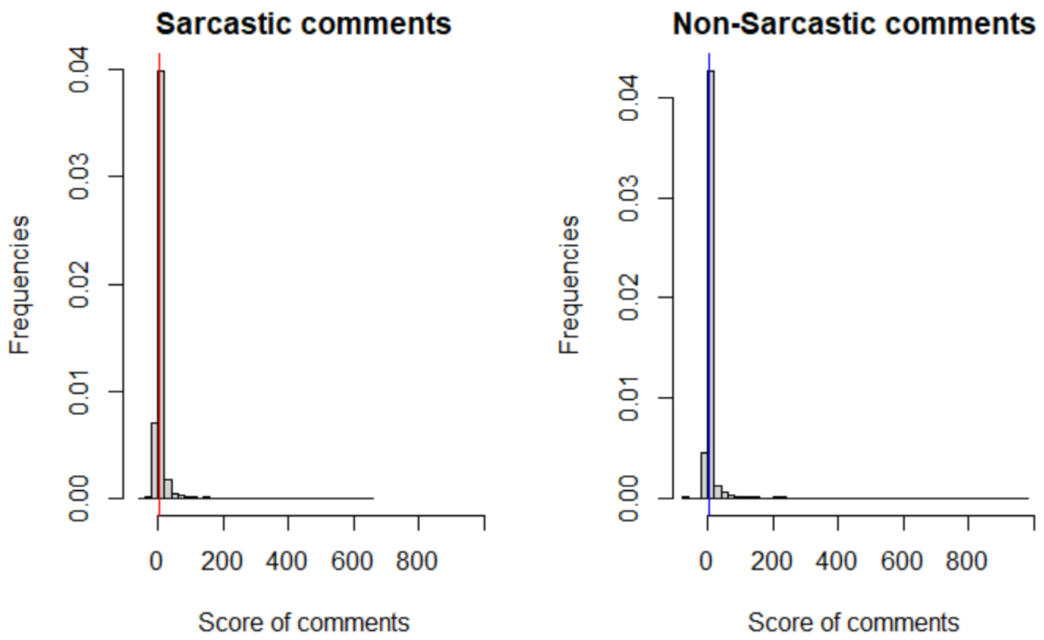


Figure 2: Distribution of the document variable 'Ups' for the sarcastic and non-sarcastic comments.

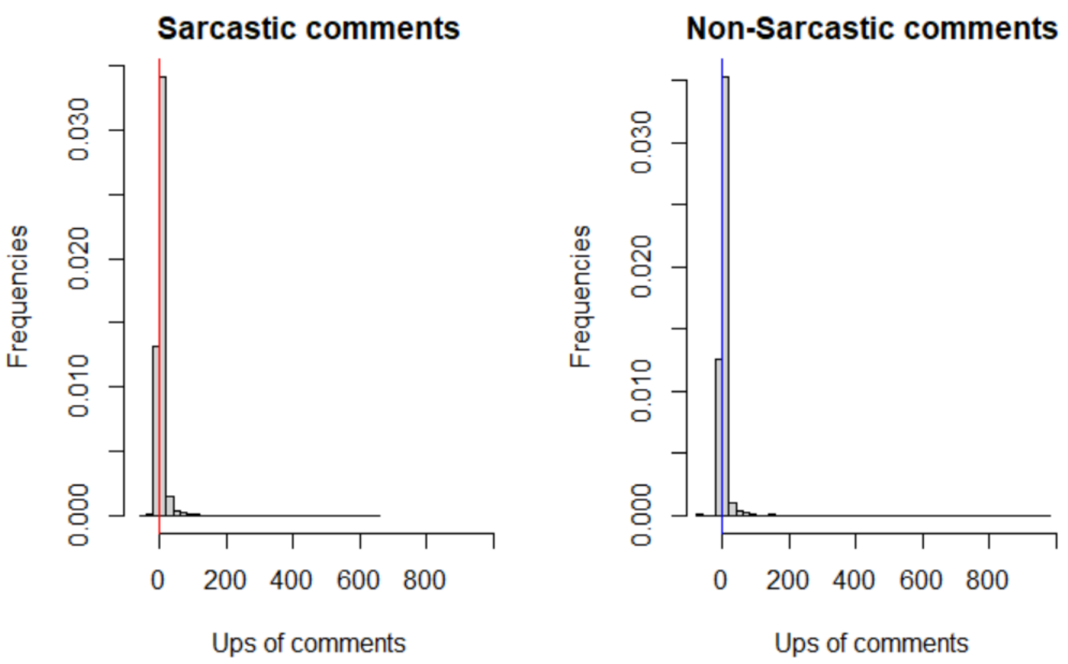
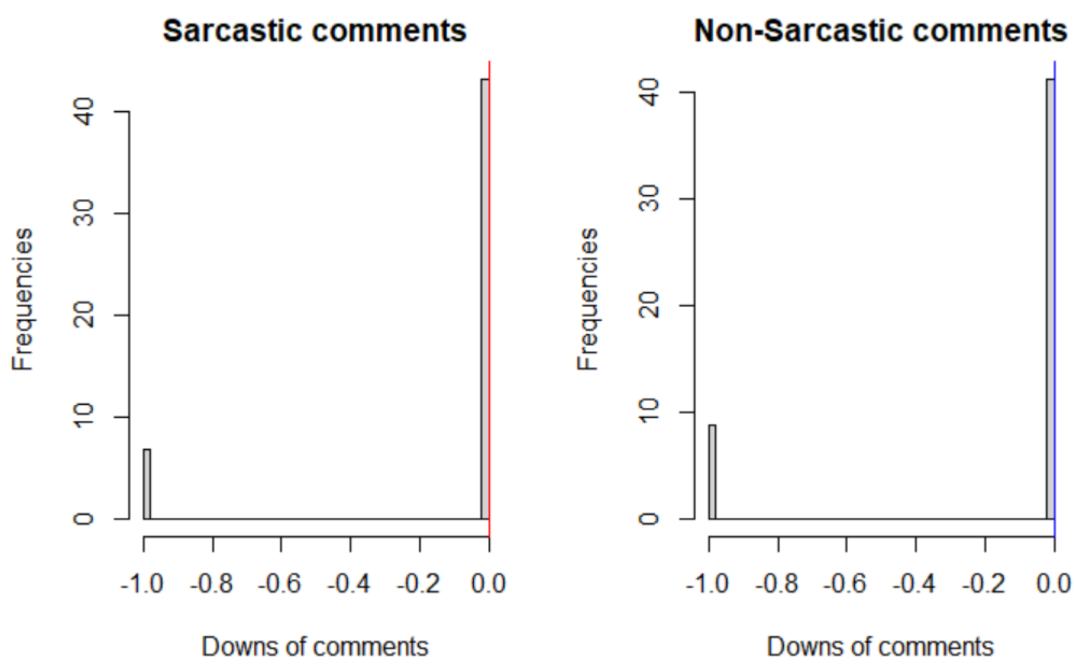


Figure 3: Distribution of the document variable 'Downs' for the sarcastic and non-sarcastic comments.



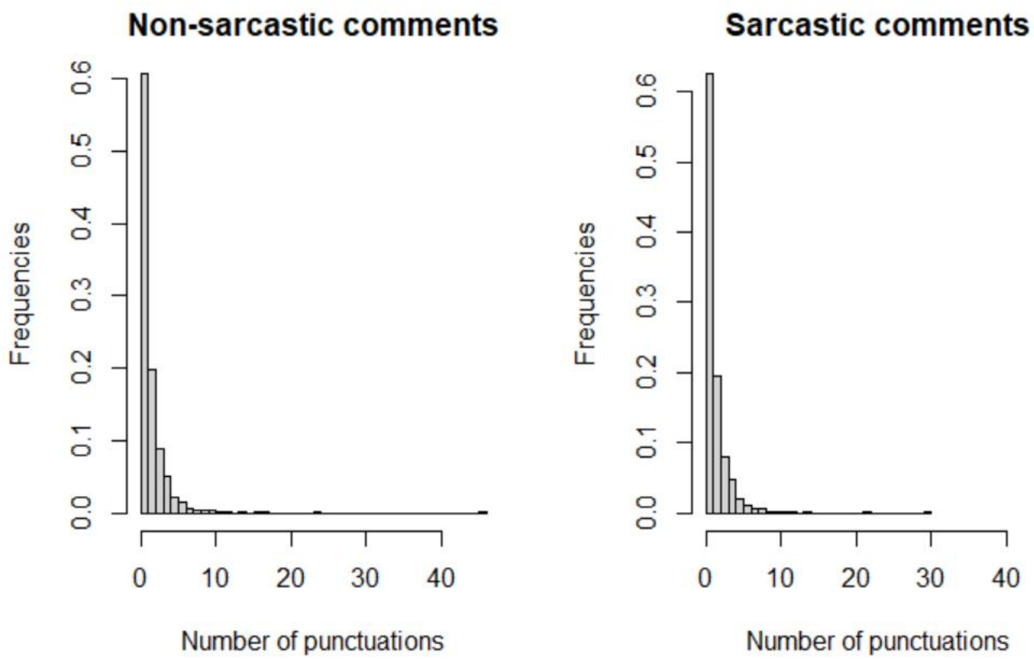
Each word is treated as a feature. In this analysis the words have been stemmed so that same stem words are treated equivalently. Therefore, the analysis focuses on the meaning of words and not on whether they are in a plural form, past tense etc. We should therefore note that stemming might be wrong in some cases. The document-feature-matrix has been created with document proportions instead of counts, to eliminate the effect of the length of a document. Also, words that appear in less than 3 documents have been removed, to reduce unnecessary complexity and avoid over-fitting of the models. Numbers, urls, symbols and stopwords have been removed, as we don't consider them good predictors of sarcasm.

Punctuation could be an indication of sarcasm, as many times 3 dots or exclamation marks indicate intentions. To explore the effect of punctuations, 100-word wordclouds of document-frequency-matrices with punctuation are shown for sarcastic and non-sarcastic comments. The same punctuations are coming up as the most frequent in both cases. The distribution of the number of punctuations for each case is drawn and is similar. However, there are a few non-sarcastic comments with extreme number of punctuations. Hence, there is no indication that punctuation should be used in our analysis.

Figure 4: Wordloud for non-sarcastic comments, when punctuations are used as features.

Figure 5: Wordcloud for sarcastic comments, when punctuations are used as features.

Figure 6: Distribution of the number of punctuations for the sarcastic and non-sarcastic comments.



After removing the punctuations, the word frequencies are shown in the wordclouds below. Common words like 'just' and 'like' are seen, but many differences are also observed, like the word 'yeah' that comes up very often only in sarcastic comments.

The distributions drawn do not uncover a particular relationship between the sarcastic comments and the metrics used. However, after performing chi-square tests for the independence of sarcastic comment to these metrics, it is seen that all p-values are below 5% and hence with a significance of 5%, there is evidence of a relationship.

Figure 7: Wordloud for non-sarcastic comments, when punctuation is removed.

Figure 8: Wordloud for sarcastic comments, when punctuation is removed.

Figure 9: Distribution of the number of tokens for the sarcastic and non-sarcastic comments.

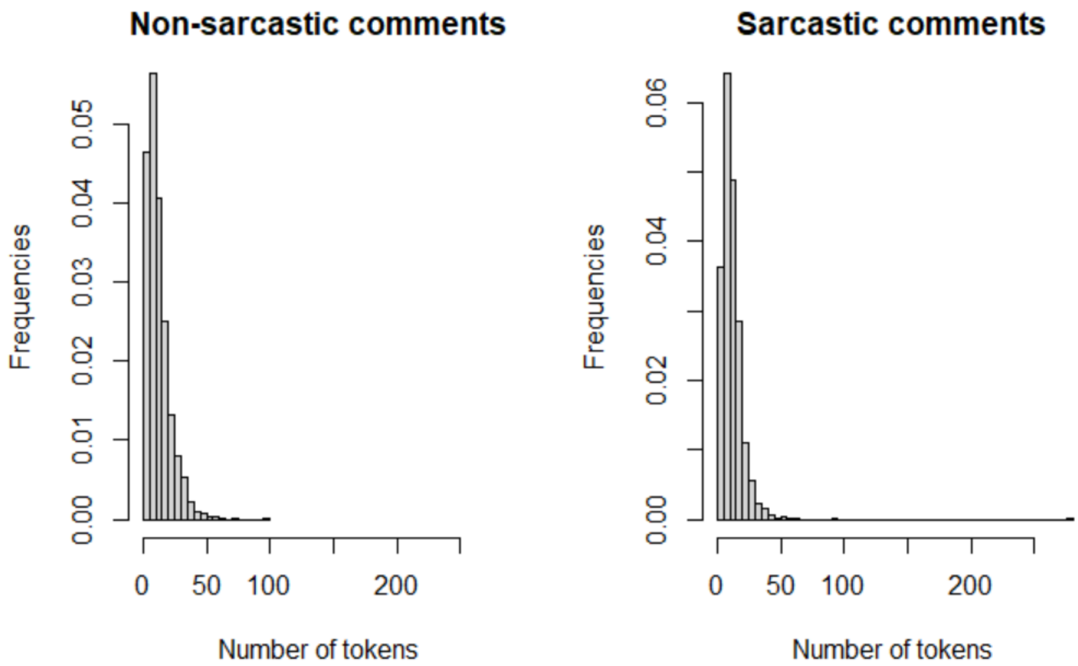


Table 2: Chi-square independence tests on sarcasm with different linguistic and document features

Null Hypothesis	Chi-squared	Degrees of Freedom	p-value
Independence between sarcasm and number of tokens	98.827	31	5.273e-09
Independence between sarcasm and number of punctuations	30.159	19	0.04981
Independence between sarcasm and score	198.54	133	0.0002001
Independence between sarcasm and ups	172.45	120	0.001218
Independence between sarcasm and downs	11.598	1	0.0006604

The **Readability** index based on Flesch-Kincaid metric is slightly higher for sarcastic comments (5.6) compared to non-sarcastic (5.4), meaning that overall sarcastic comments tend to use smaller sentences and smaller words.

The mean **Lexical Diversity** based on the Type-to-Token ratio is 98.6% for non-sarcastic comments and 99.2% for sarcastic. So, sarcastic comments use a bigger variety of words, but the difference is very small. This might be because the overall length of the comments is relatively small.

4.Methodology and Results

4.1 Dictionaries

Dictionaries will be used to explore whether sentiment and morality are correlated to sarcasm. To identify the sentiment of the comments, the LexiCoder dictionary is used, because it is built for media text and Reddit comments are more similar to that. For morality, the Moral Foundations Dictionary is used.

The approach followed was to group the documents to sarcastic and non-sarcastic and apply each dictionary to a count-unit document-frequency-matrix. The dfm is then transformed to proportions, so that the magnitude of the 2 documents is indifferent.

The output shows that 62% of the non-sarcastic comments are identified as positive, but only 57% are positive from the sarcastic ones.

Table 3: Distribution of sentiment across sarcastic and non-sarcastic comments

	Features	
Documents	**Negative**	**Positive**
Non-Sarcastic	0.3822064	0.6177936
Sarcastic	0.4326409	0.5673591

Morality is more often seen in non-sarcastic comments.

Table 4: Distribution of 2-level morality across sarcastic and non-sarcastic comments

	Features	
Documents	**Virtue**	**Vice**
Non-Sarcastic	0.5576923	0.4423077
Sarcastic	0.5022222	0.4977778

Looking into specific categories of morality, it is interesting to note that unfairness and authority are more related to sarcasm.

Table 5: Distribution of multi-level morality across sarcastic and non-sarcastic comments

	Features					
Documents	**Care-virtue**	**Care-vice**	**Fairness-virtue**	**Fairness-vice**	**Loyalty-virtue**	**Loyalty-vice**
Non-Sarcastic	0.2000000	0.1351351	0.06216216	0.03513514	0.06756757	0
Sarcastic	0.1600877	0.1469298	0.06578947	0.13157895	0.08114035	0

	Features			
Documents	**Authority-virtue**	**Authority-vice**	**Sanctity-virtue**	**Sanctity-vice**
Non-Sarcastic	0.1486486	0.002702703	0.08648649	0.2621622
Sarcastic	0.1140351	0.010964912	0.08771930	0.2017544

The same approach was applied to the '**parent-comment**' column, which was converted into a corpus and a document-frequency-matrix in the same way. As shown below, parent comments that are negative tend to receive more sarcastic comments.

Table 6: Distribution of the parent-comment sentiment across sarcastic and non-sarcastic response-comments

	Features	
Documents	**Negative**	**Positive**
Non-Sarcastic	0.4388245	0.5611755
Sarcastic	0.4694750	0.5305250

Looking into the morality of parent-comments it is seen that the more vicious a parent-comment is, the higher chances that the reply will be sarcastic. Also, the more authoritative parent-comments are, the more sarcastic comments they receive. A clear difference is also shown for the virtue of loyalty.

Table 7:Distribution of the parent-comment 2-level morality across sarcastic and non-sarcastic response-comments

	Features	
Documents	**Virtue**	**Vice**
Non-Sarcastic	0.5858283	0.4141717
Sarcastic	0.5495103	0.4504897

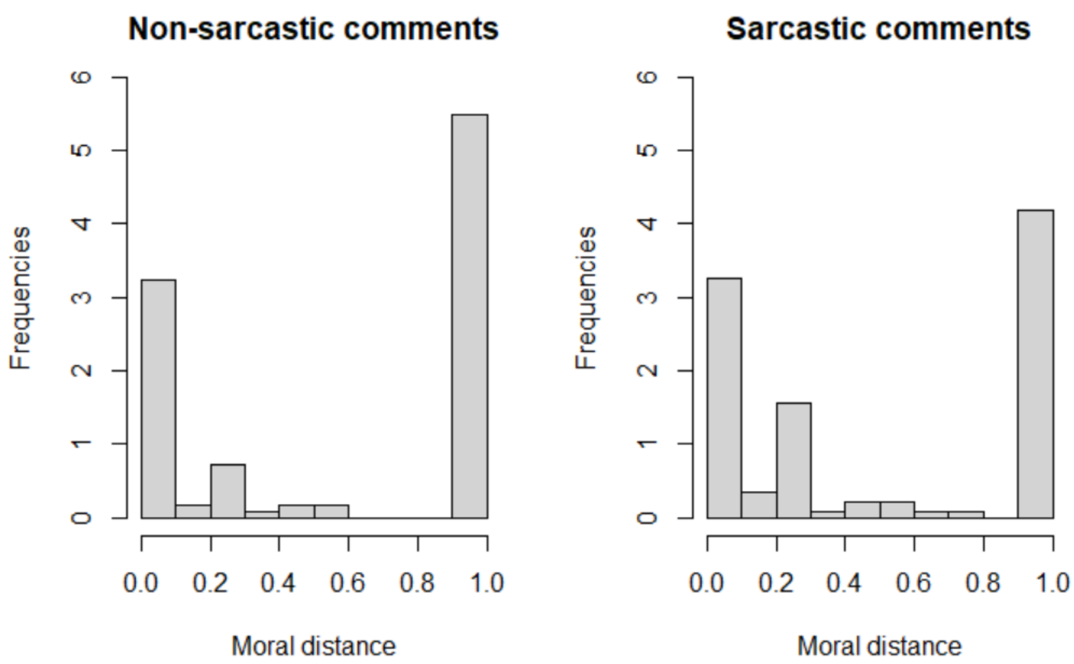
Table 8:Distribution of the parent-comments multi-level morality across sarcastic and non-sarcastic comments

Documents	Features					
	Care-virtue	Care-vice	Fairness-virtue	Fairness-vice	Loyalty-virtue	Loyalty-vice
Non-Sarcastic	0.1777996	0.1552063	0.07563851	0.04223969	0.08447937	0.001964637
Sarcastic	0.1604278	0.1497326	0.06203209	0.05668449	0.10909091	0.001069519

Documents	Features			
	Authority-virtue	Authority-vice	Sanctity-virtue	Sanctity-vice
Non-Sarcastic	0.1777996	0.1552063	0.07563851	0.04223969
Sarcastic	0.1604278	0.1497326	0.06203209	0.05668449

The last question to explore through the morality dictionary, is how distant in morality is the parent-comment and its comment and how is this related to sarcasm. The morality dictionary is applied to all comments and parent-comments. Then the cosine similarity between the parent-comment and comment is calculated and subtracted from 1, to get a morality distance measure. The figure below shows the distance distribution for each category. It is observed that non-sarcastic comments have a bigger morality distance with their parent-comment.

Figure 10: Distribution of moral dissimilarity between the parent-comments and the response-comments.



4.2 Supervised Learning

Supervised Learning is used to classify the documents into sarcastic and non-sarcastic, through a function that depicts the relationship between the sarcastic label and the features from a document-feature-matrix. The documents are therefore treated as bags-of-words. The data is split into training (80%) and test.

The first model to fit is a Naïve Bayes. Its confusion matrix and other performance metrics are seen below. The accuracy of the model is 60%, which means that 60% of the comments are correctly identified as sarcastic or non-sarcastic equivalently. Considering that the dataset is balanced and hence a random guess would return a 50% accuracy, the naïve bayes model adds some value. The precision is 61%, meaning that 61% of the comments that are predicted as sarcastic, are indeed sarcastic. Recall being 57% means that 57% of the sarcastic documents are predicted as such. F1 is a harmonic mean between precision and recall and hence will be the metric used to compare with other models.

Table 9: Confusion Matrix of the Naive Bayes Model with monograms.

		True Condition	
		Sarcastic	**Non-Sarcastic**
Prediction	**Sarcastic**	221	140
	Non-Sarcastic	165	243

Table 10: Performance metrics of the Naive Bayes Model with monograms

```
precision = 0.61
recall = 0.57
F1 score = 0.59
accuracy = 0.6
```

The posterior class probabilities are extracted for a few words that in the worcloud came up as frequent ones in one of the 2 categories. It is seen that the results align with what was seen in the worcloud, as the words 'forgot' and 'yeah' appeared in the sarcastic comments and the others in the non-sarcastic.

Table 11: Posterior class probabilities for 4 features.

Classes	forgot	yeah	good	thank
Non-sarcastic	0.0004174469	0.0009755196	0.001730241	0.002254994
Sarcastic	0.0028600951	0.0032643127	0.001354557	0.001272559

A Naïve Bayes model is also built for features being both monograms and bigrams. The F1 score and accuracy are the same. Precision has increased by 2%, since the true positives have increased from 221 to 226 and the false positives decreased by 9 units. Of course, recall is impacted. So, if we are more tolerant about wrongly identifying sarcastic comments compared to missing out sarcastic comments, we will choose the first model with the monograms.

Table 12: Confusion Matrix of the Naive Bayes Model with bigrams.

		True Condition	
		Sarcastic	Non-Sarcastic
Prediction	Sarcastic	226	131
	Non-Sarcastic	180	264

Table 13: Performance metrics of the Naive Bayes Model with bigrams.

```
precision = 0.63
recall = 0.56
F1 score = 0.59
accuracy = 0.61
```

Lasso regression is also used for bigrams. A penalty parameter (λ) is used to account for model complexity. Therefore, Lasso performs a feature selection process, as the l1 norm is minimized when many parameter coefficients are set to zero. A 5-fold cross validation process is used to estimate the best value of λ, which is 12. As it can be seen below, out of the 1734 features, only 49 were selected as significant ones.

Lasso performs worse compared to naïve bayes models, based on all metrics except precision. Its precision is 80%, meaning that even though the sarcastic predictions are few, 80% of them are indeed sarcastic.

The effect of the most important features based on Lasso is shown. Words like ‘yeah’, ‘rape’ and ‘hard’ come up as the best predictors of a sarcastic comment.

Figure 11: Lasso Regression model estimation for different values of the complexity penalty (λ)

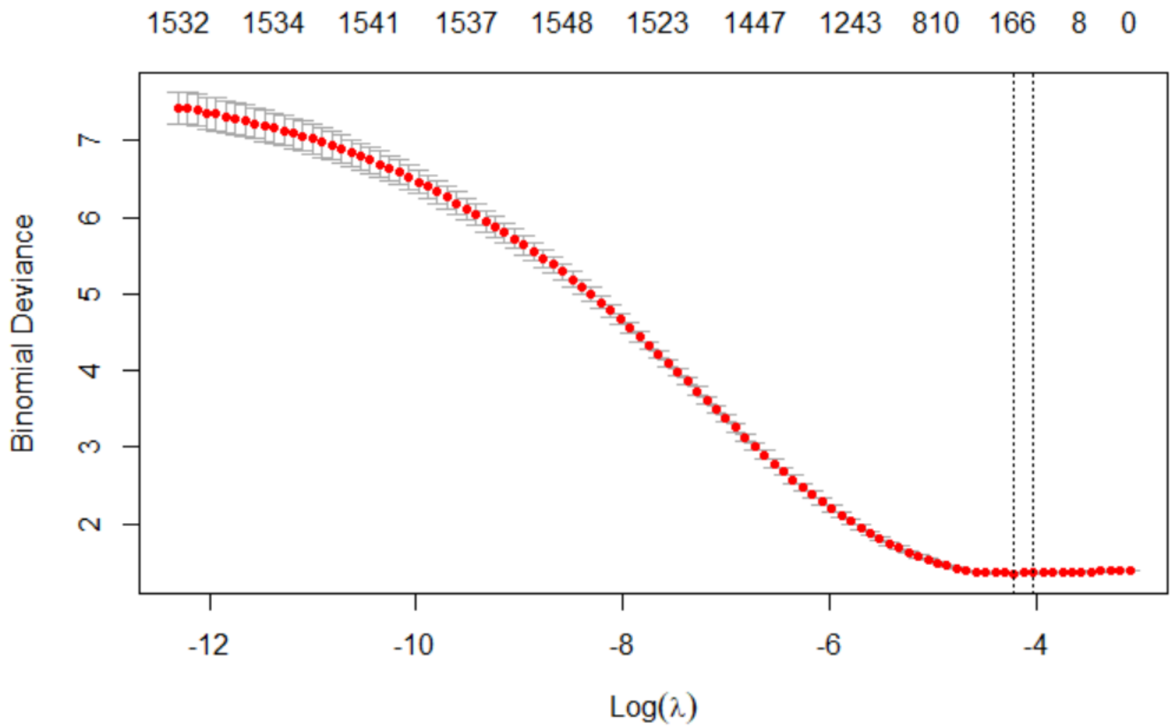


Table 14: Confusion Matrix of the Lasso Regression Model with bigrams.

		True Condition	
		Sarcastic	**Non-Sarcastic**
Prediction	**Sarcastic**	85	21
	Non-Sarcastic	321	374

Table 15: Performance metrics of the Lasso Regression Model with bigrams.

```
precision = 0.8
recall = 0.21
F1 score = 0.33
accuracy = 0.57
```

Table 16: Best predictors for non-sarcasm based on Lasso Regression model

Coefficient	Feature
-2.4950357	week
-1.4043039	entir
-1.2049988	also
-0.6435352	friend
-0.4950552	lol
-0.4295569	awesom
-0.3950272	health
-0.3347931	next
-0.2739579	leav
-0.2257979	look_like

Table 17: Best predictors for sarcasm based on Lasso Regression model

Coefficient	Feature
1.457071	obvious
1.519080	clear
1.593638	total
1.667373	racist
1.739027	caus
1.777649	just_like
1.948884	fault
2.415237	yeah
2.604776	rape
3.581333	hard

The word 'yeah' that came up as one of the most significant ones, is searched it in its context and 5 examples are shown.

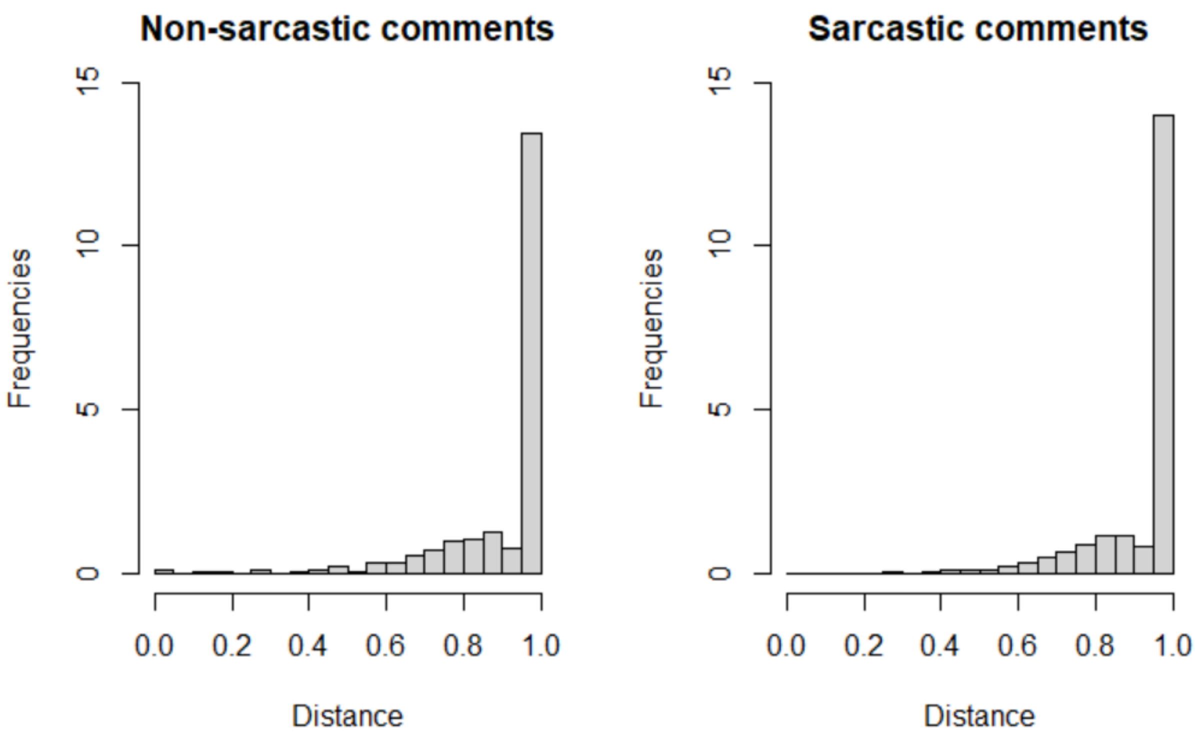
Table 18: 'Yeah' word in context on a 10-words window, for 5 documents.

Document	Word in context
1	**Yeah**, now the NSA can't steal my nudes.
2	**Yeah**, I mean, it's not like the author of
3	**Yeah**, i can't wait for the start behind safety car
4	Well **yeah**, one team is getting an Olympian and the other
5	**Yeah** riots bugsplats are certainly all our faults. **Yeah**_riots riots_bugsplats

4.3 Similarity

To answer whether comments that are different from their parent-comments in terms of their 'bag-of-words' are less likely to be sarcastic, the distance between the two is calculated. There are not many differences in the distribution of the distance, although a slightly higher distance is noted for the sarcastic comments.

Figure 12: Distribution of frequency of words dissimilarity between the parent-comments and the response-comments.



4.4 Topic Modelling

Topic modelling is used here, to uncover possible themes on the comments and explore their relationship with sarcasm. First, the optimal number of topics (k) is estimated. Then a structural topic model is fitted with the binary 'label' variable as the topic prevalence covariate. Then, a structural topic model is fitted with 'label' and a smooth version of score as topic prevalence covariates and 'label' as a content covariate. Finally, an LDA model is fitted.

Select optimal K

To select the optimal number of topics, the Spectral algorithm was used. It computed diagnostic values for models with different number of topics (K).

As it can be seen below the Held-Out Likelihood metric is highest for k=4 and very similar to 6 and 8, hence the suggested number of topics based on this metric is between 4 to 8. Therefore, for 4 number of topics the new unseen data is more possible to appear by the Structural Topic Model, compared to other number of topics.

The Residuals, have a linear decreasing pattern as the topics increase, especially after k=10. Hence, the best k is 12.

Semantic coherence is maximized when the most probable words in a topic often occur together, therefore, it is like the patterns that a human would understand (Mimno, 2011). The key is to find the trade-off with semantic coherence and exclusivity. The figure below shows high coherence for k=4 and 6. The exclusivity though drops significantly from k=8 to 6 and then to 4, so based on these metrics we can conclude that the optimal k is 4.

Each metric has suggested a slightly different number. We will continue the analysis by considering as the optimal k, the average of the above, which is 6.

Figure 13: Diagnostic values for different number of topics

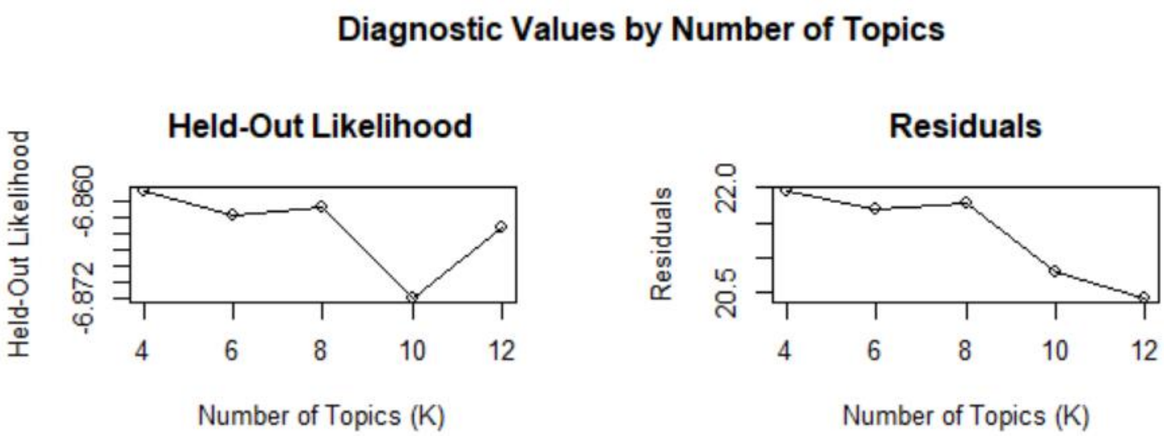
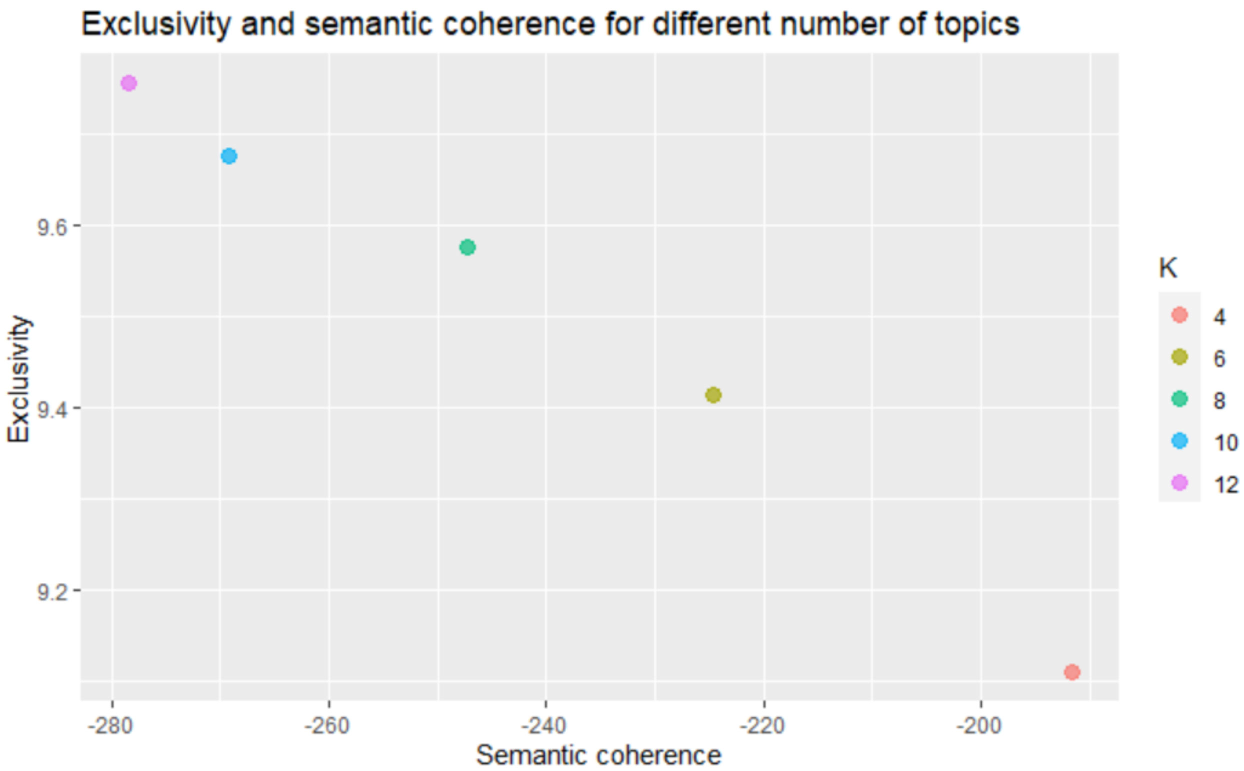


Figure 4: Exclusivity and Semantic coherence diagnostic metrics for different number of topics

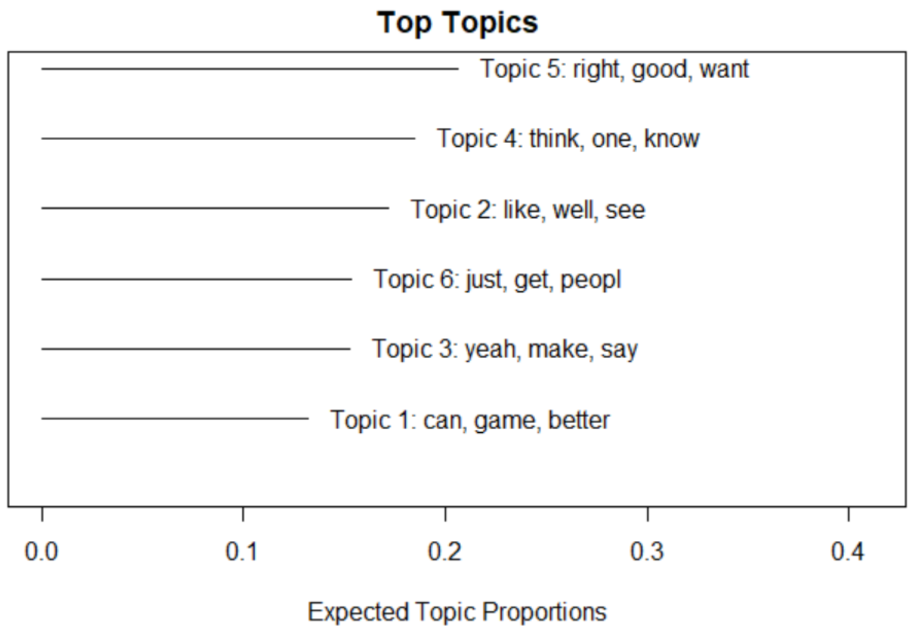


Structural topic model with topic prevalence covariate

First the `quanteda` document-frequency-matrix object is converted into an input for the structural topic model. All the documents that have zero features left are removed.

The figure below shows the most frequent words that appear in each of the six topics.

Figure 15: Most frequent words across different topics for the Structural topic model with topic prevalence covariate.



The WordClouds below, show a different structure of the words for each topic. For example, the word 'yeah' is the most common for topic 3 and it can not be seen in the top 100 words for any of the other topics. The distribution of frequencies varies for each topic. For example, topic 6 has very few words as the most frequent and all the other have a similar small frequency. On the contrary, topic 5 has many more words in high frequencies. This means that in topic 5 more words are repeated often, whereas in topic 6 there is a more even use of words.

Figure 16: Word Clouds on the 6 topics generated by the Structural topic model with topic prevalence covariate.

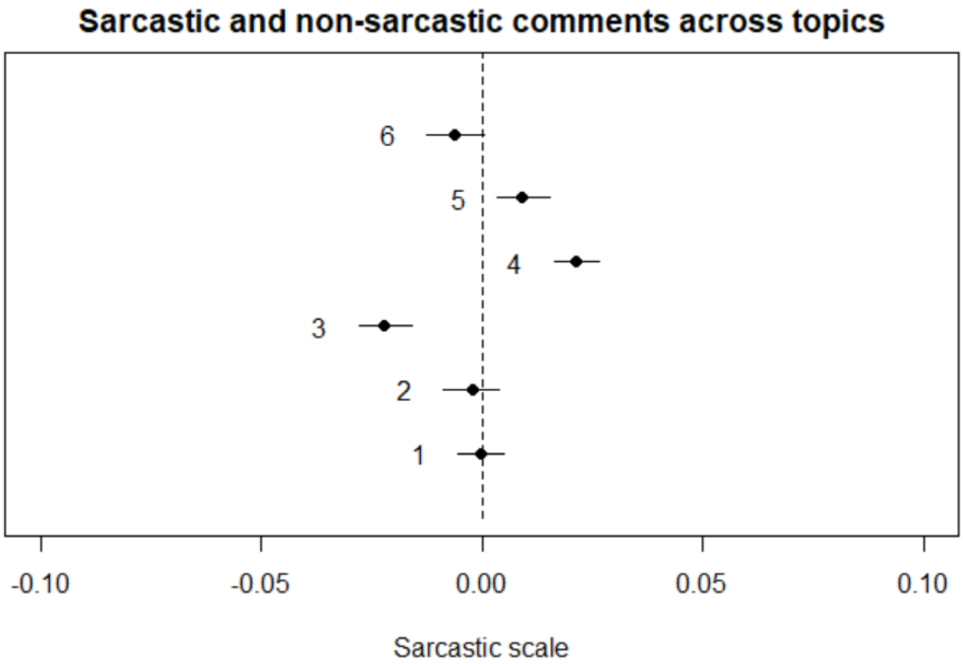
To get a better understanding of the topics, the two documents with the highest shares for topics 5 and 6 are shown. It is observed that some of the most frequent words are used accordingly, as shown in green.

Table 19: Two most representative documents for topics 5 and 6 from the Structural topic model with topic prevalence covariate

Topics	Most representative documents
5	"Good thing the Republicans are for following the law of the land, which is the Constitution and which specifically states about freedom of religion."
	"Yes, because army medics lead to a false sense of security, which leads to soldiers allowing themselves to get shot."
6	"While I really appreciate the offer, I'm just fine with what I have now, thanks :p Just keep practising, if the downvotes don't bother you each time you post you'll often get sound criticism from some people who can guide you in the right direction."
	"At least the same peep who created PPRNG has also managed to create this SAV dumper tool, for a more legit shiny egg :D"

The ‘label’ prevalence covariate indicates which topics are discussed more by which of the two types of comments. The figure below, shows that comments under topic 4 tend to be more sarcastic.

Figure 17: Sarcastic Scale for different topics as indicated by the Structural topic model with topic prevalence covariate.

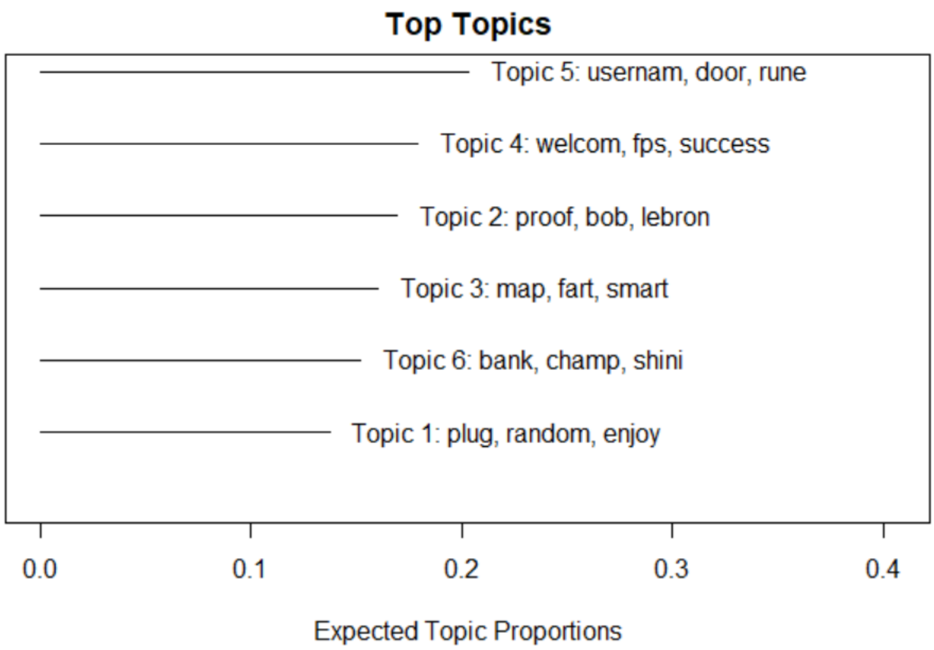


Structural topic model with topic prevalence and content covariates

A structural topic model was fitted with ‘label’ and a smooth version of score as topic prevalence covariates and ‘label’ as a content covariate. Therefore, this model takes into account that topic proportions within documents can vary based on the ‘label’ and ‘score’ and that word proportions within topics can vary based on the ‘label’.

There are no common words in the top frequencies with the previous model.

Figure 18: Most frequent words across different topics for the Structural topic model with topic prevalence and content covariate.



Latent Dirichlet Allocation (LDA) model

An LDA model, which is generative probabilistic (Blei et. Al., 2003), was also fitted. The words shown are the most frequent ones for each topic.

Table 20: Most frequent words for each topic based on LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
"see"	"look"	"peopl"	"get"	"just"	"like"
"know"	"think"	"just"	"peopl"	"like"	"well"
"guy"	"fuck"	"yeah"	"yeah"	"can"	"just"
"well"	"thing"	"can"	"time"	"better"	"right"
"want"	"just"	"one"	"just"	"know"	"go"
"need"	"good"	"good"	"make"	"game"	"can"
"just"	"way"	"game"	"even"	"everi"	"man"
"like"	"forgot"	"right"	"one"	"got"	"think"
"better"	"time"	"like"	"fuck"	"now"	"say"
"yeah"	"go"	"see"	"go"	"want"	"game"
"come"	"now"	"work"	"think"	"use"	"now"
"use"	"thank"	"make"	"thing"	"much"	"one"
"level"	"come"	"thank"	"better"	"make"	"oh"
"mean"	"yeah"	"realli"	"less"	"good"	"sure"
"year"	"call"	"say"	"well"	"yeah"	"want"
"say"	"money"	"mean"	"happen"	"yes"	"much"
"realli"	"still"	"someon"	"right"	"new"	"year"
"look"	"yes"	"keep"	"yes"	"one"	"wrong"
"alway"	"make"	"year"	"actual"	"ever"	"realli"
"never"	"lol"	"reason"	"everyon"	"go"	"pretti"

The topic with the highest proportion for each of the comments was obtained. For topic 1 and 2, 5 comments that are predicted to contain that topic in highest proportion were randomly sampled. It is shown below that their semantic content reflects the expected topic.

Table 21: Five sarcastic comments that are predicted to contain topic 1 in highest proportion.

```
 Topic 1, sarcastic comments:
[1] "Needs more james"
[2] "For the children"
[3] "What an appropriate comment in the appropriate thread."
[4] "You have really deep thoughts."
[5] "by proper disposal they mean reuse the same pill."
```

Table 22: Five non-sarcastic comments that are predicted to contain topic 1 in highest proportion.

```
 Topic 1, non-sarcastic comments:
[1] "Clue: the problem lies in the dysfunctional double-think of children being
presented as sexual yet not being allowed to be perceived as such."
[2] "People always complain that NA needs to be better towards new and young talent, and
now that a big org does it, people mean they should have picked a more experienced
player."
[3] "you have 30 boyfriends..... i see...."
[4] "No, it's the one that goes \"judge not lest ye be judged,\" which is an excellent
message coming from the part of the bible Christians are supposed to prioritize."
[5] "I know it probably won't happen, but I want more of them together season 2"
```

Table 23:Five sarcastic comments that are predicted to contain topic 2 in highest proportion.

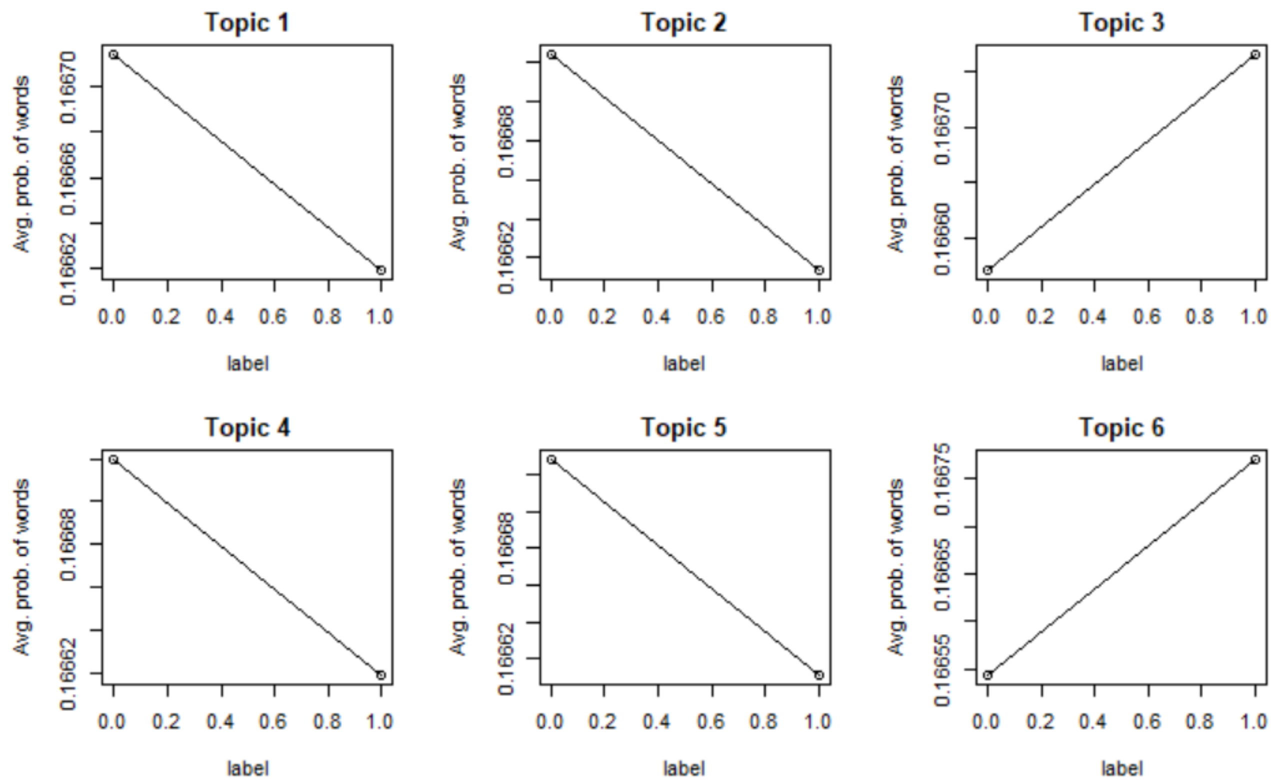
```
 Topic 2, sarcastic comments:
[1] "Look, everyone knows that interns are human garbage and they should be glad to be
paid *at all*."
[2] "But being fat is genetic and being gay is a lifestyle choice."
[3] "At last we've found our team identity post Roy-Oden!"
[4] "Nah, its not a smart idea."
[5] "Put a police officer at the finish line and watch him win"
```

Table 24: Five non-sarcastic comments that are predicted to contain topic 2 in highest proportion.

```
 Topic 2, non-sarcastic comments:
[1] "You're gold and playing in groups."
[2] "Kinda like when they showed DMX a computer."
[3] "Throw in Exodia, you'll draw all 5 cards eventually."
[4] "Some say that our plasma even disperses naturally in vacuum or atmosphere!"
[5] "And also reminds the teacher that she forgot to collect everyones homework from
yesterday right before class is over...God dammit, bitch, i forgot to do it and could
have used that extra night!"
```

The estimated proportion of words on sarcastic and non-sarcastic comments on each topic is shown. It was calculated using the theta posterior probabilities from the LDA. Topics 3 and 6 use more often the same words for sarcastic comments, opposite to all other topics.

Figure 19: The estimated proportion of words on sarcastic and non-sarcastic comments on each topic, as estimated by the Structural topic model with topic prevalence and content covariate.



5.Conclusions.

After this analysis on sarcasm, we can conclude that sarcastic comments tend to be in a language that is more simplified, with more slang words and not so much reasoning. Sarcasm is also related to negative sentiments and immorality.

Things to note for future projects is the use of emojis that could be enlightening on the subject and the use of other algorithms, like deep learning and word embeddings.

6. References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Joshi, Aditya, Pushpak Bhattacharyya, and Mark J. Carman. "Automatic sarcasm detection: A survey." *ACM Computing Surveys (CSUR)* 50, no. 5 (2017): 1-22.

Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli. "A large self-annotated corpus for sarcasm." *arXiv preprint arXiv:1704.05579* (2017).

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. "Optimizing semantic coherence in topic models." In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262-272. 2011.

Persson, Ludvig, and Jesper Larsson. "Sarcasm Detection with TensorFlow." (2018).

Wallace, Byron C., Laura Kertz, and Eugene Charniak. "Humans require context to infer ironic intent (so computers probably do, too)." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 512-516. 2014.