Federico Matteucci - 955753

Sentiment Analysis Project #8: Controversial Comments

Introduction - What makes a comment controversial?

Scope of this analysis is to identify, characterize and predict controversiality in comments to newspaper articles.

The literature in this sense identified some factors that drive *controversy*: *personalization*, *uncertainty* and *unexpectedness* more than others. [2] Instead of replicating these results, I explored the vocabulary of the comments to build a predictor for controversiality.

Research question and metodology

Random variables

Here I define the random variables used in the paper:

- $w = 1 \iff$ word w appears in the text of the comment
- $HCC = 1 \iff$ the comment is considered controversial
- $cat = C \iff$ the article to which the comment is attached falls in category C
- $kw = 1 \iff$ the article to which the comment is attached is labelled with keyword kw

Goal assessment, definition of controversiality

Goal of this report is to identify controversial comments (named `HCCs` : **Highly Controversial Comments**; in opposition to `NCs` : **Non Controversial comments**), describe some of their peculiarities and assess whether the vocabulary of a comment is enough to predict the success the comment will have. The first issue then is to assess what makes a comment controversial.

As proxies for the success (controversiality) of a comment, I had available the number of replies and recommendations a comments received. I wanted to obtain a single measure for controversiality (hereafter denoted as `CVI`, **ControVersiality Index**). A sensible approach is to rescale the two attributes and then to compute their harmonic average. Why harmonic? To give higher scores only to those comments which are both recommended and which sparked a long discussion. As harmonic mean requires non-negative inputs, the data was rescaled in standard deviations from the minimum (instead of the average).

Throughout this analysis, I decided to focus on user-user interaction, ignoring the `editorsSelection` attribute.

I finally defined an `HCC` to be a comment whose `CVI` is above the 99-th percentile (of the `CVI` distribution).

Controversial comments

I focused on the vocabulary and length of comments to study whether there are significative differences between `HCCs` and `NCs`.

As for the vocabulary, I tokenized the corpus in unigrams (as stated in [1], unigram tokenization often works well in terms of sentiment analysis), computed for each word w the probability of observing that word in the `HCC` and `NC` corpuses, $\mathbb{P}(w = 1 | HCC = i), i = 0, 1$, and eventually the **Kullback Leibler divergence** of the two probabilites, interpreted as the controversiality score of a word. The top results are the words that are significantly more common in `HCCs` than they are in `NCs`: such words can be

considered indicators of controversiality (even if, as shown below, they are not enough to predict controversiality).

I then proceeded to compare the average length of `HCCs` ad `NCs` using a t-test, under the reasonable assumptions that the length of any comment is independent from that of other comments and that the sample is large enough to overcome the normality requirement. [6]

Controversial topics and categories

I propose four ways to investigate the controversiality of topics and categories.

1. Identify **controversial keywords** (the articles are labelled with keywords extracted from the text). Computing the divergence of $\mathbb{P}(kw = 1|HCC = 1)$ and $\mathbb{P}(kw = 1|HCC = 0)$ for each keyword kw allows to score and rank keywords according to their controversiality.
2. Compute the **divergence** of $\mathbb{P}(cat = C|HCC = 1)$ and $\mathbb{P}(cat = C|HCC = 0)$ for each category C. The rationale is that a category is more controversial if it is more likely to be observed associated to `HCCs` than `NCs`.
3. Calculate the **probability of a category generating a controversial comment**. From Bayes Theorem: $\mathbb{P}(HCC = 1|cat = C) = \mathbb{P}(cat = C|HCC = 1) \cdot \mathbb{P}(HCC = 1) \cdot \frac{1}{\mathbb{P}(cat=C)}$. Note that all of the quantities on the RHS can be easily estimated with counts.
4. Rank categories based on the average `CVI` score of their comments.

Predicting controversiality

Is it possible with vocabulary alone to determine whether a comment will be controversial? To answer this question I trained a tree predictor for binary classification on the corpus using as target `HCC`.

Experimental Results

Dataset

The dataset of choice is one of the New York Times comments dataset, consisting of all the comments and articles of April 2017. [4]

After deleting some unnecessary attributes and casting the columns to appropriate types, each entry of the dataset is indexed over the ID of a comment (attribute `commentID` of the original dataset) and has the attributes below:

Column name	Description
`articleID`	Unique identifier of the article the comments belongs to
`commentBody`	Text of the comment
`snippet`	Snippet of article `articleID`
`keywords`	Keywords associated to article `articleID`
`recommendations`	Number of users' upvotes to the comment
`replyCount`	Number of replies to the comment
`editorsSelection`	1 if the comment is considered relevant by the editor
`newDesk`	Category of article `articleID`

Each column has 243830 non-null entries. As for what concerns the grammatical correctness of the texts, comments go under a process of scrutiny and grammatical correction when submitted to the NYT before they are posted. [5]

Characteristics of `HCCs`

Vocabulary

In order to study the vocabulary, the text of comments was converted to **lower case**, **lemmatized** (using Wordnet) and eventually **tokenized** in unigrams. This allows to efficiently extract the most out of a limited corpus, at the cost of some information loss (capitalization, part of speech, word dependencies, overall structure ...) common to all bag of words approaches.

I then computed the **relative frequencies** of a word w, obtaining estimates for $\mathbb{P}(w = 1|HCC = i), i = 0, 1$. Comparing these two probabilities can yield interesting results as it allows to find words that are more common in controversial comments than in unsuccesful ones. The top results of the comparison with Kullback Leibler divergence are shown in the table below.

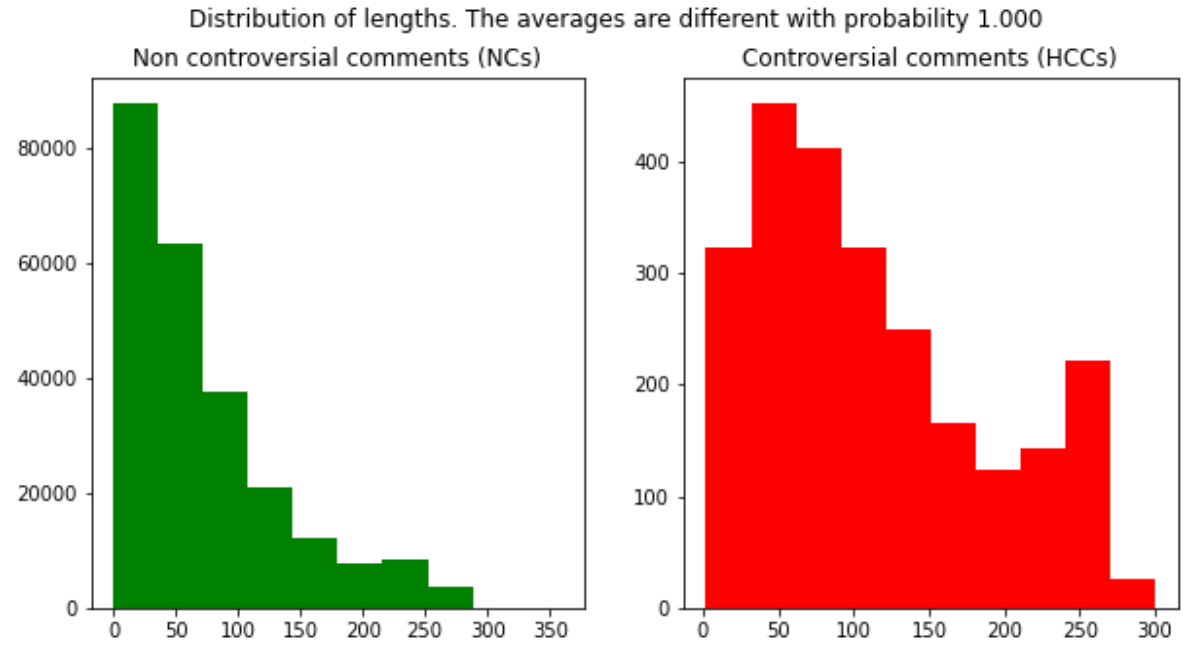
Most controversial words	**KL - divergence**
trump	0.00180
president	0.00087
warn	0.00077
republican	0.00064
american	0.00053
tax	0.00048
donald	0.00044
voter	0.00039
country	0.00032
health	0.00032

As one might have expected, the most controversial words are all related to politics, taxes and debating points. Healthcare, for instance, was a hot topic as in these year the *Patient Protection and Affordable Care Act* (also known as *Obamacare*) was under scrutiny.

Length

Is there any discrepancy in the lengths of controversial and non controversial comments? The t-test run on the two distributions of lengths shows that the average of the two distributions is the same with a probability in the order of $1e - 221, 0$ for all intents ant purposes.

This is also visually confirmed by the histograms below, as the `HCCs` have a much more probability mass in the central and right-most regions.



Controversial topics

I here examine ways of identifying controversial topics.

Keywords

The results of the keyword frequency divergence analysis:

Most controversial keywords	**KL - divergence**
United States Politics and Government	0.01918
Trump, Donald J	0.014467
Republican Party	0.005389
Presidential Election of 2016	0.00390
Federal Taxes (US)	0.00365
Income Tax	0.00356
Patient Protection and Affordable Care Act (2010)	0.00275
Ryan, Paul D Jr	0.00272
Senate	0.00267
Priebus, Reince R	0.00259

Again, the most controversial topics are related to either politics or debating points.

Divergence

The category frequency divergence analysis yields this result:

Category	KL - divergence
OpEd	0.07119
National	0.05011
Editorial	0.02287

All of the other categories have negative divergence, meaning that they are actually more commonly associated to `NCs` than to `HCCs`.

Bayes

The first three results of the Bayes approach:

Category	$\mathbb{P}(HCC = 1 \vert cat = C)$
National	0.01286
Editorial	0.01206
OpEd	0.01191

`CVI`

The three most controversial categoris according to average `CVI`:

Category	Average `CVI`
Styles	0.13726
Business	0.11094
National	0.11088

Predicting controversiality

To evaluate the performance of the tree predictor the natural choice falls on **precision**: the dataset is by design strongly unbalanced (thus making any metric that relies on true or false negatives useless) and the interest relies in identifying with certainty controversial comments from the rest of the corpus (thus ruling out recall).

Using a 5-fold cross validation, balanced in order to avoid having no `HCCs` in a training fold, I could estimate the precision of the predictor to be about 0.022. Essentially, an **awful performance** that shows how the vocabulary alone is by far not enough to achieve good results in pinpointing controversiality. This means that other factors (such as the ones found in [2]) must intervene to make a comment controversial.

I repeated the experiment by training the tree on the entire April 2017 dataset and testing on the February 2018 one, achieving a precision of around 0.93 on the training set and of 0.03 on the test set.

Concluding remarks

I could succesfully identify controversial comments in the corpus in a completely data-driven way and reproducible way.

Controversial comments showed differences in the lexicon used and in their lengths, controversial topics were mostly related to politics, ongoing debates and opinions expressed by the journalist. Overall, the most controversial categories were OpEd (Opposite the Editorial page, opinion of an independent journalist), Editorial, National, Styes and Business.

I also found that vocabulary alone is not enough to identify controversiality.

Future work involves using n-grams or skip-grams for tokenization, extending the analysis to the entire corpus of articles available (not just April 2017), using other attributes to predict controversiality, using other classifiers (hopefully less overfitting the data).

References

1. Shelke, N. M., Deshpande, S., & Thakre, V. (2012). Survey of techniques for opinion mining. *International Journal of Computer Applications*, *57*(13), 0975-8887
2. Ziegele, M., Breiner, T., & Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, *64*(6), 1111-1138
3. Mukwazvure, A., & Supreethi, K. P. (2015, September). A hybrid approach to sentiment analysis of news comments. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)* (pp. 1-6)
4. New York Times comments dataset : https://www.kaggle.com/aashita/nyt-comments
5. New York Times comments policy : https://help.nytimes.com/hc/en-us/articles/115014792387-Comments
6. Lumley, Thomas; Diehr, Paula; Emerson, Scott; Chen, Lu (May 2002). The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health*. **23** (1): 151–169