Extracting Controversy of News Comments Through Sentiment Analysis

Andrea Sala [†]

April 2021

1 Project Goals

Sentiment Analysis is a natural language processing technique that aims to identify the sentiment present in a given text, in order to classify its polarity. Such technique finds application in several fields, including the news. While numerous works have been done on sentiment analysis of news articles [1, 2], only a few put their attention of news comments. However, the set of comments under online articles constitutes an interesting dataset with possibilities of extracting useful information. Only two studies are present about sentiment analysis of news comments, retrieved respectively from *The Guardian* [3] and Chinese news websites [4]. Although both articles provide a detailed study of polarity of news comments, their only objective is to classify comments based on polarity by means of a properly trained model. No previous work has been found regarding the search for *controversy* in such comments.
The goal of this project is to determine whether an online article raised controversy. In order to do this, an index of controversy for each article was defined. Ultimately, the goal of this project is to uncover the most controversial categories of online news articles, highlighting the difference between the former and the most popular categories.

2 Dataset

The dataset used for the project is the *New York Times Comments* dataset available on Kaggle [5]. As the dataset webpage states:

> The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves. The csv files for comments contain over 2 million comments in total with 34 features and those for articles contain 16 features about more than 9,000 articles.

Among the available features, this project made use of the following:

- For the comment dataset: article ID, comment body, editors selection, recommendations, section name, desk and type of material.
- For the article dataset: article ID, author, section, desk and type of material.

The dataset was easy to upload and manipulate thanks to the `pandas` framework. However, more than half of the articles did not belong to any news section. This issue will be dealt more accurately in Sec. 4.

[†]Department of Physics, Università degli Studi di Milano

3 Methodologies

This Section presents the methodologies used to accomplish the project goals. Further details for the Sentiment Analysis techniques can be found in [6].

3.1 Data collection and preprocessing

First of all, the datasets were loaded from csv format into a Python Notebook. Since articles and comments from different months were stored in different files,they were merged into two Pandas `DataFrames`. After that, all the features considered unnecessary for the project were removed. The following step is to preprocess the comment bodies in order to prepare them for sentiment analysis. The preprocessing tasks included punctuation removal, case lowering, stop-word removal and stemming. The latter two were performed using methods from the NLTK package [7]. In particular, stemming was performed via a Porter Stemmer. Furthermore, all articles without any headline, section name and desk name were removed.

3.2 Sentiment Calculation

Once the comments were ready to be processed, a lexicon-based method was used to calculate polarity for each word in the comments. In particular, the AFINN lexicon [8] was used. This lexicon consists of a list of more than 3000 words, each associated with an integer score between -5 and 5. The score for each comment c, from now on labelled as CS, was calculated with the following formula [3]:

$$CS(c) = \frac{\sum_w S(w)}{\sqrt{N+1}} \tag{1}$$

where $S(w)$ are the single-word scores, and N is the total number of words in a comment. The definitive score for a comment, named *popularity*, was the sum of CS and a normalised count of the recommendations each comment received. Furthermore, the score was doubled if the variable `editorsSelection` was set to 1, meaning that the comment had been appreciated by the New York Times editors

$$Pop(c) = (1 + \text{editorsSelection}(c)) \left(CS(c) + \frac{\text{Recomm}(c)}{\max_c \text{Recomm}(c)} \right) \tag{2}$$

The score for each article (AS) was subsequently computed through the formula after grouping comments by the article a they belonged to:

$$AS(a) = \frac{\sum_{c \in a} Pop(c)}{\sqrt{M(a)+1}} \tag{3}$$

where CS(c) is the comment score and $M(a)$ is the number of comments related to a given article a.

3.3 Evaluation strategies

The index of controversy for each article needed to take account of two factors: popularity of the comment, and if there was any form of debate between comments from different users. The former could be estimated naively as the number of comments under each article; the latter was evaluated as the interquartile range (IQR) of the $Pop(c)$ distribution for each article. This choice was made because the IQR is an indicator of statistical dispersion. Thus, one can expect for an article with opposing opinions to present two populated regions at the extremes of the distributions. The controversy index was hence defined as

$$\text{Contr}_1(a) = \text{IQR}(a) * M(a) \tag{4}$$

The results with this strategy (see Section 4) suggested to implement another strategy: applying the natural logarithm to the number of comments, to obtain

$$\text{Contr}_2(a) = \text{IQR}(a) * \log M(a) \tag{5}$$

After defining and calculating the index of controversy, articles were grouped by different categories, such as section name, new desk and author, in order to see whether some categories were more controversial than others. All the results from this groupings are reported in Sec. 4.

Lastly, the keywords from comments of the most controversial articles were extracted by means of a TF-IDF vectorizer and a χ^2 selector. Among the best features obtained, only those present in the AFINN dictionary were kept.

4 Results

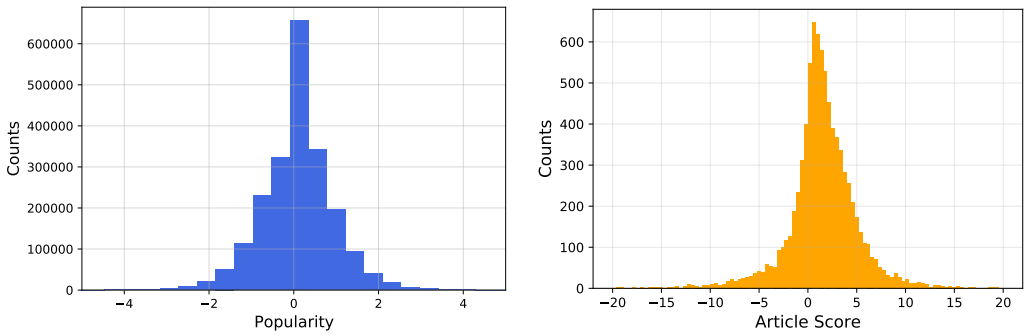
The following Section summarises the work by presenting the obtained results. Comments on the goodness of such results are also present, together with possible future developments. The popularity distribution computed through Eq. (2) shows a gaussian-like distribution centred in 0 (Fig. 1a), meaning that on average comments are neutral, with lateral tails of comments with strong polarity, both positive and negative. Recommendations play a small role, because of some outliers causing the normalised recommendation count to be quite a small number. It is fundamental for this distribution to be unbiased as it will be used to estimate controversy.

The article score distribution shows a similar behaviour; it is reported in Fig. 1b even though it will not be included in further analyses. Again, the article score calculated with Eq. 3 shows a quite symmetrical distribution, meaning that the estimator is consistent and the articles are equally distributed among the polarity spectrum.

4.1 Strategy 1 (linear)

The first strategy to estimate controversy in news articles (Eq. 4) was to multiply the number of comments related to each article with the IQR of the popularity distribution for such comments (labelled as *debate*). This calculation lead to an index of controversy peaked near 0. Furthermore, the controversy distribution looks very similar to the comment number distribution (see Fig. 2). This suggests that popularity and controversy are strongly correlated, while debate plays an insignificant role in the compute.

To corroborate this hypothesis, the correlation scatter plots for the three variables were built, and Fig. 3 shows a very strong linear correlation between popularity and controversy.



(a) Popularity distribution

(b) Article score distribution.

Figure 1: Preliminary distributions.

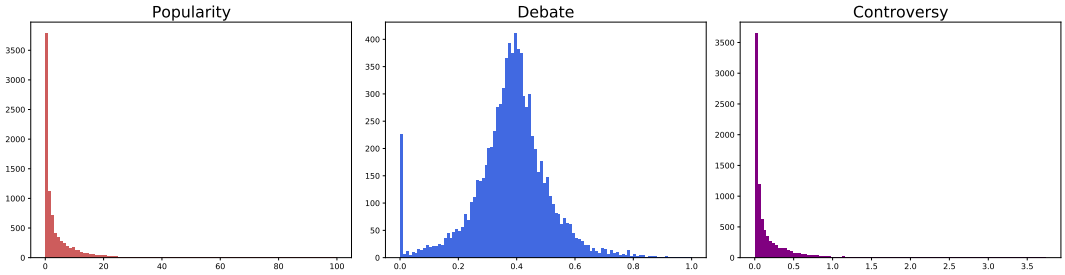


Figure 2: Popularity, Debate and Controversy distributions for Strategy 1

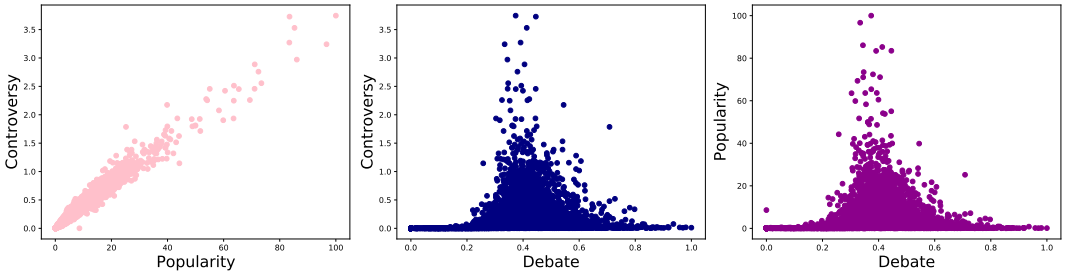


Figure 3: Correlation plots for Strategy 1.

This correlation rows against the thesis of the project, which is trying to differentiate the so-called "hot topics" from the most controversial ones. For this reason, another strategy was implemented to see if some other definition of controversy can detach more efficiently from the concept of popularity.

To go deeper into the analysis, the controversy scores for each articles were averaged among articles of the same section. This grouping was made to see if any article section is more controversial than others. Unfortunately, Strategy 1 did not prove to be useful in this case, as the most controversial topics are also the most popular one. Figure 4 shows how, for each section, the controversy and the popularity score look very similar to each other (after proper normalisation).

After that, articles were also grouped by the "new desk" index, to see if there is any category that stands out as the most controversial. Again, results (Fig. 5) indicate a high similarity between controversy and popularity, with "National" desk as the most popular desk.

4.2 Strategy 2 (log)

Strategy 2 consisted on applying the natural logarithm to the number of comments under each article, in order to dampen the strong dependencies between the former and controversy. Thanks to this new strategy, results change significantly, and they are reported in the same format as Section 4.1. First of all, the popularity distribution (hence the controversy index) are represented in their new shapes, while the debate distribution obviously does not change (Fig. 6). By looking at the correlation plots, it can be seen that the strong linear correlation between the two distributions has been removed (Fig. 7). The problem is now that distributions are not smooth for small values of popularity, because of the logarithm. However, the distribution of controversy looks smoother and more balanced compared to the one adopted in Strategy 1.

If we look at the averages on categories (Fig. 8 and 9), it can be seen that popularity and controversy are now two very different quantities. The bar plots both for section name and desk show different behaviour between the two data series. However, the problem with Strategy 2 is that the categories labelled as "highly controversial" are quite unexpected, such

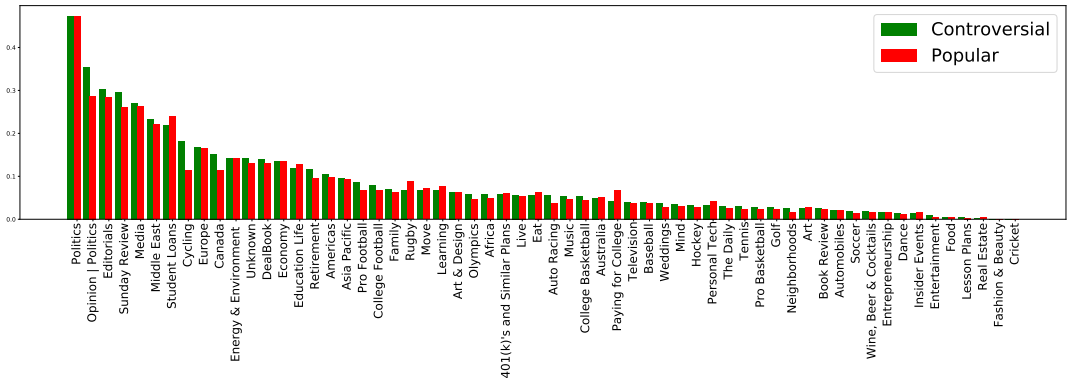


Figure 4: Most controversial sections for Strategy 1

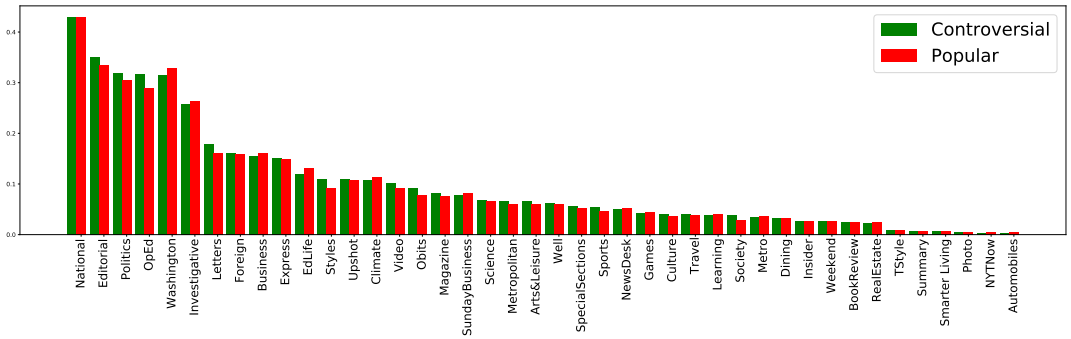


Figure 5: Most controversial desks for Strategy 1

as Cycling and Entertainment. On the other hand, controversial categories in Strategy 1 were Politics, Editorials and Middle East (categories one would expect to be controversial).

4.3 Keyword extraction

Lastly, a feature extraction was performed to check which words were common in comments related to the most controversial articles (top 10%). A corpus vectorisation was performed via a `TfidfVectorizer`, and then a $\chi 2$ analysis was invoked to select words with higher frequency. Since TF-IDF tends to boost the score of words which are very uncommon in the corpus, among the "top controversial words" there were many nonsensical words that occurred only once in the whole corpus, probably originated by typos of the comment authors. For this reason, only the words with a match in the AFINN dictionary were kept. Figure 10 depicts the words selected by the χ^2 analysis. An ample fraction of the word list is made of polarised words, indicating that polarity is at the basis of controversy.

5 Conclusions

The aim of this project was to define controversy in news articles by analysing the comments below the articles. Two different strategies were adopted, each one with its pros and cons. Strategy 1 defined controversy through a linear relationship; it was the one who guaranteed the results with most common sense (Politics and Middle East were among the top categories), but it failed to detach from popularity. On the other hand, strategy 2 guaranteed a significant shift from popularity and removed linear correlations, but the results leave us with some doubts. In any case, it was proved that in either case polarity in the comments is the key to analyse controversy. Some future developments to this work could be:

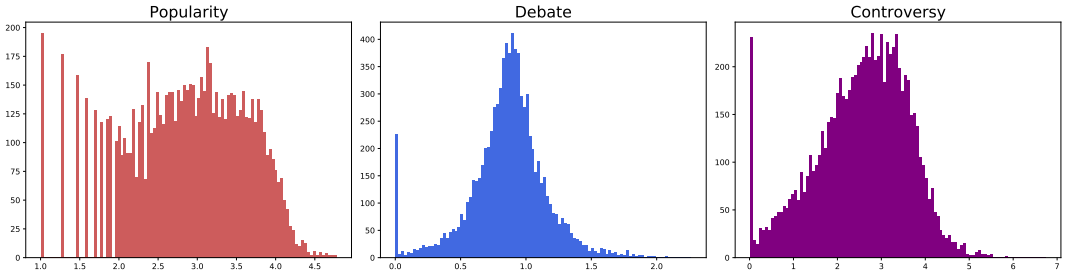


Figure 6: Popularity, Debate and Controversy distributions for Strategy 2.

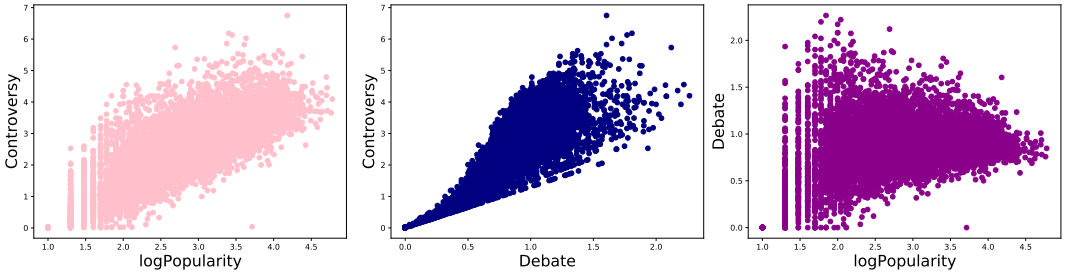


Figure 7: Correlation plots for Strategy 2.

- Perform aspect-based sentiment analysis to go deeper into the comments, and analyse which aspects within a given comment generated controversy
- Using topic modelling, extract the article section name from the comments, since nearly half of the articles in the dataset were provided with "Unknown" section name
- Demonstrate the robustness of this controversy estimator by changing the lexicon (for example, WordNet could be used instead of AFINN) and look for similar behaviour of the controversy distribution
- Find another strategy to remove correlation between popularity and controversy without the need to apply a logarithm
- Train a classifier able to predict controversy in other datasets of news and comments

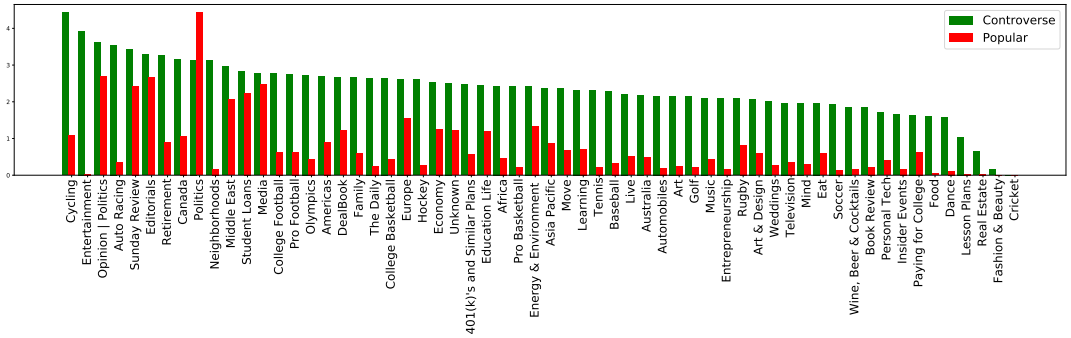


Figure 8: Most controversial sections for Strategy 2

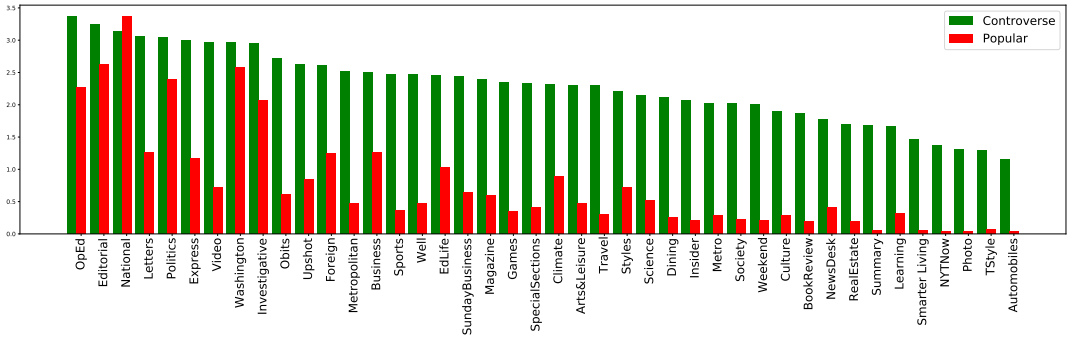


Figure 9: Most controversial desks for Strategy 2

```
Top 78 most controversial words:

anger  assault  attack  award  bad  ban  best  better  blame  bless  brilliant  care  corrupt  crime  danger  dead
death  destroy  die  disgust  enjoy  evil  fake  fan  fear  fire  fraud  fun  gift  glad  good  great  gun  ha  har
ass  hate  help  hope  joy  kill  liar  like  lost  love  luck  murder  nice  perfect  prison  problem  punish  rac
ism  racist  rape  rapist  shame  share  shoot  stupid  super  superb  support  terror  terrorist  thank  treason
true  victim  violent  war  warn  wealth  win  winner  worst  wow  wrong  xoxoxo
```

Figure 10: Most controversial words in the comment dataset.

References

[1] T. B. Magadza, A. Mukwazvure and K. P. Supreethi, *Exploring Sentiment Classification Techniques in News Articles*, IJITKM Vol. 8, no. 1, pp. 55-58, 2014

[2] P. Raina, *Sentiment Analysis of News Articles Using Sentic Computing*, 2013 IEEE 13th Int. Conf. Data Min. Work., pp. 959-962, 2013

[3] A. Mukwazvure and K. P. Supreethi *A hybrid approach to sentiment analysis of news comments*, 4th International Conference on Reliability, 2015

[4] W. Fan and S. Sun, *Sentiment Classification of Online Comments on Chinese News*, International Conference on Computer Applications and System Modeling, 2010

[5] A. Kesarwani, *New York Times Comments Dataset*, Kaggle link

[6] C. Aggarwal, *Machine Learning for Text*, Springer 2018

[7] S. Bird, E. Klein and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, O'Reilly Media Inc., 2009

[8] F. A. Nielsen, *AFINN*, Informatics Math. Model Tech. Univ. Denamrk, 2011