Maria Giustiniano

Public opinion on news

Academic year 2020/2021

Contents

1 Introduction

The New York Times comments dataset [2] contains CSVs reporting information on the editorial's articles and relative comments. The collection is divided into into two different sets of files, one grouping the articles published the other the comments published in the period that goes from January 2017 to May 2017 and from January 2018 to April 2018.
Overall the dataset is well documented, it offers both standard information as the identification ID, comments' contents, date of publication, and some contextual information like the articles topic and keywords and the comments' ”likes” count through the ”newDesk”, ”keywords” and ”recommendations” feature respectively. This data can serve the purpose of understanding and analyzing the public mood and acquire which comments first and which articles then raise the discussion the most.[1]

1.1 The New York Times editorial line

The newspaper is organized into three sections, including the magazine.

1. **News**: Includes International, National, Washington, Business, Technology, Science, Health, Sports, The Metro Section, Education(Edlife), Weather, and Obituaries.

2. **Opinion**: Includes Editorials, Op-eds and Letters to the Editor.

3. **Features**: Includes Arts, Movies, Theater, Travel, NYC Guide, Food, Home & Garden, Fashion & Style, Crossword, The New York Times Book Review, T: The New York Times Style Magazine, The New York Times Magazine, and Sunday Review.

4. **Upshot**: The New York Times website with analysis and data visualizations about politics, policy and everyday life

2 Procedure

To have a balanced comparison between 2017 and 2018 the CSVs regarding May 2017 are removed, in this way the analysis examines just the first four months of the available years.
The procedure can be synthesized in three steps:

1. Polarity detection and comments pattern reconstruction;

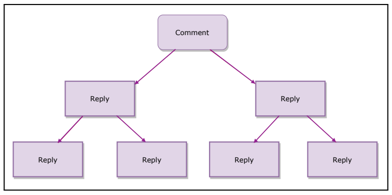
2. Comments investigation;

3. Articles analysis;

2.1 Polarity detection and comments pattern reconstruction

This phase involves as first the ”SentimentIntensityAnalyzer” tool to estimate the sentiment expressed by each comment in terms of numeric magnitude. This measure will be later exploited for catching how users emotion could be distant while arguing some particular theme.

Consequently, the main effort was to identify the latent sequence linking each comment with the corresponding replies. Within the collection, each comment type, namely: *comment*, *userReply*, *reporterReply*, is uniquely identified by the *comment ID*. Furthermore, the comments labelled as *userReply* or *reporterReply* are connected with their predecessors through the feature *parentID*. Given this peculiarity, it is reasonable to face the task using graph logic: having collected all the existing edges, that is all the comments pairs in the form of **parentID** →**replyID**, it was possible working with the *NetworkX* package to derive the underlying directed acyclic graph (DAG), in which the roots are the comments and the leaves and the intermediate nodes their replies.



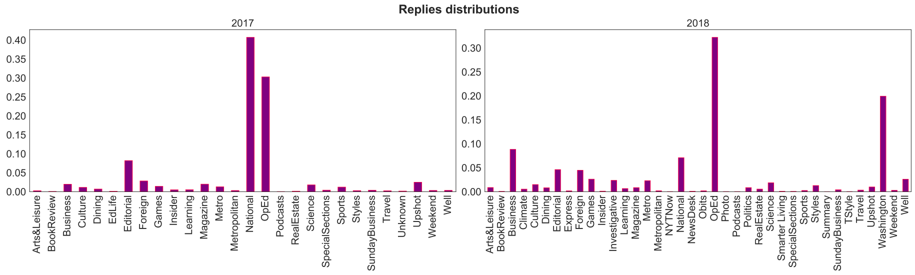
Algorithm 1 Get Successors

```
function GET SUCCESSORS(df)
    nodes= []
    for each record in df do
        if parentID exists then
            add the pair (parentID, commentID) to nodes
        end if
        create the dag starting from nodes
        retrieve all the roots and create a dataframe(DAG) with them
        successors=[]
        for root in DAG do
            add successors’ nodes to successors
        end for
        add to DAG the column with the successors of each root
        merge DAG with df
    end for
    return: df
end function
```

The procedure reveals that just for some comments the recorded reply count matches the one retrieved by the examination; nevertheless, given the wide volume of data, this lack is negligible to get a significant conclusion.
Finally, for each root, the lists with the successors' recommendations and the successors' polarities are built and the standard deviation of these sequences computed. The idea behind the adoption of this metric as a dispute measure is the following: if a comment raises contention then some of its replies will absorb more consensus than others; in social networks, consensus is often expressed in terms of ”likes” and in this case by ”recommendations”. In light of this, the more conflicting discussions could be the ones in which the recommendations distributions are skewed the most and the standard deviation is much higher with respect to the mean. The same logic is appropriate for handling and interprets polarities distributions: the higher the standard deviation, the higher the dissonance from a sentiment point of view.

2.2 Comments analysis

Firstly, comments are grouped into two Data frames accordingly the year of publication.



A preliminary analysis shows that in 2017 the section with the highest number of replies was the ”National” one, followed by the ”OpEd” and the ”Editorial”. The motivation behind this result could be that many of the articles published in 2017 were part of the ”National” segment. Differently, in 2018 the ”OpEd” section ranked as first; nevertheless, in the first four-month of this latter many sections are remarkable: first of all, there's the ”Washington” one, followed by ”Business”, ”National”, ”Editorial”, ”Foreign” and ”Investigation”. In particular, 2018 was the Midterm election year thus this event has presumably fostered discussions with a political background.
At this stage, four different metrics are computed:

1. **pol_sd**: that is the average of the polarities standard deviation per comment;

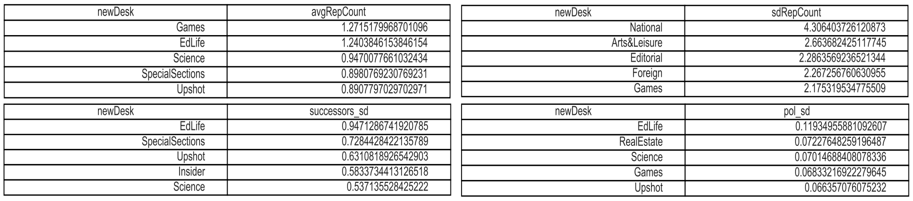
2. **successors_sd**: the average of the recommendations standard deviation per comment;

3. **avgRepCount**: the average number of replies per comment;

4. **sdRepCount**:the standard deviation of replies per comment.

The first two metrics can be interpreted as a measure of how much controversial a comment-reply conversation can be, whereas the last two as a means to get the estimated width of these discussions.

2017



newDesk	avgRepCount
Games	1.2715179968701096
EdLife	1.2403846153846154
Science	0.9470077661032434
SpecialSections	0.8980769230769231
Upshot	0.8907797029702971

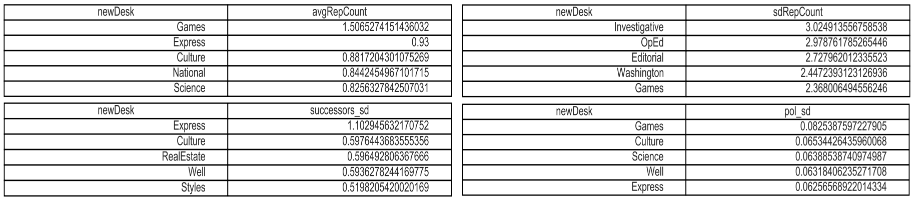
newDesk	sdRepCount
National	4.306403726120873
Arts&Leisure	2.663682425117745
Editorial	2.2863569236521344
Foreign	2.267256760630955
Games	2.175319534775509

newDesk	successors_sd
EdLife	0.9471286741920785
SpecialSections	0.7284428422135789
Upshot	0.6310818926542903
Insider	0.5833734413126518
Science	0.537135528425222

newDesk	pol_sd
EdLife	0.11934955881092607
RealEstate	0.07227648259196487
Science	0.07014688408078336
Games	0.06833216922279645
Upshot	0.066357076075232

It emerged that in 2017 the first section per *avgRepCount* is the "Game" one; nevertheless, given the large standard deviation, it is more plausible to fall into a long comment-reply conversation if we shift towards the "National" section. The *succesors_sd* and the *pol_sd* metrics suggest that the most controversial talks are within the "EdLife" section.

2018



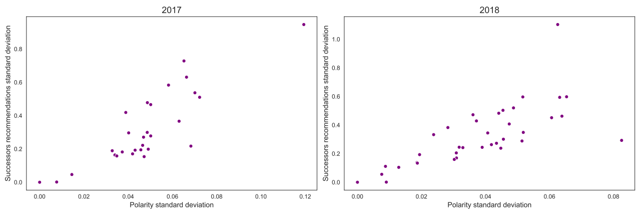
newDesk	avgRepCount
Games	1.5065274151436032
Express	0.93
Culture	0.8817204301075269
National	0.8442454967101715
Science	0.8256327842507031

newDesk	sdRepCount
Investigative	3.024913556758538
OpEd	2.978761785265446
Editorial	2.727962012335523
Washington	2.4472393123126936
Games	2.368006494556246

newDesk	successors_sd
Express	1.102945632170752
Culture	0.5976443683555356
RealEstate	0.596492806367666
Well	0.5936278244169775
Styles	0.5198205420020169

newDesk	pol_sd
Games	0.0825387597227905
Culture	0.06534426435960068
Science	0.06388538740974987
Well	0.06318406235271708
Express	0.06256568922014334

Even in 2018 "Games" reconfirms as the first section per *avgRepCount*, attesting itself as the thematic area whose comments catch the highest reply response with respect to comments of different domains. According to the *sdRepCount*, "Investigation" seems to be the area in which comment-reply debates could continue longer, even though the difference with the sections immediately following is tiny. Finally, concentrating on *pol_sd* and *successors_sd* metrics, the appearing thematic fields are "Games" again and "Express" respectively.



Lastly, focusing on *pol_sd* and *successors_sd*, it seems that an increasing level of one metrics corresponds to a growing level of the other, thus, on average, comment-reply conversations with a proper spread of consensus rate are also the ones in which the opinion is more debatable.
Given these results, the thematic area in which comment-replies conversation-like appears to be more frequent is the "Games" one. Instead, more uncertainty surrounds those in which the debate turns to be more controversial. This ambiguity is due both to a different editorial production in the years under examination, both to the news events themselves which could differently move readers judgment.

2.3 Articles analysis

Much more interesting can be the analysis concerning the articles behind the examined comments. For this purpose, the proper CSVs are updated so that each newspaper article in now associated with three synthetic measures capable of summarizing the reactions expressed by readers through the comments. Specifically, these metrics are:

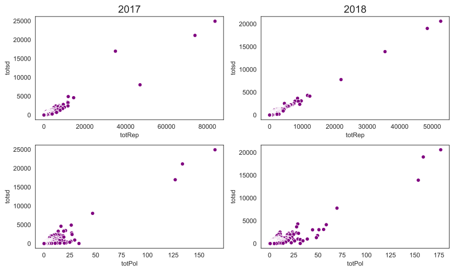
1. **polarity_sd**: that is the comments polarity standard deviation

2. **score_sd**: that is the comments recommendation number standard deviation

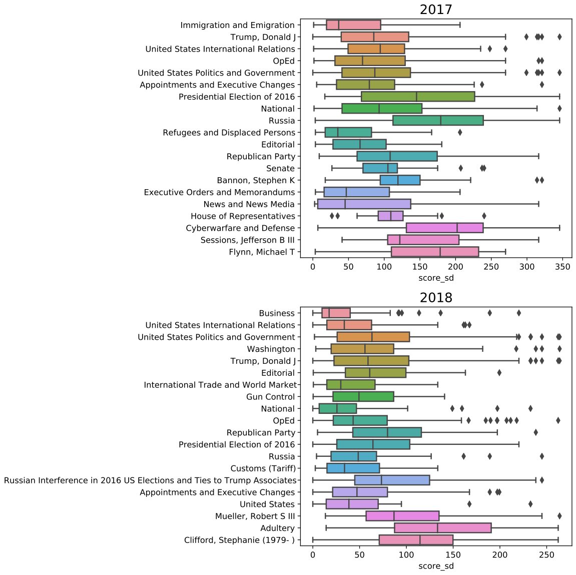
3. **totRep**: that is the sum of all comments received by an article, all updated with the relative sum of collected replies.

These metrics aim is to quantify the controversy degree that could be reached during a social networks-oriented debate by focusing on different aspects of the collected comments. To be more precisely, the *polarity_sd* attempts to isolate the controversy degree looking at the sentiment expressed by the comment content; with the *score_sd* the focus shifts on the consensus obtained by each comments, to get if some of them stand out when the number of the recommendations is take into account. Ultimately, the *totRep* observes the total number of answers per newspaper article to perceive which themes have pushed the readers to express their opinion the most.
After having grouped the newspaper articles accordingly to the year of publication, the main effort was to exploit the "keywords" feature to expand the thematic sphere touched by publications. By treating the data frames as two non-relational collections, it was possible to handle the **unwind** aggregation pipeline and unload the several lists of keywords, producing as many documents as the keywords are.Once the data frame-structure is restored, the articles are grouped by keywords. Then, for each group the sum of *score_sd*, *polarity_sd* and *totRep* is computed.

A first examination shows that both the *polarity_sd* and *totRep* sums are positively correlated with the *score_sd* sum.



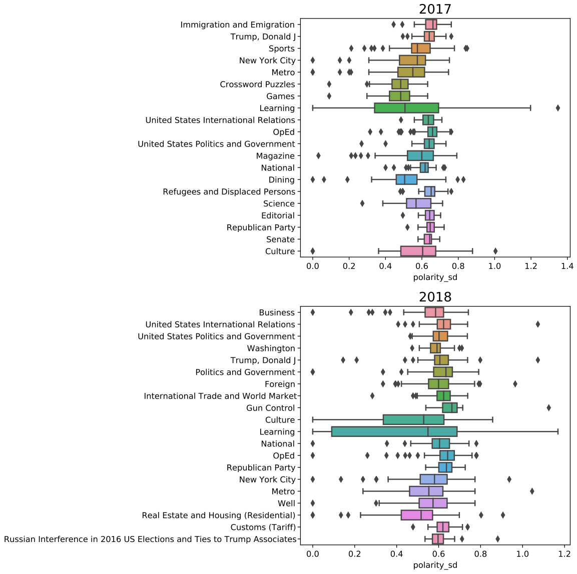
Next, looking only at the 20 keywords with the highest *score_sd* sum, it is appealing to inspect the distribution of the three different metrics grouped by keywords using boxplot.



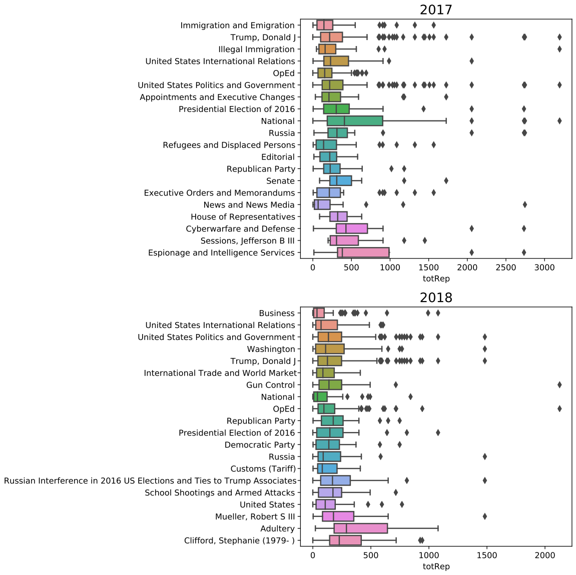
The *score_sd* distribution highlights some keywords: **"Presidential election 2016", "Russia", "Cyberwarfare and Defense", "Bannon, Stephen K", "Sessions, Jefferson B III"** and **"Flynn, Michael T"**. Indeed, 2017 has begun with the Inauguration ceremony, which marked the end of the Obama administration and the start of the Trump era. This transition could explain the growing debate surrounding both the Presidential election and the new administration policies concerning immigration or refugees. 2016 US elections were also overwhelmed by the threat of Russian interference. The so-called *Russiagate* involved Micheal Flynn and Jefferson Sessions too. The former was accused of having had affairs with the Russian ambassador before the elections; due to these suspicious, his National Security Advisor office lasted just twenty-three days. The latter, instead, caused a scandal when he refused to nominate the General Attorney to investigate the *Russiagate* case.

Another controversial character of the first four months of 2017 was Stephen Bannon, also known as *The Great Manipulator*; Bannon was the chief executive officer of Trump's 2016 presidential campaign and was appointed Chief Strategist and Senior Counselor to the President following Trump's election. His role ended in April 2017 due to some inconvenient opinion and declarations as to the one published in The New York Times on January 26th 2017, in which he stated "The media should be embarrassed and humiliated and keep its mouth shut and just listen for a while" "I want you to quote this," Mr Bannon added. "The media here is the opposition party. They don't understand this country. They still do not understand why Donald Trump is the president of the United States." [3]

In the same vein of the precedent year, the more controversial themes in 2018 share the same political background as those of 2017. This result is not surprising, since 2018 was the year of the Midterm elections. Again, looking at the selected keywords, those concerning Russia, Trump administration and his disputes are still on top. Indeed, one of the two names appearing among the 20 most contentious keywords is Stephanie Clifford, who was involved in a legal cause with the 45th US President. Robert Muller, instead, was one of the characters involved in the *Russiagate* who, after having been appointed by the Department of Justice as special counsel in 2017, published on February 16th 2018 the statement document against thirteen Russians and three organizations that played a role in the investigation.



By shifting to *polarity_sd* distribution, the most controversial themes in 2017 and 2018 seem to be coherent with those found while interpreting the *score_sd* distribution, with just some exceptions. Generally, most of them are still involved in the political sphere, but some others follow the tendency appeared also while examining comment-replies conversations. Themes as **"Games", "Science"** or **"Culture"** rise again as those that shift readers sensibility the more. Besides, even **"Sports", "Metro"** and **"Magazine"** appears as particularly relevant keywords.



Ultimately, *totRep* distributions are almost entirely congruent to the *score_sd* ones. Except for some slight differences emerging for the top keywords of 2017 (although still coherent with the political trend), those appearing for 2018 perfectly coincide.

3 Conclusion

The experiment aims to identify the most controversial topics. It does so by blending three different components acquired from the dataset: the recommendations' number, the reply count and the sentiment of the comments' content. The adopted strategy splits into two phases: one to distinguish the editorial sections with the more controversial comment-replies talks; the other, instead, to detect the keywords associated with the most discussed newspaper articles.
The results show that if the controversy is associated with the recommendations distribution, then the editorial sections raising the more contentious talks deal with, politics, education and everyday life; meanwhile, the keywords of the most debated articles mainly concern political themes. It is significant to underline that these articles were published during an uneasy period from a political point of view when one of the most contested administrations of American democracy has just born.
Alternatively, taking into account the polarity solely, the sphere of the contentious themes enriches with topics linked to videogames, culture and science world too.
In conclusion, given the two-steps' results, the most controversial topics are mainly those related to political matters and all the fields linked to this world.

References

[1] The New York Times.

[2] *https://www.kaggle.com/aashita/nyt-comments*

[3] New York Times. *https://www.nytimes.com/2017/01/26/business/media/stephen-bannon-trump-news-media.html*