Returning the dislike button

Text and Multimedia Mining

s1040068
Radboud University

ABSTRACT

Since YouTube has removed the dislike button in 2021, people have shown disapproval of this. We try to determine if this has resulted in different comments under YouTube videos, since this would be another place to give feedback on a video. We try to measure this by doing a sentiment analysis on the comments. From this, we may conclude that there seems to be an increase in positive sentiment and a decrease in negative sentiment on comments on videos from 2022 compared to similar data from 2017. It is not entirely clear whether this is the result of the removal of the dislike button, but we can say that the removal of the dislike button did not result in more negative sentiment in YouTube comments.

1 INTRODUCTION

On December 13th 2021, YouTube has decided to remove the dislike button[1]. According to YouTube, this is done to promote "respectful interactions between viewers and creators". This means that users can not easily see which videos are regarded as good or bad by other users. Furthermore, this means that users can now only express their dislike in the comment section. With this in mind, we do research to see if the removal of this has results on other elements of the YouTube page. In particular, we look at if and how comments have changed after the removal of the dislike button. We do this using a sentiment analysis tool and manually inspecting the most frequently used words.

1.1 Research Question

We present the following research question:
How have YouTube comments changed after the removal of the dislike button?

2 BACKGROUND AND RELATED WORK

2.1 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is defined as "the computational study of opinions, sentiments and emotions expressed in text."[4]. By definition of opinions, we are not necessarily looking to extract facts from our text. Since sentiment analysis excels in quantifying opinions, it is very useful for companies to investigate their reputation[2]. With enough data, they can get feedback on their products without expensive research. In our case, we apply this sentiment analysis on each YouTube comment individually, words in these comments that hold an opinion should mainly contribute to the sentiment analysis.

[1]https://blog.youtube/news-and-events/update-to-youtube/

TxMM'22-'23, Radboud University, Nijmegen, Netherlands

2.1.1 Techniques. There are two main approaches to sentiment analysis[7]. First, we have the machine learning approach. Here, we use a machine learning algorithm on linguistic features to predict the sentiment. This can be either supervised or unsupervised. When we use supervised machine learning, we have a large set of training data where each element has a label with a sentiment value. There are all sorts of classifiers that we can use, with Support Vector Machine and Naive Bayes being quite popular. However, text data is typically of a large size, so annotating everything can be very costly and time-consuming. Hence, we turn to unsupervised learning to conquer these issues. For example, we can split up the documents into sentences and perform association rule mining to determine relations[5].

On the other side, we have the lexicon based approach. Here, we either use a dictionary to give sentiment to certain words or we use a corpus-based approach to determine sentiment. Using a dictionary typically consists out of gathering a set of opinion words and then using something like a thesaurus to find synonyms[9]. When using a corpus, we are better prepared for the ambiguity of language. Because of this, we can apply it to more specific domains. A corpus can be generated using either a statistical approach, where words are found with statistical techniques, or using a semantic approach. Here, words are given a semantic value, with semantically close words having similar values. WordNet[8] can be used to give semantic relationships between words and compute sentiment polarity. When we encounter unknown words, we can use synonyms with known semantic values to determine the sentiment[3].

There are more approaches to sentiments analysis, such as a hybrid form which combines machine learning with lexicon features. There is also a BERT-based approach, which uses word-embeddings to predict sentiment[1]. These are more computationally expensive methods, but they can yield better results.

2.2 Analysis over time

We are interested in comparing the situation of two time periods using a sentiment analysis. This has been done before on a large Twitter dataset that was captured during the COVID-19 pandemic in 2020[6]. Here, they analyzed the sentiment of active Twitter users to compare the average sentiment per group per country. They presented this information using graphs to show the average sentiment. Some outliers were explained by external factors, such as the increase in infected cases.

2.3 YouTube comments ratings

YouTube comments can be rated by other users. They can get a score based on a like and dislike system. However, this is also removed by YouTube after December the 13th, 2021. The aim of Siersdorfer et al.[10] was to determine the score of these comments by looking at properties of the comments, such as their sentiment. They were able to give a reasonable prediction of the rating of comments. Using this information, they were able to predict whether a comment

would be accepted by the community. They also suggest that based on these scores, they can identify users with similar purposes and similar interests. This can be used to set up an active community of similar minded people.

3 RESOURCES

3.1 Return YouTube Dislike

Since other people did not agree with YouTube removing the dislike button, people have tried to bring it back on their own. An example of this is the browser extension Return YouTube Dislike[2]. This project also includes an API that gives access to estimated dislikes on all videos. According to their official frequently asked questions, the data is based on "A combination of archived data from before the official YouTube dislike API shut down, and extrapolated extension user behavior." This means that it is not fully reliable, but it can give us a reasonable estimate.

3.2 Video and Comments dataset

Research has been done already on the subject of sentiment in YouTube comments[3]. This was done in 2017 and it includes a dataset with around 700000 comments on 2300 (unique) videos. This data set thus has YouTube comments and video information from before December 13th 2021.

3.3 YouTube scrapers

To scrape the data from comments and videos (including dislikes), we created our own scraper[4]. Our video scraper is based on the scraper used by our other dataset[5], which is based on the YouTube video API[6]. It has been adapted to write the files in the same format as the existing dataset for compatibility reasons.

Scraping YouTube comments is a little more involved, since YouTube does not have a public API to retrieve all comments for a video. However, we can use the AJAX API that is used by the webpages to retrieve comments. Once again, our implementation is based on an existing scraper[7] and returns the comments in the same format as the existing dataset. The comment scraper scrapes comments in the order that you would see them on a YouTube page, hence we would expect that the most popular/relevant comments would be scraped first.

4 APPROACH

We begin by gathering a large dataset of videos with their comments. For this research, we limit our scope to the videos that are trending in the United States. This is because these videos will most likely be in English, this means that the comments will have the best chance of being English, which is what our sentiment analysis is best suited for. It also means that we can easily manually inspect these comments and understand what is being commented. We are interested in a large set of videos from before December 13th 2021 and another set that has been created after this date. From now on, we will refer to these datasets as B.D.R. (Before Dislike Removal) and A.D.R. (After Dislike Removal). We try to get these two datasets as balanced as possible, however, we are working with different sources, so this is not entirely possible. We at least require 200 comments and around 800 videos per dataset. These amounts should be sufficient to give us a reasonable dataset to answer our research question.

Table 1: Sentiment Library comparison

Library	NLTK	Pattern	Flair	Flair + NLTK
Precision	0.48	0.5	0.5	0.56
Recall	0.38	0.32	0.74	0.62

4.1 Gathering data

We get our B.D.R. data by using the large dataset from section 3.2. For the A.D.R. data, we use our own YouTube scraper (see section 3.3) to scrape 1000 videos and 1000 comments for each video. We also scrape the estimated dislikes from the new videos by using the Return YouTube Dislike API (see section 3.1). Both these actions were performed on the 17th of December 2022.

4.2 Preprocessing and feature extraction

To perform a sentiment analysis, we use a Python library. We can choose from Pattern[8], NLTK and Flair[9]. Both Pattern and NLTK are lexicon based, where Flair is pre-trained embedding based. It is trained on the IMDB movie reviews dataset. To decide between the three different libraries, we have annotated 25 comments from both datasets. Using these labels, we can compare the scores (see table 1). We see that Flair is the best in judging sentiments. However, it basically works binary, with only scoring almost 1 or -1. Hence, we will combine Flair with NLTK to get an accurate representation for neutral comments. Our sentiment value is computed using:

$$\texttt{sentiment} = \begin{cases} 0 & \text{if } \texttt{nltk} = 0 \\ \texttt{flair} & \text{otherwise} \end{cases} \tag{1}$$

For each video, we use the comments with their sentiment scores to compute: Total sentiment (positive, negative and combined), comments retrieved (positive, negative, neutral and combined) and sentiment per comment (positive, negative and combined). Other features that we have on our videos include: views, likes, dislikes, total number of comments and category. We also have the title and YouTube identifier of a video, but these are only interesting for a manual analysis (see section 6).

We discard our videos with 0 dislikes, because this is something that never happens in practice for popular videos. This only affects a couple of videos, so our dataset will not be changed drastically. In the end, the B.D.R. dataset has around 1600 videos and the A.D.R. has around 800 videos. Both datasets have around 750000 comments, so the A.D.R. dataset has a lot more comments per video.

[2] https://www.returnyoutubedislike.com/
[3] https://www.kaggle.com/code/tanmay111/youtube-comments-sentiment-analysis
[4] https://gitlab.science.ru.nl/trust/txmm-project
[5] https://github.com/mitchelljy/Trending-YouTube-Scraper
[6] https://developers.google.com/youtube/v3/docs/videos/list
[7] https://github.com/ahmedshahriar/youtube-comment-scraper
[8] https://github.com/clips/pattern
[9] https://github.com/flairNLP/flair

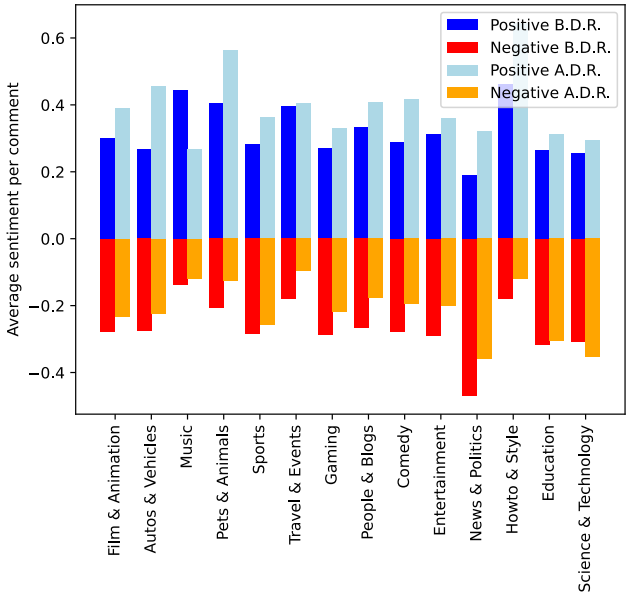

Figure 1: Average sentiment per categorie

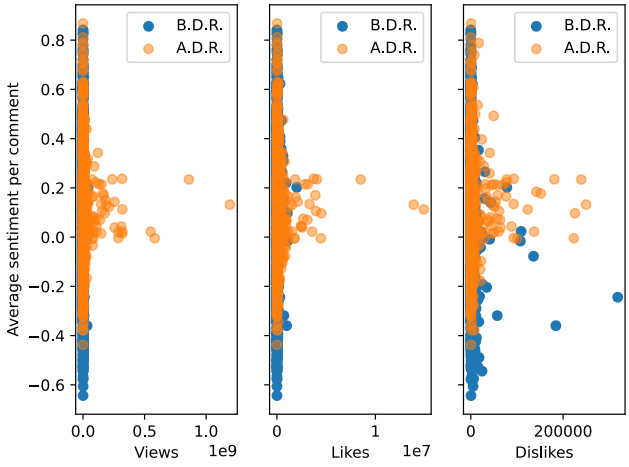

Figure 2: Sentiment and views, likes and dislikes per video

5 RESULTS

In this section, we use our features on the videos and comments to answer our research question: *How have YouTube comments changed after the removal of the dislike button?*

In figure 1 we see the average sentiment per comment, grouped by categories for both datasets. We see that the B.D.R. dataset has the least positive sentiment in almost all categories. On the negative side, we see roughly the same trend, where the B.D.R. scores lower on almost all categories. This may mean that people are overall happier with the videos that are trending right now compared to when the B.D.R. dataset was created in September 2017.

We inspect the relation between sentiment per comment and views, likes and (estimated) dislikes by plotting these together in

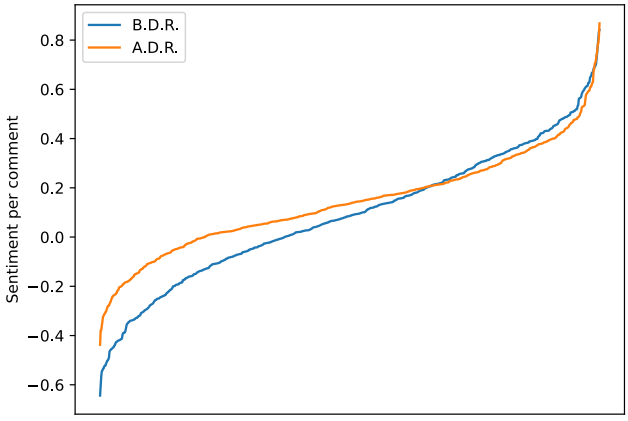

Figure 3: Distribution of average sentiment per comment

figure 2. From the figure, we can see that the most popular videos (in terms of views) have a total sentiment of around 0 to 0.4. The A.D.R. datasets seems to score a bit higher on all metrics except the dislike, which could be because YouTube has become more popular over the years. Moreover, the dislikes for the A.D.R. are provided by Return YouTube Dislike (see section 3.1), so they may not be entirely accurate. Furthermore, we see that both datasets have videos in the upper range of sentiment, whereas we see that only the B.D.R. dataset has videos in the low range of sentiment. Another interesting observation is that the videos in the B.D.R. dataset with the most dislikes also have quite a low average sentiment.

To inspect the distribution of the sentiment on comments, we plot this metric in figure 3. On the horizontal axis, we have the videos ordered by average sentiment per comment and on the vertical axis we have this average sentiment. What we can see from this is that the comments from the B.D.R. generally have a lower sentiment score than the A.D.R. dataset. An exciting observation that can be made is that the sentiment on videos from the B.D.R. dataset is less balanced (the slope of the line is higher). This could be because we have more comments per video in this dataset. Another explanation is that the videos in the B.D.R. are more different in terms of raising sentiment (positive and negative) among the viewers, which may cause the larger sentiment difference between videos.

Finally, to see how comments have changed in terms of words, we will inspect the most used words that are in our dataset. We look at the most used words using a word cloud. We only look at the words that are used in comments that are classified as negative. Since the positive comments largely are the same, and we can not really say something about all words, since this seems highly dependent on the videos that they belong to. When we look at our word clouds in figure 4 and 5, we see that the 5 most popular words are very similar in both clouds. However, when we look at swearwords, we see that they are much more present in the B.D.R. dataset. This means that commenters now use fewer swearwords than before, which may mean that they are more appreciative towards content creators and other users. We also tried looking at the most used n-grams of words, but this did not give us any results.

Figure 4: Word cloud of negative comments in B.D.R. dataset

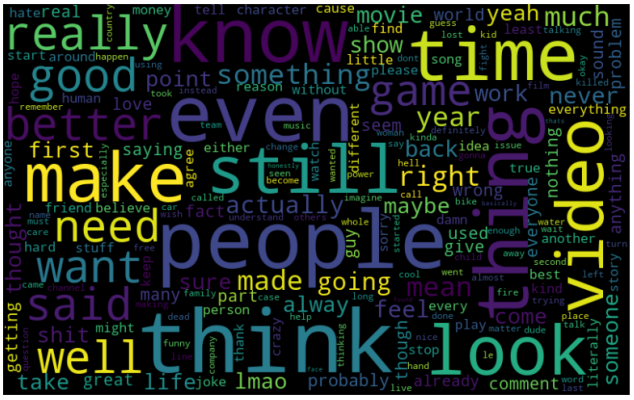

Figure 5: Word cloud of negative comments in A.D.R. dataset

6 DISCUSSION/VALIDATION

We only look at videos that were trending on two different dates (one before and one after the removal). Similarly, our small selection of videos may also not be an accurate representation of videos on YouTube. Another limitation of this work is that the comments from both datasets can not be easily compared in a way that we can draw a hard conclusion. For example, we compare different types of videos that may be aimed at different audiences. Moreover, any changes in sentiment or words usage does not necessarily need to be related to the removal of the dislike button.

On the other hand, based on the results we can say that comments on the A.D.R. dataset show more positive sentiment and the negative sentiment seems less intense compared to the B.D.R. dataset. Furthermore, we see that videos with a lower sentiment also seem more likely to gain more dislikes. These results indicate that people seem more happy in the comments on the newer videos. However, this does not necessarily mean that this is a direct consequence of the removal of the dislike button. It could for instance also be because video quality has improved over the past years. Nonetheless, we may conclude that the removal of the dislike button did not result in a significant increase in negative comments to make up for this removal.

To validate our work, we try to compare videos from the same channel that are in both datasets. We do this manually in the notebook[10]. After inspecting a few sets of similar videos of the same channel, we see the same trend that the A.D.R. dataset typically contains more positive and less negative feedback.

7 FUTURE WORK

To try to get two less diverse datasets, we may extend this work by only focussing on certain categories. Another solution would be to inspect comments from the same author. We could extend this research by using other measurements to compare comments from the two datasets. For instance, using topic modelling or by inspecting the usage of word groups such as swearwords and compliments. Another idea is to compare other features of comments, such as the number of replies or likes on comments.

8 CONCLUSION

When we compare the sentiment of comments before the removal of the dislike button to the comments on videos after this date, we can clearly see a positive change in sentiment. We have less negative comments, more positive comments and less usage of swearwords. With this research, we are however unable to say that this result is caused by the removal of the dislike button. It may also be the case that external factors have had a large influence. But in any case, we can say that the removal of the dislike button did not result in a large increase in negative sentiment in the YouTube comments.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[2] Ronen Feldman. 2013. Techniques and Applications for Sentiment Analysis. *Commun. ACM* 56, 4 (apr 2013), 82–89. https://doi.org/10.1145/2436256.2436274

[3] Soo-Min Kim and Eduard Hovy. 2004. Determining the Sentiment of Opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. COLING, Geneva, Switzerland, 1367–1373. https://aclanthology.org/C04-1200

[4] Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2, 2010 (2010), 627–666.

[5] Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (New York, NY) *(KDD'98)*. AAAI Press, 80–86.

[6] Muvazima Mansoor, Kirthika Gurumurthy, Anantharam R U, and V R Badri Prasad. 2020. Global Sentiment Analysis Of COVID-19 Tweets Over Time. https://doi.org/10.48550/ARXIV.2010.14234

[7] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[8] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (nov 1995), 39–41. https://doi.org/10.1145/219717.219748

[9] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 599–608. https://aclanthology.org/D09-1063

[10] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How Useful Are Your Comments? Analyzing and Predicting Youtube Comments and Comment Ratings. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 891–900. https://doi.org/10.1145/1772690.1772781

[10] https://gitlab.science.ru.nl/trust/txmm-project/-/blob/main/notebook.ipynb