A Hybrid Approach to Sentiment Categorization: A Vader-Bayes Sentiment Analysis for Youtube Comments on Films

18 December 2022

Jiaming Li, Mark Lung, Chris Chen, Brianna King

Source Code:
https://github.com/ljm297/NLP-VADER-BAYES

Abstract

One of the most popular social media video viewing websites, Youtube, allows its creators to use Youtube's own analytics to help grow their channel. This app, Youtube Studio, provides analytical tools to analyze audience retention, impressions, average view length, likes, and number of comments, among other things. Comments are one of the most important metrics of a youtube video's success. However, unlike movie reviews, comments do not have a corresponding rating for said video. Analyzing the comments of these videos through a hybrid machine learning and lexicon approach provides each viewers' sentiment on a video, which can allow creators to obtain positive or negative feedback. Using Sentiment VADER as a baseline, data preprocessing via tokenizing, stopwords, and other lexicon based approaches, and the machine learning Naive Bayes algorithm, the sentiment analysis program achieves an accuracy of 76.36%, Precision of 94.15%, Recall of 70.43%, and an F1 score of 80.58%.

1. Introduction

Since its inception on February 14, 2005, Youtube has been one of the most popular video viewing sites. Although YouTube is primarily a video sharing site, it also allows users to interact with its creators via comments, likes, and dislikes. Through its analytical web app Youtube Studio, Youtube has gradually increased its support for its creators over the years. Despite this, no resources are provided for the overall sentiment of a video's comment section. Furthermore, a video's statistics are only visible to the creator, and with the removal of dislikes, it becomes increasingly difficult for viewers of a specific video to understand the general sentiment of a video without consulting the comments.

This is concerning for the viewer, especially in the ever-popular "review space." When it comes to movie and technology reviews on YouTube, determining whether the audience agrees with the review is critical before making a decision. Both the YouTube review and the comments on the review are important in shaping the viewer's perception of a product.

As a result, given a YouTube video, it is critical to analyze each viewer's sentiment of a video through the comment section. However, it may be difficult to analyze because, without the assistance of sentiment analysis programs, manually analyzing each comment may prove both tedious and ambiguous. Through lexicon and machine learning algorithms, we hope to implement an NLP Youtube sentiment analysis model that will assist in overcoming the challenges of identifying sentiment in this project. The classification results will assist viewers in understanding the overall sentiment of any video they watch.

1.1 Overview of Sentiment Analysis

A sentiment is an attitude, belief, or conclusion brought on by a sensation. Sentiment analysis, commonly referred to as opinion mining, examines how individuals feel about particular things by extracting information from these individual opinions. Sentiment analysis takes place on three levels (Behdenna et al., 2018), the sentence level, the document level, and the aspect level.

Sentiment analysis at the sentence level categorizes each sentence based on the expressed sentiment. The sentence level can be classified as positive, negative, or neutral. Another form of sentiment analysis categorizes the entire document, usually as positive or negative, at the document level. The final level of sentiment analysis happens at the aspect level. Here, the opinion of a sentiment is classified rather than a general neutral,positive, or negative review.

Unlike an aspect level sentiment analysis categorization, both document and sentence level sentiment categorizations are typically classified as neutral, positive, or negative. On account of YouTube recordings, a positive opinion addresses a general preference of the video, though a negative feeling addresses a general abhorrence of the video. Investigating product sentiment provides answers to questions about how the product is performing in the market, such as whether it is receiving positive or negative responses.

Sentiment analysis is important and beneficial because it provides an efficient method of processing and analyzing large amounts of data. It can also be applied to various forms of data such as surveys, reviews, chat boxes, articles, etc. For customer-driven companies like Youtube, having the ability to quickly generate feedback and insights based on Youtube comments can help creators on the site to learn how their viewers feel about their content. As well as allowing the viewers to also observe the opinions of a video in more detail than both the comment section and the likes section could give. These creators can then provide better content and services to the consumers of their videos which can lead to greater success for both the creators and the company, as well as greater enjoyment for the users.

1.2 Techniques Used in Sentiment Analysis

1.2.1 Lexicon Approach

Sentiment analysis categorization is usually done in three different ways. One common technique is through lexicon/library based. The lexicon-based approach uses a dictionary of known words and searches the dictionary for these seed words, and usually assigns each word a score. This score is then used to calculate the sentiment expressed. However, due to the nature of this approach, it is often done by hand, and constant updates need to be made depending on the evolution of a language.

1.2.2 Machine Learning Approach

There are two approaches to sentiment classification in machine learning, the first being supervised learning. Supervised learning methods are used for datasets with labels, whereas unsupervised learning methods are used for datasets without labels. There are numerous supervised classifiers, the most common of which is the Naive Bayes classifier. Supervised

learning methods are applied to labeled datasets, thus, these learning algorithms frequently require classification algorithms. Classification algorithms are used to assign test data to specific categories accurately. It finds specific entities in the dataset and draws conclusions about how those entities should be labeled or defined. In the case of YouTube videos, these classifications are the viewers' positive, neutral, or negative sentiments. Finally, the machine learning algorithms train a subset of a very large set of data, and based on that data gathered can be used to produce sentiment analysis on actual test and real world data.

1.2.2 Hybrid Approach

Finally we have hybrid approaches, which use a combination of techniques both in the Lexicon and Machine Learning Approaches in order to reduce the limitations of the approaches (Fang et al., 2015).

2. Materials and Methods

The algorithm used to identify sentiment in this paper utilizes a hybrid approach to identify the sentiments of comments in Youtube videos.

2.1 Data Acquisition

For data acquisition, the dataset was taken from a public youtube scraper set on kaggle (https://www.kaggle.com/datasets/datasnaek/youtube?select=UScomments.csv). The set of all the videos and comments on each video was extracted from above as a csv, with the video csv file providing the video ID, video title, channel title, category ID, tags, views, and number of likes, and the comment csv providing, video ID, comment, likes and replies.

For the types of videos which we trained and tested the data on, the tests were specifically targeted at covering videos about films, and television. As a result we narrowed the dataset from twenty six categories down to category one, which are youtube videos tagged with relation to films and television shows. To retrieve only the comments of said videos, we filtered all the comments based on video ID belonging to category one for a total of 17,512 comments.

2.2 Data Classification

Next, we labeled the data as either neutral, positive or negative. As discussed before, youtube comments are not labeled with a review/grading of the video. Since we intend to use a Naive-Bayes algorithm to calculate sentiment analysis, we used a classification algorithm to assign the labels to each comment. rule-based lexicon approach Sentiment VADER (Valence Aware Dictionary for sEntiment Reasoning) is used to generate the output, and categorizes the comments as positive, neutral, or negative. In this paper the Sentiment VADER algorithm was implemented via the python package SentimentIntensityAnalyzer, which determined whether an individual's comment was positive, negative, or neutral, through the polarity scores assigned. Cut-off values for positive sentiments were polarity scores greater than 0.2, and negative sentiment values were scores less than -0.2. Scores with a value of in between were given a neutral value.

2.3 Preprocessing and Training

After that, we split the data into a 70:30 split, for both training and testing respectively. The data is then preprocessed for the Naive Bayes algorithm. To begin, because the algorithm only deals with positive and negative sentiments, every neutral sentiment was removed in preparation for the Naive Bayes algorithm. After that, the training data is further split into training and development set using ten-fold cross validation. The training and test data are then pre-processed separately in order for the Bayes algorithm to work more effectively.

2.3.1 Stop words

First, the algorithm identifies stop words (words with no polarity: the,then,as,etc.), which are then removed from the training data set using the WordNet package for English words, as such words are not helpful for analyzing sentiment. The algorithm also recognizes the top ten most frequently used words and removes them from the training set. These words occur very frequently in the comments, but don't provide the algorithm with much information related to the topic and may cause noise when the algorithm analyzes data. The following displays the top ten most frequent words removed from the data set.

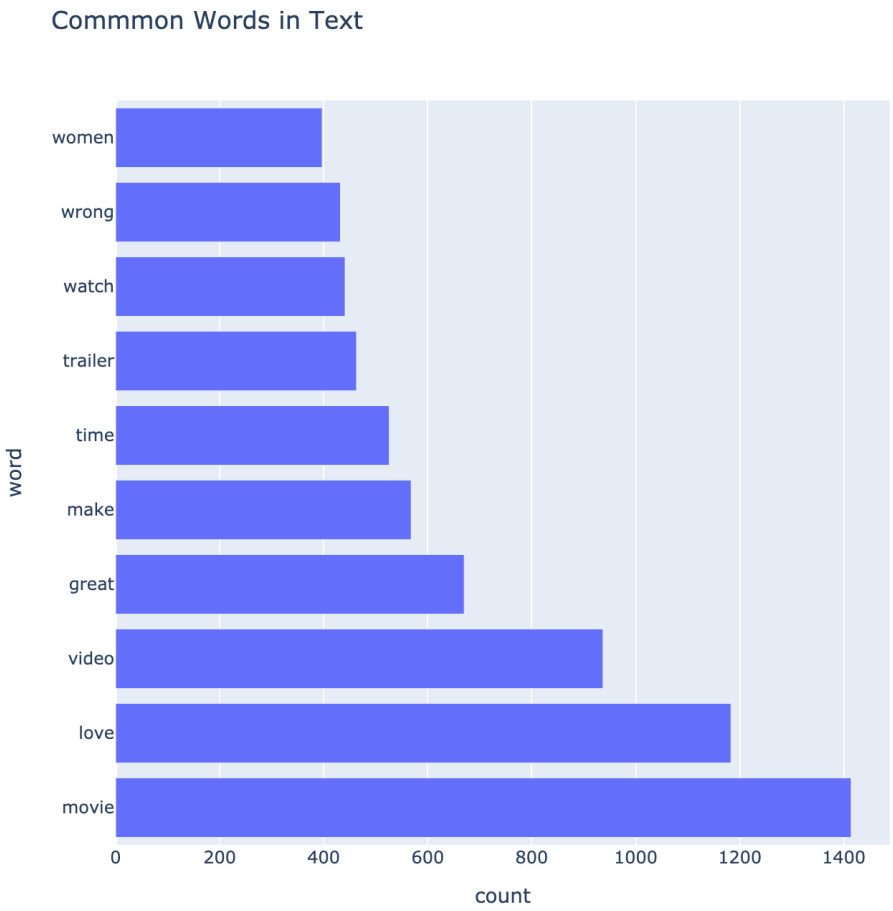


Figure 1. top ten common words in training data

2.3.2 Stemming

Stemming is introduced to identify the root of each word. The Porter Stemmer algorithm is used to accomplish this, which reduces the time required by the algorithm to train all tenses of the word to be positive or negative. Finally, the pre-processed data is fed into the Naive Bayes algorithm, which categorizes the sentiment of each comment.

2.3.3 Tokenization

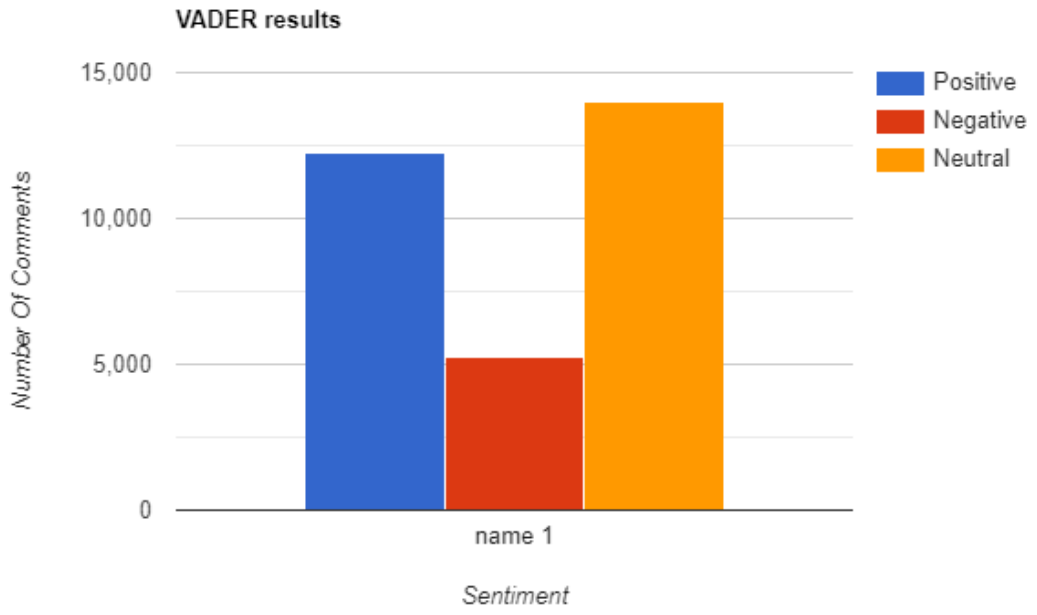
In the last step of preprocessing the data, the data is tokenized. Due to the difficulty of training on an entire comment, the algorithm tokenized each word so that the Naive Bayes algorithm could train on each word rather than each comment.

2.3.4 Undersampling

To remove bias from the Naive Bayes model (Positive: 8593, Negative: 3664) undersampling is performed on the training set (after separated from the development set in 10-fold cross-validation). The undersampling method we chose for the model is Near-Miss 1.

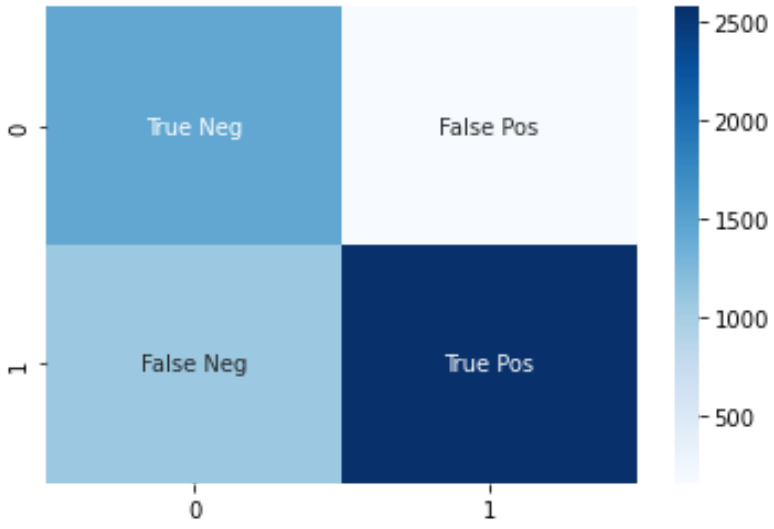
3. Results and Figures

The outcome of using the Sentiment VADER on the dataset. The number of positive, negative, and neutral values in the labeled data is counted. The neutral elements are removed when testing and training data.



Once properly trained, the Naive Bayes algorithm is finally run on the test set. The Naive Bayes classifier is trained using the preprocessed data in the development set as discussed

before. The classifier is then applied to the testing dataset, and its accuracy is found to be 76.36%, Precision of 94.15%, Recall of 70.43%, and an F1 score of 80.58%.



Accuracy	.7636086791016369
Precision	.9415418341249543
Recall	.7042907898332877
F1-score	.8058161350844278

4. Conclusion

A hybrid machine learning and lexicon approach can be effectively used to analyze comments on YouTube videos and determine the sentiment of viewers. By using Sentiment VADER as a baseline and applied data preprocessing techniques, such as tokenizing and removing stopwords, the accuracy of the sentiment analysis model can be improved. Applying the Naive Bayes algorithm will classify the sentiment of comments as positive, negative, or neutral. Through this approach, it is possible to achieve an accuracy of accuracy of 76.36%, Precision of 94.15%, Recall of 70.43%, and an F1 score of 80.58%.

The ability to accurately analyze the sentiment of comments on YouTube videos can be valuable for creators, as it allows them to understand the positive or negative feedback they are receiving from viewers. This information can be used to improve the quality of their videos and tailor them to their audience's preferences. It can also be helpful for viewers, as it allows them to understand the overall sentiment of a video without having to manually read through the

comments. This can save time and provide a more concise overview of the public opinion on a particular video.

Overall, the aforementioned approach to sentiment analysis for YouTube comments shows promise in providing a useful tool for both creators and viewers. Further research and development in this area may lead to even more accurate and efficient methods for determining the sentiment of YouTube comments.

5. Acknowledgements and Limitations

Since the dataset (UScomments.csv) we chose is filtered for one video category, the imbalance between the number of positive and negative comments is significant, which exists both in the training and the testing set. The Naive Bayes classifier we constructed is supposed to be bias-free due to the usage of undersampling, which means the imbalance in the training set is removed before use. Therefore, the model is later tested on a set which has a different distribution than its training set, thus resulting in a higher false negative number (less favoring the majority class).

6. References

A. Amin, I. Hossain, A. Akther and K. M. Alam, "Bengali VADER: A Sentiment Analysis Approach Using Modified VADER," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679144.

Behdenna, S., et al. “Document Level Sentiment Analysis: A Survey.” *EAI Endorsed Transactions on Context-Aware Systems and Applications*, vol. 4, no. 13, 2018, p. 154339., https://doi.org/10.4108/eai.14-3-2018.154339.

Fang, X., Zhan, J. Sentiment analysis using product review data. *Journal of Big Data* 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

Owoputi, Olutobi, et al. "Improved part-of-speech tagging for online conversational text with word clusters." *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2013.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." *arXiv preprint cs/0205070* (2002).

Talbot, Ruth, Chloe Acheampong, and Richard Wicentowski. "Swash: A naive bayes classifier for tweet sentiment identification." *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015.

Uma, J., and Dr K. Prabha. "Machine Learning Algorithm for Sentiment Analysis in Twitter Data." *Int. J. of Aquatic Science* 12.3 (2021): 640-651.