Advances in Computer and Communication, 2025, 6(1), 35-40
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Sentiment Analysis on YouTube Data: A Comparison of TextBlob and VADER

Haowen Zhang

University of California, San Diego, San Diego, CA 92092, USA.

How to cite this paper: Haowen Zhang. (2025) Sentiment Analysis on YouTube Data: A Comparison of TextBlob and VADER. *Advances in Computer and Communication*, 6(1), 35-40.
DOI: 10.26855/acc.2025.01.006

Received: January 12, 2025
Accepted: February 7, 2025
Published: March 4, 2025

*Corresponding author: Haowen Zhang, University of California, San Diego, San Diego, CA 92092, USA.

Abstract

This article tried to compare the performance of TextBlob and VADER on analyzing comments under YouTube videos and measured the correlation between using sentimental analyzers to predict the quality of videos and using the like/views ratio to compare the quality of videos. People are becoming busy, and they rely more on watching videos to have fun, acquire new knowledge, and exchange thoughts. Finding a way to estimate the quality of the videos becomes important. 18000 comments are collected from 60 videos from 6 categories. TextBlob and VADER are used to analyze these comments. Scores are given by these two sentimental analyzers. The scores of each video are compared to the likes/views ratio to predict the quality of the videos. Charts are made to analyze the trends. The result is that TextBlob is more suitable for this model because the scores it gives are more stable and have a positive correlation to the likes/views model.

Keywords

Sentiment analysis; Data processing; Correlation analysis

1. Introduction

The amount of digital data being produced, in real time, has been exploding at an unknown rate [1]. People are using phones to chat with each other, watching videos on the internet, working on the internet, and so on. Every tap on the phone and computer will construct a transmission of data. The society creates a huge amount of data flow every day. Large data leads us to a busy time. People have limited time and unlimited wants. Therefore, it is important for the users to do things efficiently. More and more people love watching videos instead of reading books because watching videos is a more efficient way to acquire information. The video producers help their viewers summarize their thoughts. Students can absorb most of the knowledge in the textbooks within hours by watching videos, while it takes weeks if they read thick books. Watching videos has become an ideal way to study and relax. Helping users to find high quality video is becoming more and more important.

YouTube removed the public dislike counter from all videos on the 10th of November 2021 to protect the video producers. This makes it harder for the users to decide whether a video is good or not. The users need to watch part of the video in order to know the quality of the video. This takes time and is not efficient. Therefore, we need other methods to let the computer judge the quality of the video for us.

The ratio of number of likes and number of views can be used to predict the quality of videos. The higher the ratio, the more likely the video has high quality.

Comments can be used to predict the quality of videos. Other users express their thoughts towards the video. Sentiment analysis can be used to analyze the comments. Sentiment analysis would analyze the given content, which might be an opinion about some entity and would pertain this content to be either positive or negative in nature, and sometimes it is also known as opinion mining. It will analyze the words in the comments and can give the polarity of the sentences. The resulted data can be used to analyze the attitude of other users and predict the quality of the videos.

However, it's difficult to judge whether a video is good or not because it is subjective. We can only compare the methods to see whether they have a correlation with each other.

2. Related Works

Sentiment analysis is one of many areas of computational studies that deal with opinion oriented natural language processing [2].

The sentiment analysis can be categorized into lexicon sentiment analysis, machine learning-based sentiment analysis, and hybrid techniques. The lexicon sentiment analysis relies on the polarity of words in a given text. The machine learning-based sentiment analysis represents a document with a feature set and performs classification on the feature set to predict the polarity of the document [3].

It has turned out as an exciting new trend in social media with a gamut of practical applications that range from applications in business, applications as subcomponent technology to applications in politics [4]. Social media became more and more prevalent in society. Therefore, analyzing how the public feels about certain products, topics or people has become more and more important. Sentiment analysis is mostly used in analyzing comments under the videos.

There are challenges to using sentiment analysis on social media [5]. Grammar rules are not always followed on social media, and abbreviations and Internet slang are always used. This means that the rulesets need to be kept up to data. The model needs to be improved frequently.

2.1 Experiment Setup

The data used in this research cannot be found from any existing datasets because we need to use the comments, number of views, and likes. There are YouTube comments datasets on Kaggle, but the datasets do not contain the name and URL of the videos. So, Python programs were written to fetch the data. After the comments were fetched, sentiment analyzers were used to evaluate each of the comments and give scores to them. After that, the overall score of the video will be calculated. The overall score of the video will be compared to the likes/views ratio to see whether the model has a positive or negative correlation to the likes/views model.

2.2 YouTube

YouTube is a global video-sharing platform. It was launched in 2005 by three PayPal employees and became a sentral hub for video content quickly [6]. It allows users to upload, view, and share videos. It also allows users to leave comment under the video to express their thoughts and attitudes towards the videos. The language used in the comments is very fashionable and up to date [7].

3. Data

3.1 Data Description

The data are separated into three parts: the comments under the videos, the views and likes of the videos. There are several steps to fetch the data for this experiment. The first step was to choose the videos to analyze. We choose the top 10 most viewed videos for each of the video types [15]. The video types are: #Music, #Gaming, #Podcasts, #Rhythm & Blues, #Thoughts. We record the links of the videos and store them in an Excel file. The comments of the video will be fetched using the YouTube API. Since we choose the top 10 most viewed videos for each of the video types, there must be a lot of comments to analyze. It takes too much time to analyze. Therefore, we only choose 300 of the comments for each of the videos.

3.2 Data Collection

The data were collected from the YouTube video platform by using YouTube's API. Web scraping is used to fetch the comments, username, time, number of views, and likes of the YouTube videos. The comments are stored in a dictionary with the links of the videos being the key of the dictionary. Short words with a character length of less than three are removed from the sentences [8]. Stop words are also removed [9]. Then a sentiment analysis tool such as TextBlob and VADER is used to analyze whether the comments are positive, negative or neutral. Then the result data are stored in a dataframe in an excel. Some of the videos have closed the comments section, so other videos were selected instead.

3.3 Data Processing

The dataframe has 11 columns. They are the types of videos, links of the videos, title of the videos, number of views of the videos, number of likes of the videos, tb_positive (comments analysis by textblob and shows positive), tb_negative (comments analysis by textblob and shows negative), tb_neutral (comments analysis by textblob and shows neutral), vd_positive (comments analysis by VADER and shows positive), vd_negative, vd_neutral. Then the column likes/views (%) is added by dividing the number of likes by the views. This means the percentage of viewers that like this video and can somehow show the quality of the videos. The higher the percentage, should mean that the video has a good quality. We also sum up all the scores of the comments of the video and take the average to get the score of the video. The formula is (sum of the scores of the comments)/number of comments. The correlation coefficient between the scores of the video given by TextBlob and the like/views ratio will be calculated using the python numpy tool. The correlation coefficient of the scores of the video given by VADER and the like/views ratio will also be calculated.

3.4 Method and Models

The experiment was conducted in PyCharm [9]. Data processing and visualization were done through Excel and Matplotlib. The requests library was used to fetch the comments and number of views and likes. Two sentiment analyzers were used. They are TextBlob and VADER.

TextBlob is one of the tools to process textual data. It is commonly used for tasks such as part-of-speech tagging, noun phrase extraction and sentiment analysis. It is a Python library used to process language data. It measures the polarity of a text and returns a score ranging from -1 to 1 with -1 being very negative and 1 being very positive and 0 being neutral [10]. Each word is assigned a score that indicates whether it is positive or negative or neutral based on its meaning. TextBlob measures the word frequency, intensity and semantic relations between the words [11]. It treats the text as a bag of words and assigns scores to each of the words and calculates the overall sentiment score. After TextBlob assigns scores to each word, it calculates the overall score by taking the average of the scores of the words [12].

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically designed to analyze text from social media and other informal writing sources [13]. Developed by C.J. Hutto and Eric Gilbert in 2014, VADER is especially attuned to the sentiment expressed in social media contexts and is capable of handling various aspects of natural language, including slang, emoticons, capitalization, and punctuation. It also assigns a score ranging from -1 to 1 to the text with -1 being negative, 1 being positive and 0 being neutral, but it is different from TextBlob. It first assigns a valence score to each of the words in the text and then computes the compound score by summing up and normalizing the valence scores. It assigns emotion intensity to the words in the text and sums up the scores. It also returns a dictionary that contains three different scores (positive, neutral, negative) and the compound score of the text. The compound score is calculated by normalizing the three individual scores [17].

4. Results

TextBlob and VADER are two different sentiment analyzers [16]. The overall trend is shown in the table below. All the comments are merged into a single file, and we calculate the average and standard deviation of the scores TextBlob and VADER give to all the comments. The scores given by VADER have an average of 0.2041 and standard deviation of 0.4331 while the scores given by TextBlob have an average of 0.1 and standard deviation of 0.2697. This means that VADER might give a higher score to a given text, but with larger fluctuation. VADER assigning emotion intensity to the words in the text might cause this difference. It tends to assign more extreme scores to words that are very positive or very negative, and when many positive/negative terms are used, the overall sentiment score becomes more extreme. TextBlob is more conservative and often balances out scores with neutral language. It does not over-emphasize emotional punctuation or slang, resulting in sentiment scores that are often more moderate and less extreme, producing a smaller standard deviation.

Table 1. Performance of TextBlob and VADER on analyzing the comments table

	TextBlob	VADER
Average	0.1003901785	0.2040694583
Standard deviation	0.2697471562	0.4331091647
Min	-1	-0.9996
Max	1	1

The table below shows the correlation between the scores given by TextBlob and VADER.

Table 2. Correlation between results from TextBlob and VADER table

	TextBlob	VADER
TextBlob	1	0.4611112179
VADER	0.4611112179	1

Then we calculate the sum of all the scores of the comments of the videos. Then divides by 300 to evaluate the score of the video. The graph below shows the scores of the videos using sentiment analyzers and using the likes/views ratio.

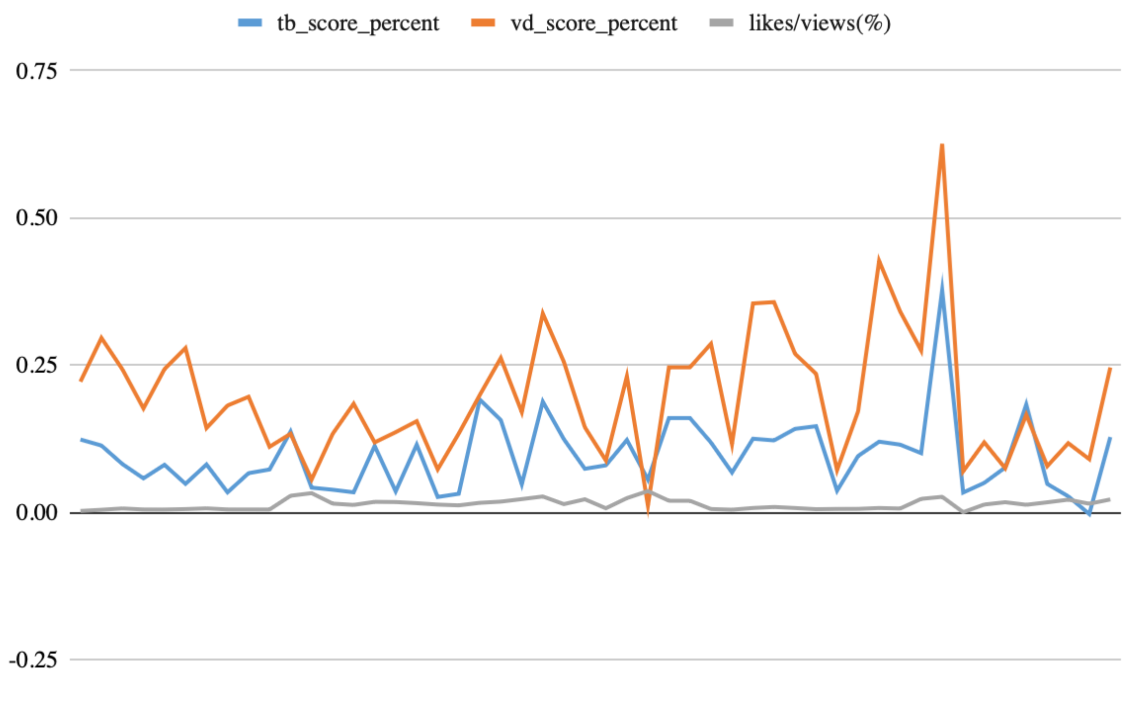

Figure 1. Line chart comparing performance of TextBlob and VADER and likes/views ratio.

The correlation coefficient between each of the sentiment analyzers and likes/views ratio is also calculated. The scores of the videos calculated by TextBlob and the likes/views ratio have a correlation coefficient of 0.21336633492498294. The scores calculated by VADER have a correlation coefficient of -0.07012087724065634. This shows that TextBlob has a positive correlation to the likes/views model, while scores calculated by VADER have a negative correlation to the likes/views model. Thus, TextBlob is more suitable for this model because the scores it gives are more stable and have a positive correlation to the likes/views model.

5. Discussion

The result shows that using the TextBlob as the sentiment analyzer is better than VADER for this model, but there are some limitations. Firstly, the likes/views model is not perfect because we do not know the exact number of dislikes of the videos. Some of the videos may have little neutral viewers. All the viewers will choose to like the video or dislike the video. The video may have a larger like/views ratio than the other videos, but the rest of the viewers dislike this video. We still cannot identify this video as good. Secondly, not all users like to leave comments under the videos. Only a small percentage of the users will engage with every video they watch. Also, if a user decides to leave a comment under a video, they must have a strong subjective feeling to the video. This makes the scores of the comments to be extreme. However, these viewers who leave extreme comments may watch and analyze the videos more carefully than the other viewers. Therefore, their comments are more valuable and should be considered more. In this case, using the VADER sentiment analyzer may be better. Finally, although the TextBlob model has a positive correlation to the likes/views ratio model, the correlation coefficient is small. This shows that the correlation between them is not strong.

This research can be improved in the future. Firstly, we can use advanced mathematical statistics models and past data to predict the dislike numbers of the videos and then compare the accuracy of different sentiment analyzers. Secondly, the number of comments is small. More comments from more types of videos will be fetched and joined in the model in the future. Finally, we may need to judge the content of the comments with the help of AI because some of the comments are not related to the videos at all, and this will affect the overall score of the videos.

References

[1] Devakunchari R. Analysis on big data over the years. Int J Sci Res Publ. 2014 Jan;4(1).

[2] Kumar A, Sebastian TM. Sentiment Analysis: A Perspective on its Past, Present and Future. Int J Intell Syst Appl. 2012;4(10):1-14. https://doi.org/10.5815/ijisa.2012.10.01

[3] Aljedaani W. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Knowl-Based Syst. 2022 Nov 14;255:109780. https://www.sciencedirect.com/science/article/pii/S0950705122009017.

[4] Rochadiani TH. Sentiment Analysis of YouTube comments toward Chat GPT. J Transformatika. 2023;21(1). https://doi.org/10.26623/transformatika.v21i2.7033.

[5] Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowl Inf Syst. 2018;60(2):617-663. https://doi.org/10.1007/s10115-018-1236-4.

[6] Jones T, Cuthrell K. YouTube: Educational Potentials and Pitfalls. Comput Sch. 2011;28(1):75-85. https://doi.org/10.1080/07380569.2011.553149.

[7] Berger I. YouTube as a source of data. PsyPAG Q. 2012 Jun;(83):9. https://d1wqtxts1xzle7.cloudfront.net/30868316/PsyPAG-83-June-2012-libre.pdf?1392134977=&response-content-disposition=inline%3B+filename%3DYoutube_as_a_source_of_data.pdf&Expires=1731808685&Signature=bPJFa7SAMZrtdGABzgyDv5J6eSlqrJ3q4fH6CpFdZnN30GGktyLAX-Mi4cubRk-KB8eQJxY93d64UnT~xZ9PK~qfSacsj4hjtggc6SwMQZfmMhzvzP85eFPku2CedKCr-Jv5cyz2FNG7~9MqR8a-A6Kr7ElQMFjOF~x0tziLvvjC~L1cCjyifrwJzqa3ztQ-mA0~C-myuMNeFJLL1~o-OhzJbwvZWV7ygmMlT3~ExDghcFPPm0Xnu~IZbwjAuOxQ5OAIXiMXrgbmOrhDwcDlbG5O6c22KAHSrvtAzTYTzqH85bSuSFB-wZc25uZWWMHQBQqkcK58L-rVQvgIyZk8xg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA#page=11.

[8] Umer M. Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. Comput Intell. 2021 Feb;37(1):409-434. https://onlinelibrary.wiley.com/doi/full/10.1111/coin.12415.

[9] Nguyen TH, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction. Expert Syst Appl. 2015;42(24):9603-9611. https://doi.org/10.1016/j.eswa.2015.07.052.

[10] Chandra RV, Varanasi BS. Python requests essentials. Packt Publishing Ltd; 2015.

[11] textblob.readthedocs.io [Internet]. Available from: https://textblob.readthedocs.io/en/dev/.

[12] Aljedaani W, Rustam F, Mkaouer MW, Ghallab A, Rupapara V, Washington PB, et al. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Knowl-Based Syst. 2022;255:109780. https://doi.org/10.1016/j.knosys.2022.109780.

[13] Marsanich G. Comparing Predictions of YouTube Video Like to Dislike Ratios Using Sentiment Analysis Tools [thesis]. Tilburg University; [cited 2023]. Available from: http://arno.uvt.nl/show.cgi?fid=161975.

[14] Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for. In: Eighth Int AAAI Conf Weblogs Soc Media. 2014. p. 18. Available from: https://ojs.aaai.org/index.php/icwsm/article/view/14550.

[15] Rousidis D, Koukaras P, Tjortjis C. Social media prediction: a literature review. Multimed Tools Appl. 2020;79(9-10).

[16] Abiola O, et al. Sentiment Analysis of Covid-19 Tweets from Selected Hashtags in Nigeria Using Vader and Text Blob Analyser. J Electr Syst Inf Technol. 2023 Jan 16. https://doi.org/10.1186/s43067-023-00070-9.

[17] Devi BL, Bai VV, Ramasubbareddy S, Govinda K. Sentiment analysis on movie reviews. In: Advances in Intelligent Systems and Computing. 2020. p. 321-328. https://doi.org/10.1007/978-981-15-0135-7_31.