Sentiment Analysis on Youtube comments using Numerical Methods

Alejandro Villada Toro
avilladat@eafit.edu.co

Cristian Alzate Urrea
calzateu@eafit.edu.co

Harold Steven González Ossa
hsgonzaleo@eafit.edu.co

Simón Álvarez Ospina
salvarezo1@eafit.edu.co

November 21, 2021

Abstract

Sentiment analysis, or opinion mining, is one of Natural Language Processing's techniques related to evaluating emotions and attitudes users express on online resources such as social media. In recent years, the popularity of social media has increased significantly and consequently has attracted many users towards video sharing sites, such as YouTube. In the present work our approach to understand this phenomenon through the study of different songs' comments using diverse classifier methods is given.
Keywords: *Numerical methods; Sentiment Analysis; Natural Language Processing; Machine Learning.*

1 Introduction

The modern world brings with it a huge amount of new information every day, and social media is one of its major contributors, so it is important to analyze what advantages and disadvantages it brings and how it can affect people's lives. Some researchers have tried to understand social media's impact, like Wang, McKee, Torbica, and Stuckler (2019), who talk about how misinformation is often more popular than reliable information and the topics where it's possible to meet with it more frequently. On the other hand, Akram and Kumar (2017) show how convenient social media is by knowing consumer's opinion about products and widening market's vision.

Therefore, it's essential to understand how people's feelings are affected due to their interaction with platforms like Youtube. Some researchers have shown the influence of music in people, such as Chen, Zhou, and Bryant (2007), who talk about how music can affect people's mood: it was found that sad people tend to listen to songs that accompany this feeling and eventually begin to listen to more upbeat songs, which shows that listening to sad songs helped them to cope with that sadness. In our approach we take comments from different websites and apply different classification techniques to comprise them into indicators easier to comprehend so the manipulation of this information becomes simpler.

1.1 Theoretical framework

Natural Language Processing (NLP) is a set of techniques whose intention is to propose algorithms to analyze human language. There are many applications of this topic in different fields, such as the information field, where scientists use NLP methods to analyze polarity and subjectivity of opinions people have about different subjects and to make conclusions. Other applications consist in making robots interact with humans or building tools to facilitate diverse activities with a writing component.

In this paper, we aim to use NLP to make a sentiment analysis on Youtube comments using different numerical methods such as the logistic regression and Naive Bayes (NB) criteria for classification. Suárez and Monroy (2020), Kaewpitakkun, Shirai, and Mohd (2014) and Saif, He, and Kundi (2011) have done different approaches to the sentiment analysis on Twitter. Suárez and Monroy (2020) implemented a random forest algorithm, an NB algorithm and a Support Vector Machine (SVM) algorithm to identify the impact on people's opinion about a colombian institution based on the opinions of the tweets and found that the Random Forest Algorithm obtained the best results with a precission of 74.56%.

Kaewpitakkun et al. (2014) and Saif et al. (2011) proposed a lexicon interpolation method and a semantic smoothing method with Naive Bayes classification method respectively. They found that their algorithms had an accuracy around 75%. Leonardo (2019) used, besides the previous methods, a neural network algorithm to make a sentiment analysis on social media in general. He concluded that it doesn't matter which algorithm is used because they all have a similar performance, however it is very important to make a good processing of the texts.

There are some previous works that made a sentiment analysis with Youtube comments. Tehreem and Tahir (2021) implemented five algorithms to classify the positivity of Youtube comments in Roman Urdu language, finding an accuracy of 64% using an SVM algorithm. On the other hand, Asghar, Ahmad, Marwat, and Kundi (2015) found some problems when they applied different NLP algorithms to Youtube comments of a particular video, such as colloquial speaking and the use of different languages.

Traditionally, an approach to make language analysis is to decompose the process in five stages: text preprocessing to understand how the phrases could be segmented, lexical analysis for sequences of words because the words are not atomic and the union of two words could have a different meaning with one of the words, syntactic analysis to identify the structure of the sentences, semantic analysis to understand their meaning and, finally, pragmatic analysis to make conclusions over the phrases given the context due to ambiguous expressions (Indurkhya and Damerau (2010)).

NLP proposes different algorithms able to do this analysis, using techniques such as tokenization to do the text preprocessing, eliminating "stop words" by doing a previous syntactic analysis and discarding the words that are not necessary to understand the meaning of a sentence, and "stemming" to transform words into their roots (González (2021)).

Many numerical classifier methods are used for sentiment analysis such as logistic regression that consists of, when given a value for a variable, the probability of occurrence for another variable is given, and the last variable returns a probability between 0 and 1. For this method two numerical techniques are used: Implementing a gradient ascent's method where the program tries to approach to the maximum log-likelihood using little steps in the gradient's direction, while the second technique is about approaching towards the maximum with Newton's method using the log-likelihood's Hessian inverse matrix. There are other methods like Support Vector Machine, that constructs a hyperplane or a set of hyperplanes that separate the dataset with one condition: if the data is not linearly separable, the algorithm allows the use of Kernel functions that enable them to become linearly independent; and Naive Bayes that uses features to classify data supposing strong independence between them.

This article describes the implementation of a sentiment analysis algorithm for Youtube comments and compares the effectiveness of different classification models that use the numerical methods described above. Besides, a brief analysis to a video is made according to the results of the most efficient model of those studied.

1.2 Objectives

Make an optimized and effective NLP algorithm to classify Youtube comments and measure the positivity of Youtube videos.

The specific objectives are:

- Implement different classification algorithms and find the most efficient by comparing other methods.
- Analyze the possible relation between the videos' perception and popularity.
- Propose a scheme to describe the feeling conveyed in a video according to its perception and popularity.

2 Methodology

This section explains how comments from YouTube videos are extracted and processed for classification. In addition, the libraries used for the implementation of the numerical

methods are mentioned.

2.1 Comments extraction and preprocessing

Google offers a wide range of services for programmers; among them is access to the APIs of the different platforms it handles. To extract the comments from a YouTube video, the YouTube API offered by Google is used and, as not all the information about the comments is needed, only some of the characteristics of each one is stored, such as its identification, content and number of 'likes'. This process is supported by Pandas library for Python.

2.2 Cleaning and classification of comments

The TextBlob library for Python is used to classify the extracted comments. To make use of it, it is necessary to clean up the comments first by eliminating emojis and characters such as commas and periods. Demoji library and the Python Regular Expressions module are excellent for this task. In addition, as TextBlob only classifies texts in English, it is necessary to filter the comments, so the Language Detection library for Python is used.

TextBlob measures the polarity of the comments (that is, how positive or negative they are) using values between -1 and 1. To facilitate the use of numerical methods, comments are divided into two categories based on the polarity: if the polarity is less than or equal to 0, then the comment is negative and is represented by '-1', otherwise the comment is positive and is represented by a '1'.

Finally, the Natural Language Toolkit library for Python is useful to delete the 'stopwords' from the comments. The dataset is divided into two sets: training and testing; those will be used in the implementation of numerical methods to classify new comments.

2.3 Vectorization of comments and implementation of numerical methods

Because texts are not quantitative data, using numerical classification methods with them is a pointless task, therefore the comments are then vectorized (that is, converted to vector data) using the Term frequency - document inverse frequency (Tf-idf) method (Stecanella (2019) explains this technique in detail).

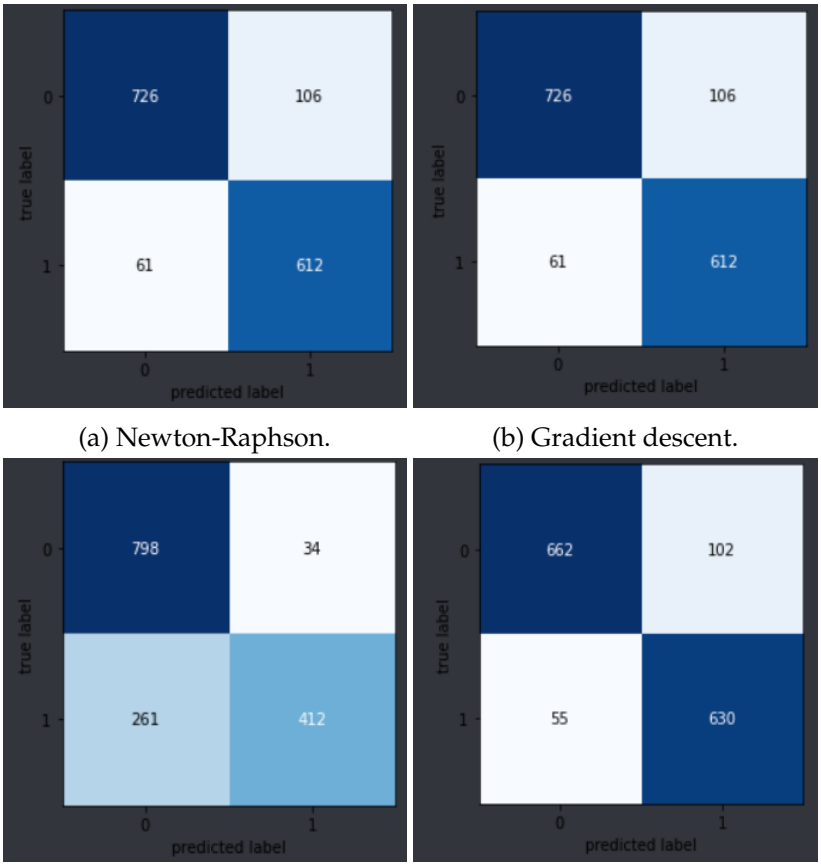
The SciKit Learn library for Python brings the tools to vectorize the comments and implements the logistic regression, Naive Bayes and SVM methods to build the classification models compared.

2.4 Results comparison

Comments from 5 Youtube videos are extracted and processed following the steps described above[1]. Then, the model is tested with the comments from a different video. The resulting precision for the models implementing the different numerical methods are: 88.9% for Logistic regression with gradient descent, 88.9% for Logistic regression with Newton-Raphson, 80.3% for Naive Bayes and 92.2% for SVM.

It is clear that the model using SVM is the most efficient of those studied according to the results, but the efficiency of the other methods is not devalued since they also yield very good results.

The confusion matrices obtained from the different methods are shown in the Figure 1.



(a) Newton-Raphson. (b) Gradient descent.

(c) Naive Bayes. (d) SVM.

Figure 1: Confusion matrices of the results obtained from the studied models.

[1]The videos used are in the bibliography (see Passenger (2012), Roses (2009), Robinson (2014), Cash (2018), Gray (2020) and Odell (2012))

3 Brief analysis of Youtube videos perception

The impact of a Youtube video is measured with the equation:

$$\frac{Number-of-positive-comments+Number-of-likes-of-positive-comments}{Number-of-comments+Number-of-likes-of-comments}$$

And the popularity is defined by:

$$\frac{Number-of-likes-of-the-video}{Number-of-votes}$$

The impact and popularity of the video used for tests is calculated. It is obtained that the impact is 0.99 and the popularity is 0.9704. To make statistical affirmations about the relation of these measures, it is needed to analyze more videos that is left to further research.

A hypothesis proposed to describe the feelings conveyed in a video relate these measures. If the impact is high and the popularity is high, then it could be concluded that the video convey joy, satisfaction, etc. If the impact is not high and the popularity is high, then it could be concluded that the video is controversial and convey many diferent emotions. Finally, if both of the measures are not high, then the video do not convey strong feelings or it is not interesting.

4 Conclusions

In this research, an algorithm to make sentiment analysis to Youtube comments is proposed. The effectiveness of three classifications models is compared according to the precision obtained with the numerical methods used (logistic regression, Naive Bayes and SVM). Implementing SVM to the algorithm returned the best results from the methods studied. In further research, it is recommended to study the precision of other methods such as random forests or combinations of the methods used in this article.

A study to the impact of a Youtube video is realized. There is not enough evidence to think that the impact of the video affects its popularity, so a scheme to relate these measures is proposed as a method to describe the feeling conveyed in the video. However, it is subjected to the authors opinion, so it requires a more rigorous research.

References

Akram, W., & Kumar, R. (2017). A study on positive and negative effects of social media on society. *International Journal of Computer Sciences and Engineering*, *10*(5), 351-354.

Asghar, M. Z., Ahmad, S., Marwat, A., & Kundi, F. M. (2015, 11). Sentiment analysis on youtube: A brief survey. *MAGNT Research Report*(1), 1250-1257.

Cash, F. (2018). *flora cash - you're somebody else (lyric video).* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=qVdPh2cBTN0

Chen, L., Zhou, S., & Bryant, J. (2007, 12). Temporal changes in mood repair through music consumption: Effects of mood, mood salience, and individual differences. *Media Psychology*, *12*(1), 695-713.

González, I. (2021). *Procesamiento del lenguaje natural | aplicaciones y conceptos.* Retrieved 2021-10-10, from https://youtu.be/mRQORkK3ZYk

Gray, C. (2020). *Conan gray - heather.* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=24u3NoPvgMw

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing* (2nd ed.). Chapman & Hall.

Kaewpitakkun, Y., Shirai, K., & Mohd, M. (2014, December). Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In *Proceedings of the 28th pacific asia conference on language, information and computing* (pp. 204–213). Phuket,Thailand: Department of Linguistics, Chulalongkorn University. Retrieved from https://aclanthology.org/Y14-1026

Leonardo, P. D. (2019). *Análisis de sentimientos: aplicación sobre textos en redes sociales.*

Odell, T. (2012). *Tom odell - another love (official video).* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=MwpMEbgC7DA

Passenger. (2012). *Passenger — let her go (official video).* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=RBumgq5yVrA

Robinson, P. (2014). *Porter robinson - goodbye to a world (official audio).* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=W2TE0DjdNqI

Roses, G. N. (2009). *Guns n' roses - november rain.* Retrieved 2021-11-10, from https://www.youtube.com/watch?v=8SbUC-UaAxE

Saif, H., He, Y., & Kundi, H. A. (2011, 10). Semantic smoothing for twitter sentiment analysis. *10th International Semantic Web Conference*(1), 1250-1257.

Stecanella, B. (2019). *Understanding tf-id: A simple introduction.* Retrieved 2021-11-10, from https://monkeylearn.com/blog/what-is-tf-idf/

Suárez, E., & Monroy, A. F. (2020). *Implementación de un modelo de análisis de sentimientos con respecto a la jep basado en minería de datos en twitter.*

Tehreem, T., & Tahir, H. (2021). *Sentiment analysis for youtube comments in roman urdu.* Retrieved 2021-10-10, from https://arxiv.org/abs/2102.10075

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science*

& Medicine, 240, 112552. Retrieved from https://www.sciencedirect.com/science/article/pii/S0277953619305465 doi: https://doi.org/10.1016/j.socscimed.2019.112552