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

October 10, 2021

Abstract

Pending...
Keywords: *key1; key2; key3; key4.*

1 Introduction

he 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 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 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; Support Vector Machine 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 independence between them.

This article describes the implementation of different algorithms that use these numerical methods, explaining the tools used and their applications; then, a comparison between the properties of the algorithms, such as their effectiveness, is made.

1.2 Objectives

The general objective is to make an optimized and effective NLP algorithm to classify Youtube comments and measure the positivity of Youtube videos.

The specific objectives are:

- To build a dataset with the Youtube comments using Youtube API and tagging manually the polarity to each comment.
- To implement different classification algorithms and find the most efficient by comparing other methods.
- To analyze how the people's perception can affect Youtube videos popularity.

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.

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`

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.*

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

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