SOCIOECONOMIC ANALYTICS

Predicting sentiments in Spotify comments: A comparative analysis of Machine Learning Models

Filipe Augusto Felix de Queiroz [1]*
Igor Barbosa Negreiros [2]
Giovana de Souza [3]
Débora Cordeiro de Sousa [4]
Sílvio Fernando Alves Xavier Júnior [5]

[1 2 3 4 5] Departamento de Estatística, State University of Paraíba, Campina Grande, Brazil
Emails: [1] filipe.queiroz@aluno.uepb.edu.br; [2] igor.negreiros@aluno.uepb.edu.br;
[3] giovana.souza@aluno.uepb.edu.br; [4] debora.cordeiro@servidor.uepb.edu.br; [5]
silvio@servidor.uepb.edu.br.
*Corresponding Author

How to cite this paper: de Queiroz, F. A., Negreiros, I. B., de Souza, G., de Sousa, D. C., Xavier Júnior, S. F. A. (2024). Predicting sentiments in Spotify comments: A comparative analysis of Machine Learning Models. *Socioeconomic Analytics,* 2(1), 107-113. https://doi.org/10.51359/2965-4661.2024.265070

RESEARCH ARTICLE

Socioeconomic Analytics
https://periodicos.ufpe.br/revistas/SECAN/
ISSN Online: 2965-4661

Submitted on: 11.11.2024.
Accepted on: 15.12.2024.
Published on: 27.12.2024.

Abstract

Using data from user sentences on Spotify, this work explores through Natural Language Processing positive and negative sentiments in each comment. We compare different statistical modeling and Machine Learning techniques, identifying the ones with the greatest accuracy in predicting sentiments. As a result, the assessment supports most of the sentences presented with negative connotations. As for modeling, the Logistic Regression and Random Forest models resulted in better accuracy.

Keywords

Sentiment Analysis, Natural Language Processing, Machine Learning, Logistic Regression, Random Forest, Spotify

1. Introduction

In the current scenario, opinions about services or products are essential to market performance. People report their experiences with certain brands or companies on a daily basis, whether such experiences are good or unpleasant, after all, those who do not first research the comments about it before buying a product. With the high use of social media, there was also the naturalization of said feedback, however, the feedbacks portray not only people's satisfaction, but also their feelings in relation to the product purchased or the service provided. In this sense, it is necessary to understand the feeling shown by consumers when disseminating their reviews.

In view of this, Cambria & White's (2014) techniques address, in their research, several advances and challenges in the field of Natural Language Processing, emphasizing the main trends and innovations that have driven the progress of this area. They highlight the relevance of combining disparate approaches, such as rule-based and statistical approaches, to overcome past limitations and achieve significant advances in natural language understanding and processing. They will be applied in order to understand the emotion expressed by Spotify users in their reviews of the app. In addition, the Logistic Regression, Random Forest, Support Vector Machine (SVM) (Lorena & de Carvalho, 2007) and Gradient Boosting models will be applied, to identify the one that best predicts the sentiment of the users of the application.

2. Literature review

Machine learning (ML) can be described as a set of statistical techniques together with computing, in which through algorithms it can identify patterns to make decisions or make predictions. The term Machine Learning was first used in 1969 by Arthur Samuel, a pioneer in the field of artificial intelligence. He used the term when developing software that played checkers and improved his moves over time, learning from experience. Samuel defined Machine Learning as the field of study that gives computers the ability to learn without being explicitly programmed (Samuel, 1959).

2.1 Machine Learning

New technologies are typically designed to make the process more manageable, accurate, faster, or less expensive. They also allow us to carry out or develop work that was previously unattainable in certain circumstances. One of the rapidly expanding scientific procedures for practical use in recent years has been Artificial Intelligence (AI). ML is an application of AI in which algorithms trained on data produce AI. Artificial intelligence and machine learning have grown in popularity over the past decade thanks to significant advancements in computing technology. This popularity has resulted in dramatic advances in the ability to gather and analyze massive amounts of data (Alkhatib *et al.,* 2023).

In Supervised Learning, machines are taught to solve problems with the help of humans, who gather and identify data and then “feed” it to the systems. A computer is given specific data characteristics to examine to detect patterns, classify items, and assess whether its prediction is correct or incorrect. In Unsupervised Learning the methods are primarily concerned with the grouping of uncategorized data. In this category of learning, machines learn to identify patterns and trends in unlabeled data without anyone being supervised by humans. Finally, in Reinforcement Learning in a confined environment that is foreign to them, models must solve a problem through a series of trial and error. Machines are punished for mistakes and rewarded for good attempts, similar to a scenario in many games (Arulkumaran *et al.*, 2017).

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is an area of computer science and artificial intelligence that is dedicated to the development of systems capable of understanding, interpreting, and generating human language. This area combines knowledge of linguistics, computer science, statistics, and machine learning to facilitate interaction between humans and machines (Silva & Souza, 2014).

NLP is used in various activities related to language, questions, questioning, sorting in various ways, and conversation between people. Figure 1 shows practical examples of NLP applications below.

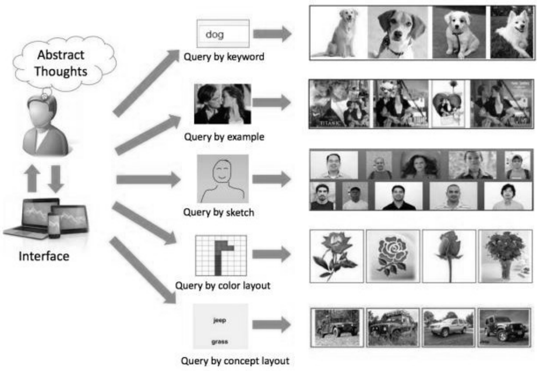


Figure 1 - NLP applications (Adapted from Piplani & Bamman, 2018).

NLP models work by finding relationships between the constituent parts of language - for example, the letters, words, and phrases found in a text dataset. NLP architectures use various methods for data preprocessing, feature extraction, and modeling.

The first step in understanding NLP is to focus on the meaning of dialogue and discourse in a specific context. The primary purpose is to facilitate meaningful conversations between a voicebot and a human. For example, when giving commands to chatbots, such as "show me the best recipes" or "play music for the party", we are within the scope of this step. This involves understanding and responding to user requests within the context of the ongoing conversation.

Before a model processes text for a particular task, it is common for text to be preprocessed to improve the performance of the model or to convert words and characters into a format that the model can understand. Several techniques can be used in this data pre-processing.

The process of dividing text into smaller units called tokens, which can be words, subwords, or characters. They are tokenization by words: It consists of dividing the text into individual words; sentence tokenization: Involves dividing the text into individual sentences;

Subword tokenization: This refers to the division of words into smaller units, such as prefixes, suffixes, or even individual characters.

Bag-of-Words is a widely used technique in natural language processing to represent texts in numerical form. In this model, a text is represented by the frequency of occurrence of its words, disregarding the order and grammatical structure. As highlighted by \cite{harris1954}, this approach is based on word distribution, where meaning is inferred from the contexts in which words appear.

The Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure that evaluates the importance of a word in a specific document, considering its frequency in the document and its occurrence in a set of documents (*corpus*). Harris (1954) explains that the TF-IDF combines two metrics: Term Frequency (TF): measures the frequency of a word in a document. The more times the word appears, the higher the TF; and Inverse Document Frequency (IDF): Evaluates the rarity of a word in the *corpus*. Common words in many documents receive a low IDF, while rare words receive a high IDF. The multiplication of TF and IDF results in the TF-IDF, which highlights frequent words in a document, but rare in the *corpus*, indicating potentially more informative terms for the characterization of the document's content.

Finally, Part of Speach (POS) tags are labels assigned to each word in a text, indicating its grammatical function, such as noun, verb, adjective, among others. The identification of these tags is essential for understanding the syntactic and semantic structure of a sentence, aiding in tasks such as syntactic analysis, machine translation, and information extraction. Salton & Buckley (1988) proposed a rule-based tagging method that demonstrated effectiveness in assigning POS tags, contributing significantly to the advancement of natural language processing techniques.

3. Methodology

The data used refers to comments from users of the Spotify application, are phrases and reports of the application's performance and customer satisfaction. In this sense, statistical methods will be used to calculate basic information that includes the number of negative and positive sentences, techniques for processing sentences, removing stop words and cleaning up information will also be applied. Subsequently, two word clouds will be created, one for each type of feeling, and finally the comparison of models will be carried out. The models used will be Logistic Regression (Dayton, 1992), Random Forest (Biau & Scornet, 2016), SVM and Gradient Boosting (Natekin & Knoll, 2013). All these methods have a high predictive capacity, and the criterion of choice used will be accuracy.

4. Analysis and discussions

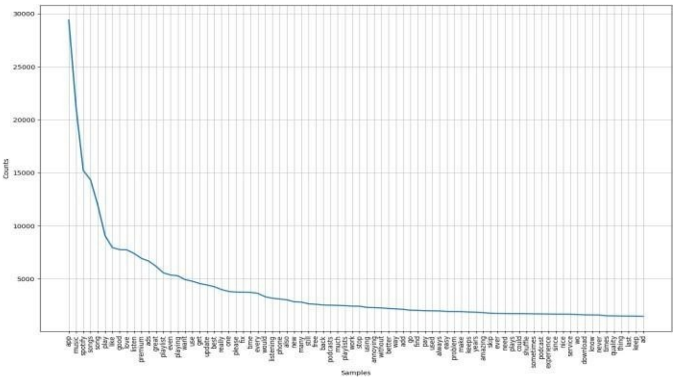
The objective was to identify the feelings transmitted through the comments of Spotify users, dividing them into positive and negative feelings. In addition, it was also of interest to check the modeling with greater ability to predict reviews. The following table presents the main results inherent to the analysis.

Table 1 - Number of sentences per feeling.

Positive	**54046**
Negative	**62469**

Through Table 1, it is possible to notice that most of the phrases about Spotify show negative sentiment, about 54%. While 46% of comments have positive sentiment.

Figure 2 - Word frequency.



Through Image 1, it is noticeable that the words most frequently used by Spotify customers are positive. However, the words alone do not correspond to the feeling that the whole sentence presents.

Figure 3 - Word cloud for positive and negative feelings.

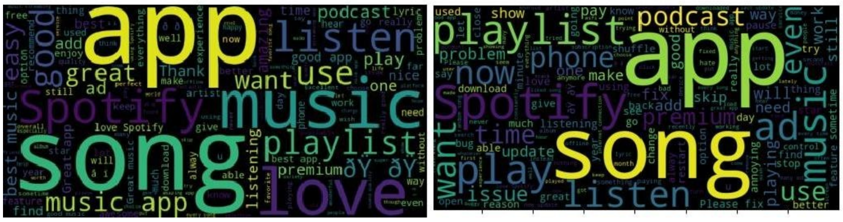


Figure 2 shows the main words contained in the positive and negative sentences. Notice that in the cloud on the left there are only optimistic words, corroborating the notion of positivity of the phrases. While in the cloud on the right are contained words related to negative sentiment.

Table 2 - Accuracy of the models.

	Accuracy
Logistic Regression	0.8826
Random Forest	0.8633
SVM	0.8798
Gradient Boosting	0.8277

Table 2 shows that all the models applied obtained good accuracy. Among them, Logistic Regression and the SCV model stand out, presenting the highest accuracy values, being the best fit models. While Gradient Boosting resulted in the lowest accuracy value.

5. Conclusions

The Natural Language Processing method showed good ability to identify the feelings of the sentences, through the word clouds it was possible to verify that the sentences identified as positive or negative were composed mostly of words that reflected such emotions.

The use of Machine Learning models to predict sentiment from Spotify reviews demonstrated great performance, along with Natural Language Processing techniques. All models showed good accuracy, with emphasis on Logistic Regression and the SVM model.

References

1. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227
2. Lorena, A. C., & De Carvalho, A. C. P. L. F. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada, 14*(2), 43–67.
3. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210
4. Alkhatib, R., Sahwan, W., Alkhatieb, A., & Schütt, B. (2023). A brief review of machine learning algorithms in forest fires science. *Applied Sciences, 13*(8275). https://doi.org/10.3390/app13148275
5. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240
6. Silva, E. M. da, & Souza, R. R. (2014). Fundamentos em processamento de linguagem natural: Uma proposta para extração de bigramas. *Encontros Bibli:Revista Eletrônica de Biblioteconomia e Ciência da Informação, 19*(40), 1–31.
7. Piplani, T., & Bamman, D. (2018). DeepSeek: Content-based image search & retrieval. *arXiv*. Disponível em https://arxiv.org/abs/1801.03406
8. Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520
9. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
10. Dayton, C. M. (1992). Logistic regression analysis. *Stat, 474*, 574. Washington, DC.
11. Biau, Gérard, & Scornet, Erwan. (2016). A random forest guided tour. Test, 25, 197–227. Springer.
12. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 21. Frontiers Media SA. https://doi.org/10.3389/fnbot.2013.00021