ISSN 1870-4069

Social Media Sentiment Analysis

Mariana Edith Antonio Aranda, Brenda Sunuami González López,
María Guadalupe Pineda Arizmendi

Tecnológico de Estudios Superiores de Tianguistenco,
Mexico

{mariana.editharanda15, brenda.sunuami.gonzalez.lop,
mariaguadalupe.pineda.arizmendi}@gmail.com

Abstract. The purpose of this paper is to analyze the feelings that people express on social networks such as X (formerly Twitter), Facebook and Instagram. To achieve this, natural language processing techniques and machine learning algorithms were applied. With these tools, the publications were classified into three types of feelings: positive, negative and neutral. Data already available in the English language were used. For the analysis we used the VADER tool (an analyzer based on dictionaries and rules) and an algorithm based on Vector Support Machines. The results show that the model works best with neutral and positive publications, but has some difficulties in identifying negative ones. This type of analysis can be used to better understand public opinion and help make decisions in areas such as communication, marketing or attention to social problems.

Keywords: Machine learning, vader, natural language processing, vector support machines.

1 Introduction

Social networks have become a widely used medium for expressing opinions, emotions and experiences. Every day, millions of people post content that reflects their mood, personal experiences or reactions to social events. This has generated a large volume of data that can be used to understand collective feeling.

Sentiment analysis is a tool that, through the use of natural language processing and machine learning models, identifies whether a publication expresses a positive, negative or neutral opinion. This classification is useful for companies, institutions or researchers who want to know people’s views on a particular subject.

On various platforms such as X, Facebook and Instagram users make a lot of posts that reflect different feelings, generating a large amount of textual data that is useful for their analysis.

The problem is that, although these publications contain valuable information, their analysis and efficient classification faces several challenges. For example, the language used in social networks is extremely varied, with informal expressions, abbreviations, emojis, etc., making it difficult to interpret feelings.

The use of natural language processing techniques and machine learning algorithms allows large volumes of data to be automatically processed and classified, identifying whether the post is expressed in categories as positive, negative and neutral.

The general objective of this work is to implement artificial intelligence techniques, such as natural language processing algorithms and machine learning, to analyze publications made on X platforms, Facebook and Instagram and classify them based on three polarities: positive, negative and neutral.

2 Theoretical Framework

Several authors have addressed the analysis of feelings using different techniques. Salgado and Trujillo (2024) used neural networks and SVM and Bayesian classifiers, achieving an accuracy of 80% on Twitter. Lazo and Rodas (2024) evaluated student comments on social networks of universities in Cuenca, finding a predominantly positive perception.

Moreno, Ávila and Ramírez applied techniques such as Python, NLTK and TextBlob to identify business opportunities through sentiment analysis on Twitter. Cardoso, Talame, Amor and Neil (2019) categorized tweets on emotions as fear, anger and happiness using NoSQL databases. Granados (2020) developed a model based on recurrent neural networks and GRU, using the TASS corpus to detect opinion trends in Spanish.

Maldonado (2022) used TextBlob and Tweepy in policy and marketing contexts, evaluating the accuracy of models. Henríquez, Pla, Hurtado and Guzmán (2017) applied SVM along with ontologies to classify opinions on products and services. Aguado et al. (2012) used language rules and tools like Freeling 3.0 to classify emotions in Spanish.

Calvo Madurga (2020) compared models of emotional classification in Spanish using techniques such as Bag of Words and Word Embeddings. Scotto (2021) developed a model based on polarity dictionaries for texts in Spanish.

Rojo, Pollo y Britos (2020) adapted a corpus to the Spanish of Rio de Janeiro in order to improve the analysis of feelings on Twitter. Fernández, Gutiérrez, Gómez and Martínez (2015) created a web application for real-time monitoring of opinions using sentiment dictionaries and machine learning.

3 Methodology

The proposed method describes a detailed process for performing sentiment analysis on social media, ranging from data collection and preparation to the interpretation of results. This process involves several interconnected stages, such as the cleaning and preparation of textual data, its processing using linguistic techniques, and the application of machine learning models to classify the content according to its polarity positive, negative or neutral. (See Fig. 1).

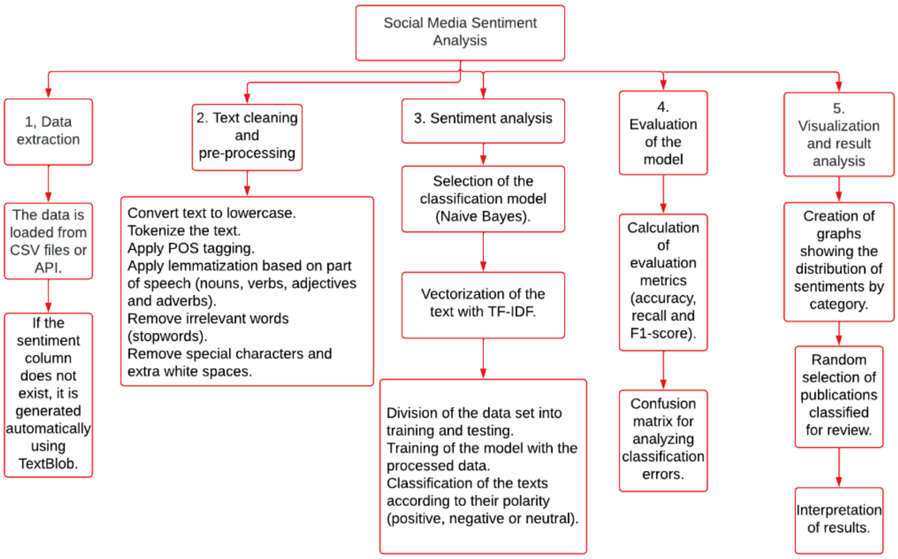
Social Media Sentiment Analysis
1, Data extraction
The data is loaded from CSV files or API.
If the sentiment column does not exist, it is generated automatically using TextBlob.
2. Text cleaning and pre-processing
Convert text to lowercase.
Tokenize the text.
Apply POS tagging.
Apply lemmatization based on part of speech (nouns, verbs, adjectives and adverbs).
Remove irrelevant words (stopwords).
Remove special characters and extra white spaces.
3. Sentiment analysis
Selection of the classification model (Naive Bayes).
Vectorization of the text with TF-IDF.
Division of the data set into training and testing.
Training of the model with the processed data.
Classification of the texts according to their polarity (positive, negative or neutral).
4. Evaluation of the model
Calculation of evaluation metrics (accuracy, recall and F1-score).
Confusion matrix for analyzing classification errors.
5. Visualization and result analysis
Creation of graphs showing the distribution of sentiments by category.
Random selection of publications classified for review.
Interpretation of results.

Fig. 1. Diagram of the proposed method.

4 Results

We work with three datasets, each platform consisting of a number of publications as shown in Table 1.

As mentioned before, posts are not tagged, so Vader is used to define the polarity of each post. A small count was made to verify how many positive, negative and neutral posts there are on each platform. Thus, giving the following data shown in Table 2.

The SVM model had an acceptable performance in general, highlighting its effectiveness in the classification of neutral publications, although with some difficulties in identifying negative texts. The evaluation metrics and confusion matrix obtained in the model test are presented below:

Accuracy: 0.60, indicating that approximately 60% of the total predictions were correct.

The model divided the data into 80% for training (2,400 data) and 20% for testing (600 data).

Evaluation metrics such as Accuracy (Precision), Recovery (Recall) and F1-Score were used as show in Table 3. The model makes predictions about the polarity of the texts, as a first result it has the evaluation of metrics show in Table 3.

Precision (Precision): Indicates the proportion of instances classified as a specific category that actually belong to that category.

Negative:0.83 this means that 83% of posts classified as negative are actually negative. Neutral: 0.60 this means that 60% of the publications which were classified as neutral are actually so. Positive: 0.59 this means that 59% of the publications that were classified as positive are actually positive.

Table 1. Number of posts from each social network.

	X	**Facebook**	**Instagram**
Number of posts	1,000	1,000	1,000

Table 2. Number of positive, negative and neutral publications on each platform.

	X	**Facebook**	**Instagram**
Positive publications	282	474	338
Negative publications	170	149	53
Neutral publications	548	377	609

Table 3. Evaluation metrics.

	precision	**recall**	**f1-score**	**support**
negative	0.83	0.12	0.21	83
neutral	0.60	0.89	0.72	311
positive	0.59	0.36	0.45	206

Recall (Recovery or Sensitivity): Measures the proportion of true instances of a class that the model was able to identify correctly Negative: 0.12 this means that 12% of all really negative posts were correctly identified, Neutral: 0.89 this means that 89% of the really neutral posts were correctly classified. Positive: 0.36 this means that 36% of the really positive posts were rated correctly.

F1-Score: It is the harmonic mean between precision and recall, used when classes are unbalanced.

Negative: 0.21 this suggests that the model does not classify negative posts very well. Neutral: 0.72 this means that there is an acceptable performance in the category. Positive: 0.45 indicating performance similar to the neutral class.

Support: Number of real instances in each category.

- Negative: 83 posts are negative.
- Neutral: 311 publications are neutral.
- Positive: 206 publications are positive.

A confusion matrix was used to analyse classification errors. (See Fig. 2).

Class "negative":

- 10 publications were correctly classified as "negative".
- 55 publications that were actually "negative" were incorrectly classified as "neutral".
- 18 publications that were "negative" were incorrectly classified as "positive".

Class "neutral":

- 278 publications were correctly classified as "neutral".
- 0 publications that were "neutral" were incorrectly classified as "negative".

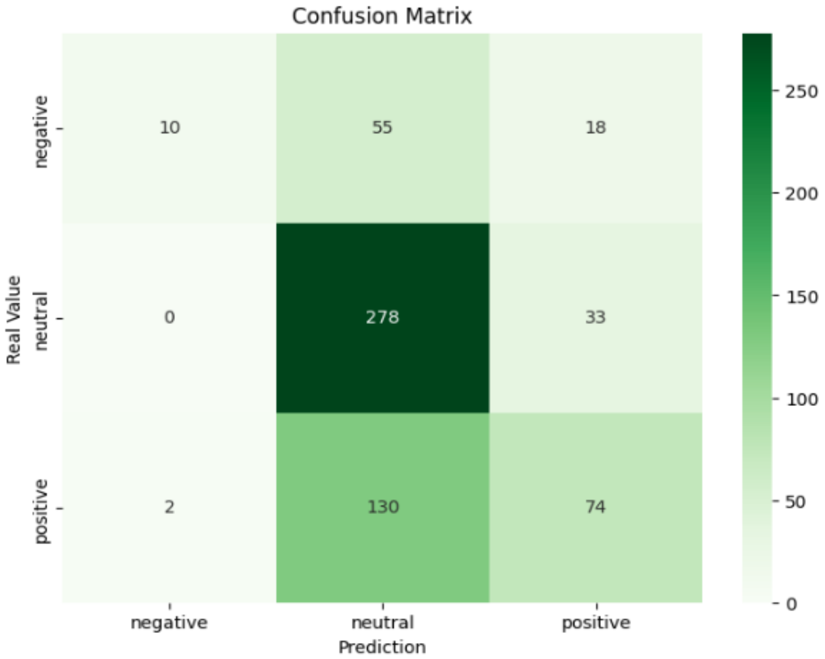

Fig. 2. Confusion Matrix.

- 33 publications that were "neutral" were incorrectly classified as "positive".

Class "positive":

- 74 publications were correctly classified as "positive".
- 130 publications that were "positive" were incorrectly classified as "neutral".
- 2 publications that were "positive" were incorrectly classified as "negative".

As extra information, several publications were shown before and after being preprocessed, in order to show more detail how the preprocessing part works.
The first example will show in detail how preprocessing works, then for the following examples only the original and preprocessed text will be shown without details.

Original text: Hello everybody! Im back on the job, back from the children camp (someone had to look after my sis pupils)... and totally exhausted.

Details of text pre-processing:

- **Lower case conversion:** The text is transformed to lowercase.

hello everybody! i`m back on the job, back from the children camp (someone had to look after my sis' pupils)... and totally exhausted"Tokenization: the text is divided into individual words.

- **Post tagging:** each word is tagged with its grammatical category.

“hello”, “everybody” → Nouns or greetings.

“im”` → Stopword (removed).
“back” → Adverb.
“on”, “the”, “from”, “to” → Stopwords (removed).
“job”, “camp”, “pupils” → Nouns.
“children” → Adjective (headword: “child”).
“someone”, “my”, “sis'” → Stopwords (removed).
“look” → Verb.
“had”, “after” → Stopwords (removed).
“totally” → Adverb.
“exhausted” → Adjective (keyword: “exhaust”).

- **Lemmatization:** words are reduced to their base form.

"children" → "child".
"pupils" → "pupil".
"exhausted" → "exhaust".

- **Words and special characters** that do not contribute meaning for analysis are eliminated.

Irrelevant words such as “im”, “on”, “the”, “from”, “to”, “my”, “and”, “had”, “after” are removed.
Characters such as “!”, “,”, “(”, “)”, “...” and other punctuation marks are removed.
Result of the preprocessed text: hello everybody back job back child camp someone look si pupil totally exhaust.

Examples of X

- Sentiment: positive

  Original text: You are welcome, xuxu

  Processed text: welcome xuxu

  Original Text: mmmmm my hair smells guuud. the wonders of 'pantien' ;D

  Processed Text: mmmmm hair smell guuud wonder

  Original Text: Pepsi throwback, you taste so good in my belly.

  Processed Text: pepsi throwback taste good belly

- Sentiment: negative

  Original Text: Irony: Inventor of Ford Mustang can`t keep his car http://tinyurl.com/lpmvtk via:

  Processed Text: irony inventor ford mustang keep car http via

  Original Text: meh! You should try the one on commercial drive with all the cats

  Processed Text: meh try one commercial drive cat

  Original Text: Consider yourself lucky. It hasn`t rained here in ages. It`s depressing.

  Processed Text: consider lucky rain age depressing

- Sentiment neutral:

Original Text: just woke up, I`m starving

Processed Text: wake starve

Original Text: at the drive ins with daa crewww

Processed Text: drive ins daa crewww

Facebook Examples

- Sentiment: positive

Original Text: Breaking: Edwin Díaz will be suspended 10 games after his ejection for having a foreign substance on his hand, per Jeff Passan. New York Mets | MLB

Processed Text: breaking edwin díaz suspend 10 game ejection foreign substance hand per jeff passan new york mets mlb.

Original Text: Welcome back to another edition of Bet.! Doug Kezirian is joined by Ohm Youngmisuk and Nick Friedell as they preview Game 3 of the NBA Finals. Will the Heat continue to thrive in their role as the underdog? Plus, Ian Parker joins the show to preview UFC 289.

Processed Text: welcome back another edition doug kezirian join ohm youngmisuk nick friedell preview game 3 nba final heat continue thrive role underdog plus ian parker join show preview ufc 289.

- Sentiment: negative

Original Text: Game 4. Denver Nuggets up 4 at the half 👀 Gordon and Jokic lead all scorers with 16 points 🤝

Processed Text: game denver nugget 4 half gordon jokic lead scorer 16 point

Original Text: Breaking: Tonight's Chicago White Sox-New York Yankees and Detroit Tigers-Philadelphia Phillies games have been postponed due to poor air quality in the NY and Philly areas. Both games have been rescheduled for Thursday.

Processed Text: breaking tonight chicago white york yankee detroit phillies game postpone due poor air quality ny philly area game reschedule thursday

- Sentiment: neutral

Original Text Philadelphia Phillies third baseman Alec Bohm will be participating in the 2024 Home Run Derby, the team announced on Friday 🤩

Processed Text:philadelphia phillies third baseman alec bohm participate 2024 home run derby team announce friday

Original Text The Emirates NBA Cup West groups are set 👏

Processed Text:emirate nba cup west group set

Instagram Examples

- Sentiment: positive

Original Text awww @shania.mooree so cute

Processed Text:awww cute

Original Text I have them & they’re so good!!! 😍😍😍

Processed Text:good

- Sentiment: negative

Original Text I have oily skin, and I have such a hard time finding one. I just wanna have a velvety matte skin!! And for not have my makeup separate

Processed Text:oily skin hard time find one wan na velvety matte skin makeup separate

Original Text Why do you half of your plans have an entryway coming into a living room or kitchen with no closet?

Processed Text:half plan entryway come living room kitchen closet

- Sentiment: neutral

Original Text:My fave is demolition

Processed Text: fave demolition

Original Text:Want it

Processed Text: want

Original Text:Do you guys do commercial roofing?

Processed Text: guy commercial roofing

5 Conclusions

The model performs well in the "neutral" category, both in precision and recall. Has difficulties in the "negative" category, where recall is very low (it identifies only 12% of real ones). In the "positive" category, performance is acceptable, but there is still room for improvement. The overall accuracy of the model was approximately 60%, indicating reasonable performance.

This work demonstrates that sentiment analysis using natural language processing (NLP) and machine learning (ML) techniques on various digital platforms is an extremely useful tool for interpreting the emotions expressed by online users. Through this approach, it is possible to gain a clearer and more accurate understanding of users' opinions and attitudes towards certain topics, products, or services. The results obtained show that social media, blogs, forums, and other online platforms constitute a rich and dynamic source of emotional data that can be classified into sentiment categories such as positive, negative, and neutral. This classification is essential for companies, institutions, and analysts looking to understand public reactions and improve their communication or marketing strategies.

The implementation of advanced text processing techniques such as tokenization, lemmatization, and stopword removal significantly improves the quality of text preprocessing, which in turn makes sentiment classification more accurate and effective:

- Tokenization allows the text to be divided into meaningful units such as words, phrases, or even characters, making it easier to identify key elements in the analysis.
- Lemmatization helps reduce words to their base or root form, preventing the distortion of meaning due to morphological variations in words, which improves data understanding.
- Stopword removal, by removing words that do not carry significant semantic value, allows the system to focus on terms that truly influence the expressed sentiment.

When applied correctly, these techniques not only optimize the accuracy of the analysis but also reduce noise in the data, making machine learning models more effective in predicting the polarity of emotions (positive, negative, or neutral) in an automated manner. This type of analysis can be used in a wide range of applications, such as improving customer experience, evaluating advertising campaigns, analyzing public opinions, or even detecting crises on social media. In summary, the use of NLP and ML in sentiment analysis represents a key tool for understanding user emotions and attitudes in the digital age.

References

1. Salgado, N., Trujillo, G.: Sentiment Analysis in Social Network Data: Application of Natural Language Processing and Machine Learning Techniques to Analyze Opinions and Sentiments in Social Network Data in the Context of Information Systems. Dominio de las Ciencias, 10(1), pp. 314–327 (2024) doi: 10.23857/dc.v10i1.3714.
2. Lazo-Calle, A.A., Rodas-Calle, E.L.: Sentiment Analysis in the Social Networks of the Universities of Cuenca. Institutional Repository of the Universidad Politécnica Salesiana (2024) http://dspace.ups.edu.ec/handle/123456789/26900
3. Moreno, L., Ávila, F., Ramírez, A.M.: Business Opportunities. Analysis-of-Feelings-on-Twitter-Social-Networks.pdf
4. Cardoso, A.C., Talame, L., Amor, M.: Opinion Mining: Sentiment Analysis in a Social Network. Institutional Repository of the UNLP. Opinion Mining: Sentiment Analysis in a Social Network (2019)
5. Granados, J.D.: Application of Machine Learning Techniques to Analyze the Polarity of Sentiments in Text to Detect Trends of Opinion on Online Platforms. [Degree project, Santo Tomás de Aquino University, Faculty of Electronic Engineering, Bogotá D.C]. Academic Repository (2020)
6. Maldonado, E.S.: Sentiment Analysis on the Twitter Social Network using Natural Language Processing. [Degree Thesis, National University of Chamborazo, Riobamba, Ecuador]. UNACH Digital Repository: Sentiment Analysis on the Twitter Social Network Using Natural Language Processing (2022)
7. Henríquez, C., Pla, F., Hurtado, L.F.: Sentiment Analysis at the Aspect Level Using Ontologies and Machine Learning. Natural Language Processing, (59), pp. 49-56 (2017)
8. Aguado, G., Barrios, M., Socorro, M.: Sentiment Analysis in a Social Network Corpus. Degree thesis. Polytechnic University of Madrid, Complutense University of Madrid (2012)
9. Calvo Madurga, A.: Analysis of Feelings and Emotions on Social Networks Using ML. [Final degree project, University of Valladolid], UVa. Analysis of Feelings and Emotions on Social Networks Using ML (2020)

10. Scotto, J.: Sentiment Analysis of Opinions on Social Networks Using Natural Language Processing Techniques. Degree thesis, University of Belgrano, Buenos Aires, Argentina, Faculty of Engineering and Computer Technology, Computer Engineering (2021)
11. Rojo, V., Pollo-Cattaneo., Ma. F., Britos, P.: Sentiment Analysis on Twitter: Development of Resources in Rioplatense Spanish from Argentina. Institutional Repository of the UNLP (2020)
12. Fernández, J., Gutiérrez, Y., Gómez, J.: Social Rankings: Visual Analysis of Sentiments in Social Networks. Spanish Society for the Processing of Natural Language, (55), pp. 199–202 (2015)