Text Classification — Sentiment Analysis (MI201 Project 3)

1st Carlos Adrian Meneses Gamboa
Ingénieur Degree Programme in STIC
ENSTA Paris
Paris, France
carlos.meneses@ensta-paris.fr

2nd Jose Daniel Chacon Gomez
Ingénieur Degree Programme in STIC
ENSTA Paris
Paris, France
jose-daniel.chacon@ensta-paris.fr

3rd Santiago Florido Gomez
Ingénieur Degree Programme in STIC
ENSTA Paris
Paris, France
santiago.florido@ensta-paris.fr

***Abstract*—This work presents a sentiment analysis system for short English texts and compares classical machine learning methods with a transformer-based approach. Using standard vector representations (bag-of-words and TF-IDF), we train and evaluate several classifiers and benchmark them against a model leveraging BERT embeddings. Results are reported with accuracy and macro-F1, highlighting differences in performance, robustness, and computational cost. The study provides practical guidance on selecting an appropriate sentiment classification pipeline under typical resource constraints.**

***Index Terms*—sentiment analysis, NLP, text classification, TF-IDF, BERT**

I. Introduction

Sentiment analysis of short texts becomes fundamentally important when perceptions are considered a critical information asset for product and service owners [1]. This is especially relevant in the development of emotion-driven systems, which can yield meaningful insights to improve the user or customer experience. For example, these insights can lead to adjustments in customer-support strategies or to more targeted marketing campaigns [2]. As a conceptual input for such improvements, search systems or sentiment-analysis approaches can be adapted to focus on the emotions expressed by the target population. In this context, social networks—and more specifically short messages such as tweets and comments on multimedia platforms—are among the most commonly used sources for conducting this type of analysis.

This project focuses on the automatic sentiment analysis of short English texts. First, an exploratory phase is conducted in which the dataset content is preprocessed, and a preliminary analysis of the information is performed using traditional machine-learning methods. Subsequently, the classification stage is carried out with standard classifiers such as Naive Bayes, Logistic Regression, and Linear SVM, using multiple text representation schemes, including bag-of-words, word-level TF–IDF, and character-level TF–IDF. Model performance is reported using accuracy, macro-F1, and complementary metrics to ensure a fair comparison.

Next, a multilayer perceptron (MLP) trained on vectorized text is evaluated, and an alternative based on BERT embeddings is studied to capture contextual semantics. To this end, the performance of MLP models built for each vectorization approach is compared across four network architectures, each adapted to the amount of information provided by the corresponding vectorizer or by BERT, and oriented toward a final three-class classification. In addition, an appropriate depth is defined according to the level of detail in the input representation in order to reduce overfitting on the training data. Dropout layers are also incorporated between hidden layers to further control overfitting and overtraining.

Finally, in order to improve message classification, strategies based on large language models (LLMs) were evaluated by using the API version of the Gemma 3-4b-it (Gemini) model to compare its performance as a short-text classifier against the previously trained models. In addition, LoRA was used to perform an efficient fine-tuning of BERT-based transformers [3].

II. Q0-Dataset analysis with different classical machine learning model

The Sentiment Data Analysis dataset to which access is available is composed of sets of short tweets in English that are classified into three categories according to their polarity as positive, neutral, or negative in the “sentiment” column. Additionally, there is information related to the tweet metadata in terms of the time of day when it was published, the user’s age, and their country of origin in the columns “Age of User” and “country”.

It is initially proposed to perform preprocessing of the text field, starting with the removal of null values and then, using NLTK, removing stopwords, which allows obtaining the processed text column without very frequent words in English that do not have a significant semantic contribution [4].

A. Exploratory data analysis (EDA)

An analysis of the class distribution in the training data is performed. As shown in Fig. 1, although the presence of neutral-polarity data exceeds the other classes, there is no substantial imbalance in the training dataset that could be associated with bias in the classifiers to be developed.

It is useful, prior to implementing some machine-learning strategies, to examine the behavior of groups of terms (bag-of-words) with respect to polarity classes, mainly from a frequency-based perspective.

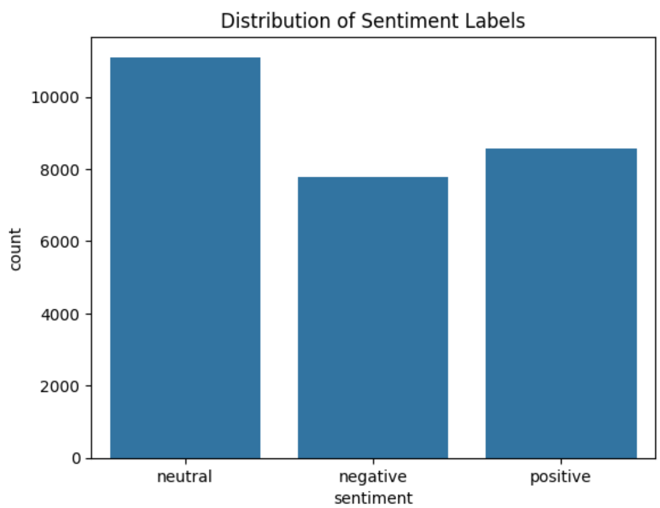


Fig. 1: Distribution of sentiments in training dataset.

This highlights the importance of the concept of a vectorizer, which is responsible for converting collections of texts—such as the content of the "processed text" column—into numerical vectors. This implies obtaining a sparse representation of each document after applying the vectorizer. This process is precisely known as vectorization, and depending on the way the numerical representation is constructed in the sparse matrix, it can be classified into several types. In this work, the following approaches are considered: BoW (Bag of Words), TF–IDF, and Char TF–IDF [5].

- **BoW.** Represents a document by the occurrence of words in n-grams, ignoring the positions they occupy within the document. It builds a dictionary/vocabulary from the words in the corpus, assigns an index to each term, and then counts how many times each word appears in the document, storing those counts in the document vector. In scikit-learn, this is implemented by `CountVectorizer`, which "converts a collection of text documents to a matrix of token counts" [5].
- **TF–IDF.** Starts from a principle similar to BoW in the sense that it also produces a sparse matrix representation; however, instead of relying only on raw frequency counts, it incorporates the inverse document frequency (IDF) term, which penalizes terms that appear in many documents [5]. Conceptually, the TF–IDF weight of a term t in a document d is computed as in Eq. (1), where $\mathrm{tf}(t, d)$ corresponds to the (raw) term frequency in d, and $\mathrm{idf}(t)$ assigns lower weights to terms that are widely distributed across the corpus.

$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \times \mathrm{idf}(t). \tag{1}$$

In practice (and as implemented in common libraries such as scikit-learn), a smoothed version of IDF is typically used, defined in Eq. (2), where n is the total number of documents in the corpus and $\mathrm{df}(t)$ is the number of documents that contain the term t.

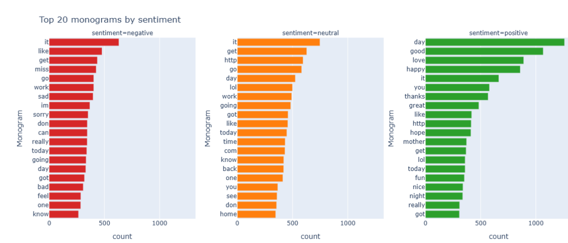
$$\mathrm{idf}(t) = \log\left(\frac{1+n}{1+\mathrm{df}(t)}\right) + 1. \tag{2}$$

Finally, after computing TF–IDF, it is common to normalize each document vector to control for differences in document length and to stabilize the scale of the features. In this work, ℓ_2 normalization is considered, as shown in Eq. (3), where $\mathbf{v}$ denotes the TF–IDF vector of a document.

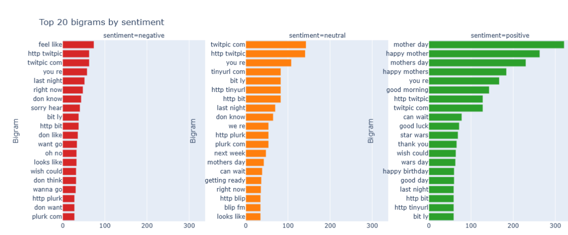
$$\mathbf{v}_{\mathrm{norm}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \frac{\mathbf{v}}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}}. \tag{3}$$

- **Char–IDF.** Consists of applying TF–IDF over character n-grams rather than word n-grams, which in scikit-learn can be controlled through the `analyzer` parameter of the vectorizer [6].

After defining the vector representation schemes for the documents, it is proposed to use BoW to extract the 20 most frequent unigrams and bigrams within the training set. In the case of TF–IDF, the objective is to display the 20 unigrams and bigrams with the highest weights in the training dataset. The use of character n-grams is not proposed for this step because, by not forming complete words, the resulting features are less intelligible for the intended analysis.



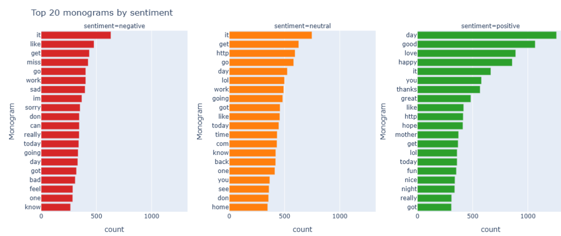
(a) Monograms



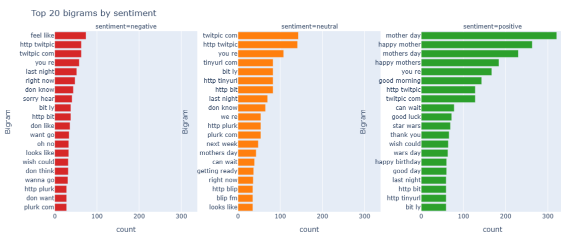
(b) Bigrams

Fig. 2: Top 20 n-gramas with BoW.

By analyzing the results shown in Figs. 2 - 3, it can be observed that some n-grams appearing under both methods correspond to words that, in everyday language use, are commonly associated with the polarity class in which they become most representative. This is the case for the unigram *bad* or *sorry* within the negative polarity class, which appears

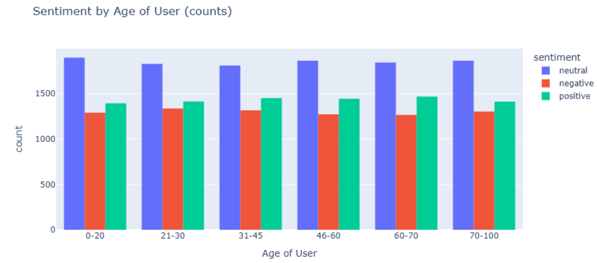


(a) Monograms

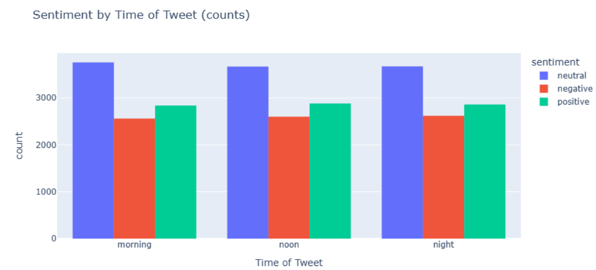


(b) Bigrams

Fig. 3: Top 20 n-gramas with TF–IDF.



(a) Sentiments by User



(b) Sentiments by type of tweet

Fig. 4: Comparative sentiment distribution by category, stratified by user and tweet type

among the top terms for both vectorization methods. Nevertheless, it is evident that, since these are purely statistical approaches, words that were not removed during preprocessing but are very frequent in English usage—such as *it*—may also appear. Similarly, certain bigrams that a speaker might classify as neutral, but that become frequent due to the data-collection context (e.g., *feels like*), appear among the most representative n-grams across multiple classes. This anticipates one of the effects addressed in later stages: training models using frequency-based vectorizations, rather than embeddings that incorporate semantic information into the sparse vector representation of the documents.

In addition to the processed text column, which is the main focus of this work, it is of interest to determine whether some of the other dataset columns exhibit any relationship with sentiment classification. To this end, the class distributions are analyzed as a function of the user's age and the time of day at which the tweet is published, as shown in Fig. 4 However, the class distribution for these categorical variables remains very similar across their respective domains in both cases; therefore, they are not considered relevant for training the sentiment classification model.

In the case of the country in which the tweet is published, a distribution of the countries with the highest number of posts is analyzed in Fig. 5. It is observed that the tweet counts do not vary in a markedly representative way among them, and although the class distributions are slightly different across countries, this feature is not included in later stages. One reason is that the model may "memorize" country-specific patterns that do not hold outside the dataset or that change over time. Additionally, given the shape of the country distribution, many countries have only a small number of examples, which generates rare features and increases the probability of overfitting. For these reasons, this column is ultimately not included in the analysis.

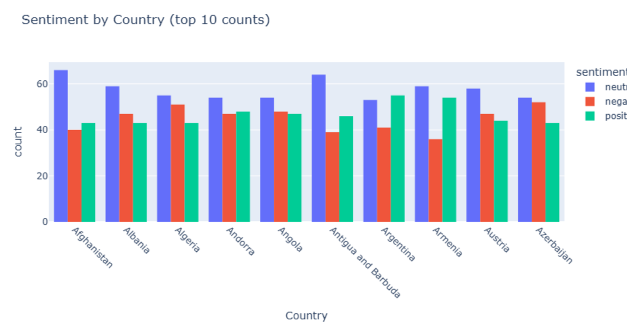


Fig. 5: Distribution of sentiments in the top 10 countries.

Finally, as a last analysis of the dataset, it is proposed to examine whether the length of the preprocessed text has any relationship with the class, and thus whether it could be worth including it as a variable during the training of the classification models. However, as shown in Fig. 6, the distribution of the number of words per message for each class is comparable (or effectively equivalent). For this reason, this variable is also not considered relevant for the objective of the models constructed in the next stage.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content

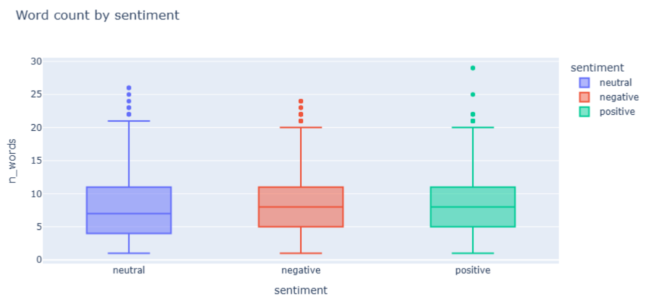


Fig. 6: Word count distribution by sentiment

and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “$\mathrm{Wb/m}^2$” or “webers per square meter”, not “$\mathrm{webers/m}^2$”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{4}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(4)”, not “Eq. (4)” or “equation (4)”, except at the beginning of a sentence: “Equation (4) is . . .”

D. LaTeX-Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

BibTeX does not work by magic. It doesn’t get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can’t read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it’s supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won’t be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word “data” is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).

- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [**?**].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. **??**", even at the beginning of a sentence.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] T. Finn and A. Downie, "How can sentiment analysis be used to improve customer experience?," *IBM Think*, accessed Jan. 18, 2026. [Online]. Available: https://www.ibm.com/think/insights/how-can-sentiment-analysis-be-used-to-improve-customer-experience

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint* arXiv:2106.09685, Jun. 2021, doi: 10.48550/arXiv.2106.09685.

[4] The NLTK Project, "Sample usage for corpus," *NLTK Documentation*, accessed Jan. 18, 2026. [Online]. Available: https://www.nltk.org/howto/corpus.html

[5] scikit-learn developers, "Feature extraction," *scikit-learn Documentation*, accessed Jan. 18, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html

[6] scikit-learn developers, “sklearn.feature_extraction.text.TfidfVectorizer,” *scikit-learn Documentation*, accessed Jan. 18, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.