IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 November 2023, accepted 13 December 2023, date of publication 25 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3346675

RESEARCH ARTICLE

Detecting Topics and Polarity From Twitter: A University Faculty Case

ALMUDENA SÁNCHEZ RUÍZ[1], DANIEL GALAN[2], ÁNGEL GARCÍA-BELTRÁN[1], AND JAVIER RODRÍGUEZ-VIDAL[1]

[1]Departamento de Automática, Ingeniería Eléctrica y Electrónica e Informática Industrial, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, 28006 Madrid, Spain
[2]Centre for Automation and Robotics (UPM-CSIC), Universidad Politécnica de Madrid, 28006 Madrid, Spain

Corresponding author: Javier Rodríguez-Vidal (javier.rodriguez.vidal@upm.es)

This work was supported in part by the Research and Development Project "Cognitive Personal Assistance for Social Environments (ACOGES)," through Ministerio de Ciencia e Innovación (MCIN)/Agencia Española de Investigación (AEI)/10.13039/501100011033 under Grant PID2020-113096RB-I00; and in part by the European Science Foundation (ESF) Investing in your Future.

ABSTRACT Social networks have become a powerful communication tool, with millions of people exchanging information, opinions, and experiences daily. Companies, organizations, and even people have turned this tool into a marketing platform to position themselves and gain popularity. However, not only do companies present products or services to society, but society also provides feedback. This feedback also has a significant impact. It is impossible to process all this vast information manually in time, but it is crucial. This information is precious even to governmental or public entities such as universities. Potential future students will use social media to learn about the general feel of the institution. Therefore, this study presents a new dataset called CEIMaT2021, which compiles all tweets in Spanish related to the Technical School of Industrial Engineering of the Universidad Politécnica de Madrid (ETSII-UPM). This dataset is designed for two main tasks of Online Reputation Management: 1) automatic detection of topics and 2) polarity. Furthermore, this study shows that the BETO model obtains better performance for topic detection for these tasks. Meanwhile, the MarIA model obtains better results for polarity detection.

INDEX TERMS Dataset, information retrieval, polarity, social network analysis, topic, Twitter, web and social media search.

I. INTRODUCTION

The emergence and development of the Internet have been one of the greatest revolutions in recent history, directly impacting every aspect of our society. The one that has experienced the most changes and has benefitted from these communication advances between people. This evolution is due to the development of social networks in 1997. These networks arose from the need to connect people anywhere in the world and to share information. According to the Cambridge dictionary: "*a social network is a website or computer program that allows people to communicate and share information on the Internet using a computer or mobile phone*". Millions of people exchange comments, opinions, personal experiences, and audiovisual content daily. Information is generated, multiplied, and transmitted quickly from public profiles anywhere. The development of social networks has been such that their use today has many applications, from leisure and culture to work. For example, there are platforms to share photos and videos, such as Instagram (https://www.instagram.com) or TikTok (https://www.tiktok.com), other for job searches, such as LinkedIn (https://www.linkedin.com), others for communicating with friends and family, such as Whatsapp (https://www.whatsapp.com) and others for watching and commenting movies or videos, such as YouTube (https://www.youtube.com).

The flow of information has no limits. Users of social networks express their opinions on any topic: new products, restaurants, hotels, politics, or the economy. Even the concept of marketing has changed. Social networks are essential for brands to know how the general public reacts to their products. However, it is not only the industry that suffers from the impact of the opinions expressed by its customers.

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

The academic world does not escape this impact either. Students at different universities post their thoughts on their social networks: faculty, facilities, and food served in their cafeterias, among others. Ignoring these opinions could lead to reputational problems resulting, for example, in a reduction in the number of students enrolled in universities.

This paper focuses on i) detecting the topics that affect the Technical School of Industrial Engineering (henceforth, ETSII) of the Universidad Politécnica de Madrid (from now on, UPM) and ii) establishing the polarity that these topics have on the ETSII in social networks, in particular, Twitter. Information appears on Twitter before in other networks; therefore, the problems that may affect ETSII will appear on Twitter before in other networks [1]. In addition, another feature that makes Twitter more attractive than other networks, such as Facebook, is that it has a public nature [1]. Some of the objectives that will be addressed in this paper are the following:

a) To generate an original dataset to study both i) the automatic detection of Twitter topics affecting ETSII and ii) the establishment of the polarity of the previously detected topics.
b) To generate baselines using state-of-the-art algorithms, which will allow establishing a starting point for future studies.
c) To analyze and compare the results obtained from the baselines.

This study aims to lay the foundation for a framework applicable to different universities or schools that allows detection and decision-making in the face of reputational threats.

The rest of the paper is organized as follows. First, some related work in the area is presented. Second, the CEIMaT2021 dataset is introduced: i) retrieved data, ii) data preprocessing, iii) the way the dataset was annotated, and iv) the preliminary analysis of the dataset. Third, the experimental framework is presented: i) a description of the baselines, ii) a description of the metrics employed, iii) a description of the balanced data methods used, and iv) a description of the features used. Fourth, the results achieved are presented and discussed. Finally, the main conclusions of the study and the outline of future work are drawn.

II. RELATED WORK

Entities (e.g., companies, products, people) need to know their positioning or reputation in social networks; this is important because social network users are capable, with their opinions, of making or losing money. Therefore, experts must be able to locate and neutralize reputational threats as quickly as possible. One of the main tasks in Online Reputation Management (ORM) is to locate the different topics of conversation concerning the entity about which users are talking. These topics can be automatically extracted by processing the information collected through social networks. The use of embeddings such as the Cbow Topic Model [2] or embeddings obtained directly through a bag of words or "tf-idf" [3], [4] together with machine learning techniques, such as Support Vector Machine [5], makes it possible to create clusters from which to obtain the topic common to these texts: war, economy, politics, etc. One of the main challenges in the cluster task is using short texts in small collections due to of the vocabulary mismatch between these types of texts and the insufficient dataset-based statistics [6]. Among the multiple applications of topic detection, one has appeared in recent years: the damage of to the reputations of entities by spreading hoaxes and fake news. Different models, such as diffusion models, basic epidemic models, or independent cascade models, can be used to detect such sources of disinformation. Other techniques, such as neural networks: Convolutional Neural Networks [7], Long Short-Term Memories [8], ensemble methods or attention mechanisms can be used to detect false information disseminated through different media [9]. Twitter allows real-time knowledge of global or related country trending topics, unlike other social networks. These characteristics are exploited along with different methods, such as Naïve Bayes [10], to know the different topics of conversation.

It is equally crucial for a company's reputation to know the topics of conversation of users as it is to know the intentionality (polarity) with which messages were written. If the intention is primarily negative, it can mean a significant economic loss for the entities. Ideally, the aim is to maximize the positive influence of the entities in the networks while minimizing negative comments. To this end, the Polarity-related Independent Cascade (IC-P) diffusion model [11], techniques based on emotional calm states in which disruptive users are added as potential customers for the entities [12], are proposed. The characteristics found in the texts and the features of the social networks: URLs, hashtags, and emoticons, among others, deserve a detailed analysis to discover the polarity of texts [13]. Lexicons created by experts are used to construct dictionaries of newly coined words and emoticons to classify tweets emotionally [14]. Lastly, employing aggregation methods to develop topic models over time can contribute to creating higher-quality topics and understanding user preferences and intentions [15], [16], [17].

As far as we know, the study of conversation topics and polarity detection in Spanish academia is limited. Only one dataset has this information in English and Spanish. The RepLab dataset in its 2013 and 2014 versions [18], [19] contains a domain-specific to universities. However, this domain is only exploratory and has yet to be used in the tasks proposed for this initiative [1].

III. THE CEIMAT2021 DATASET

In this section, the CEIMaT2021 (*Corpus de la Escuela de Industriales de la Universidad Politécnica de Madrid en Twitter versión 2021*) dataset (https://zenodo.org/record/714918

TABLE 1. Tweet download results (September 2010-July 2021).

Category	#Tweets
Published by @industrialesupm	3147
Mentioning @industrialesupm	10987
With #industrialesupm	39
With #etsii	4798
Total	**18971**

TABLE 2. Tweets after the data preprocessing step.

Category	#Tweets
Published by @industrialesupm	3016
Mentioning @industrialesupm	7162
With #industrialesupm	39
With #etsii	1200
Total	**11417**

The reason for the number of tweets exceeding 11014 is that certain tweets are included in two different categories. For instance, this occurs when tweets are both published by @industrialesupm and mention the same account.

8#.Yz51DnZByUk) is introduced. This dataset contains all the texts written in Spanish related to the ETSII Twitter account:

1) Tweets published by the official account of ETSII (@industrialesupm).
2) Tweets in which @industrialesupm is mentioned, regardless of the user who posted them.
3) Tweets that include the hashtag #industrialesupm.
4) Tweets that include the hashtag #etsii because it is the acronym for Technical School of Industrial Engineering in Spanish.

The Python library called "snscrape" [20] was used to extract data from Twitter. The number of tweets downloaded (between September 2010 and July 2021) was 18971, used as a starting point for the dataset generation. The number of tweets obtained in each category is shown in Table 1.

A. DATA PREPROCESSING

One of the main drawbacks observed was that some texts were not directly related to the ETSII-UPM since, for example, the hashtag #etsii may refer to other universities (e.g., the School of Computer Engineering of the University of Seville). Furthermore, the language used to write the tweets differed from Spanish. Therefore, to obtain the final dataset, a filtering stage was necessary for those tweets: i) written in other languages, ii) duplicated, and iii) unrelated to ETSII. After this process, the number of unique tweets recovered was 11014. The number of tweets after the data preprocessing step is shown in Table 2.

B. ANNOTATIONS

Once the tweets have been collected, they need to be labeled. This task was carried out by three different experts who manually annotated each tweet. These annotators were selected because of their background in social networks, specifically in retrieving and annotating tweets. The odd number of annotators used fulfills the objective of maintaining the fairness of the data annotations [21]. To ensure that there is no bias in the annotations, clear instructions were dictated to the three annotators, who were guided through a training process. Doubts were consulted and discussed throughout the annotation process. For this case study, two tags per tweet have been added:

- **Polarity:** determines the intention or attitude of the tweet author or the character the comment acquires. It can take the following values:
  1) *Positive:* whether or not any references suggest a favorable judgment towards ETSII.
  2) *Negative:* whether or not any references suggest a judgment against ETSII.
  3) *Neutral:* whether or not the texts do not contain references that suggest any sentiment towards ETSII.
- **Topic**: refers to the subject matter of the text. After studying the available data and based on the context of the study, the eight different topics are described:
  1) *Events:* events of some relevance, e.g., conferences or awards.
  2) *Exams:* anything related to written or oral tests, e.g., grades or reviews.
  3) *Computing:* everything related to computer equipment, Internet or ETSII computer applications.
  4) *Teaching and research:* duties related to teaching and research, e.g., lectures, professors' work, or theses.
  5) *Institution:* everything related to the functioning of the ETSII, e.g., information on bachelor's and master's degrees, academic calendar, or scholarships.
  6) *Services:* set of activities to satisfy the needs of a customer, e.g., secretary's office or cafeteria.
  7) *Infrastructure:* technical means necessary for developing an activity, e.g., indoor and outdoor spaces, facilities, or cleanliness.
  8) *Other:* tweets that cannot be included in the previous categories.

After completion of the labeling task, the quality of the annotations of the three experts was evaluated using Cohen's kappa coefficient [22]. The formula is described in Equation 1:

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{1}$$

where, P_0 is the proportion of observed agreements among annotators and P_e probability of agreement by chance. Equations 2 and 3 describe how to calculate P_0 P_e respectively:

$$P_0 = \frac{predicted\_yes + predicted\_no}{N} \tag{2}$$

$$P_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \tag{3}$$

where N is the number of observations to categorize, $predicted\_yes$ is the number of inter-annotator matches in

TABLE 3. Results obtained for the Cohen's kappa coefficient.

Task	Agreement
Polarity	0.62
Topic	0.64

TABLE 4. Tweets obtained for each topic category.

Topic	#Tweets
Events	5595
Exams	138
Computing	140
Teaching and Research	734
Institution	879
Services	1202
Infrastructure	141
Other	2185
Total	**11014**

TABLE 5. Tweets obtained for each polarity category.

Category	#Tweets
Positive	695
Negative	262
Neutral	10057
Total	**11014**

which an observation belongs to a category, *predicted_no* is the number of inter-annotator matches in which an observation does not belong to a category, k is the number of categories and n_{ki} is the number of times that annotator i predicted category k. If the annotators are in complete agreement, $\kappa = 1$, while $\kappa = 0$ means that the annotators are in complete disagreement. Since equation 3 is originally defined for two annotators while there are three in this context, the calculation of inter-annotation agreement is adjusted by first computing the pairwise agreement and then determining the average agreement across all pairs. The results obtained are shown in Table 3:

According to [23], the agreement between the annotators is substantial and therefore has been considered sufficient.

C. PRELIMINARY ANALYSIS

In this section, a preliminary analysis of the CEIMaT2021 dataset is presented. Table 4 summarizes the tweets belonging to each topic category once the dataset is tagged:

During the annotation step, it has been detected that tweets published by the ETSII inform about talks, conferences, or fairs that will take place on it, to which many users also react by mentioning @industrialesupm. This situation makes the category ''Events'' group most of the tweets. Therefore, table 5 summarizes the tweets belonging to each polarity category:

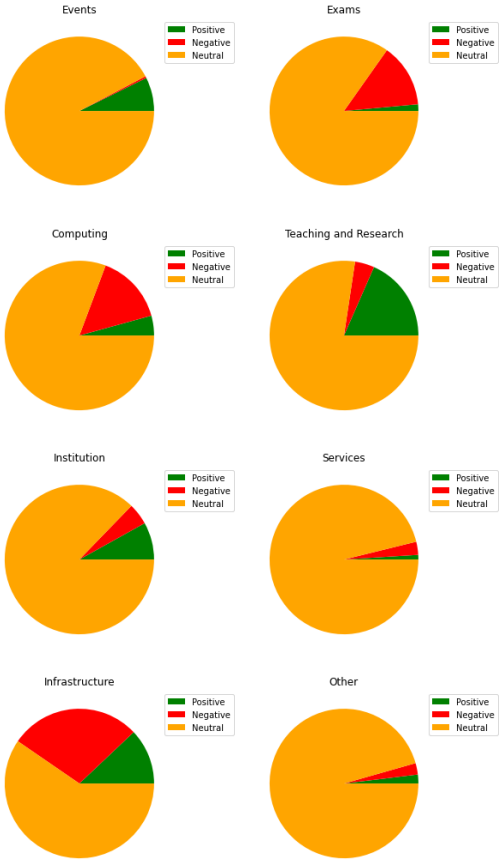

FIGURE 1. Tweet distribution in topics ("Events", "Exams", "Computing", "Teaching and Research", "Institution", "Services", "Infrastructure", "Other") and polarity ("Positive", "Negative", "Neutral").

As can be seen, the ''neutral'' category groups the vast majority of elements because most of the tweets published about the ETSII are informative, where neither positive nor negative sentiments are expressed. The graphs in Figure 1 show the distribution of tweets into topic and polarity classes. It may be observed that most tweets belonging to a topic are neutral, but, in most cases, it is followed by negative and, finally, positive. This was expected since most users of ETSII are students, and they want better conditions in their daily life on campus: better infrastructures (crowded classes) or services (cafeteria). These issues can lead new students to think about other universities in which to enroll.

IV. EXPERIMENTAL FRAMEWORK

This section introduces the algorithms for automatically detecting topics and polarity, the metrics, the methods used to balance the data, and the selected features for both tasks.

A. BASELINES

This study uses different learning algorithms as a starting point for future work. These methods were employed in work related to topic classification tasks [24]. In particular, the following algorithms have been tested:

- **Naïve Bayes (NB):** is a simple and fast [25] probabilistic classifier based on Bayes' theorem (Equation 4) that assumes that the value of a particular characteristic is independent of the value of any other feature, given the class variable. It is suitable for binary and multiclass classification tasks [26], [27].

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{4}$$

  where $P(A \mid B)$ is the probability that the hypothesis A given B; $P(B \mid A)$ is the probability of B where A is true; $P(A)$ and $P(B)$ are the probabilities that A and B are true respectively.
- **Support Vector Machine (SVM):** are supervised learning algorithms that use geometric characteristics of the input data (vectors in an n-dimensional space) to separate them into different classes using hyperplanes. This classifier aims to find the best hyperplane that classifies the data, separating the groups from each other. Thus, when new unannotated data from the test set are introduced, the algorithm studies its position and decides which side of the hyperplane it is located on. This hyperplane maximizes the distance or margin to any point, known as *Maximal Margin Hyperplane*.
- **Extreme Gradient Boosting (XGB) [28]:** based on *Decision Trees* [29] which improves the performance of most algorithms, is designed to reduce both the bias and variance of supervised learning. Traditional decision trees are based on giving a binary answer (yes or no) to an established question. In XGBoost, on the contrary, the decision nodes contain real values that determine the category to which each object belongs. This is because CART trees (Classification and Regression Trees) [30] are used.

This experiment was carried out using the Sklearn library (https://scikit-learn.org/stable/) and the default parameters for each classifier. Furthermore, to train them, the *K-Fold Cross Validation* [31] technique has been used, which is an iterative process that divides the total available data set into k groups of equal size (in this study, $k = 5$). $k - 1$ groups are used to train the algorithm, while the remaining group is used for validation. The process is repeated using a different test group until all k groups have participated in the validation of the model. This process is illustrated in Figure 2.

Additionally, other baselines related to generative AI models and pre-trained models in Spanish were included:

- **GPT-2 Small Spanish (SGPT-2) [33]:** is a Spanish-language model derived from the GPT-2 Small model. It underwent training on Spanish Wikipedia using Transfer Learning and Fine-tuning techniques, starting from the English pre-trained GPT-2 Small model. This versatile model exhibits proficiency in both text generation and classification tasks.
- **BETO [34]:** is a pre-trained transformer-based language model for the Spanish language. It was pretrained using different Spanish texts extracted from Wikipedia and the OPUS[1] project. The model used was uploaded to HuggingFace and has a similar size of BERT-base: 12 self-attention layers, 12 attention heads each, a hidden size of 768, and a total of 110M parameters.
- **MarIA/RoBERTa-base (MarIA) [35]:** is a pre-trained transformer-based language model for the Spanish language. It was pretrained using a Spanish text web crawled from the National Library of Spain (Biblioteca Nacional de España). The model used was uploaded to HuggingFace, and it is based on the RoBERTa-base model: 12 self-attention layers, 12 attention heads each, a hidden size of 768, and a total of 125M parameters.

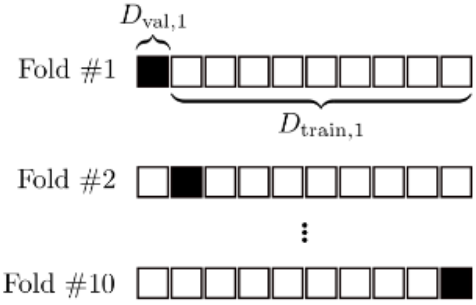

FIGURE 2. 10-Fold Cross-Validation. (Source: [32]).

All models were trained using Google Colab[2] free GPU. The same settings were applied to all previous systems: a) epochs: 2,3,5,6 and 10; b) batch size: 16 and 32; c) learning rate: 1e-5, 3e-5, 5e-5 and 1e-6 and d) weight decay: 0.01.

B. METRICS

To evaluate the performance of the different machine learning algorithms, we use accuracy, precision, recall, and F-measure, as traditionally done in supervised classification [36]. All these metrics are defined as follows:

- **Accuracy:** is the proportion of true results (both true positives and true negatives) among the total number of cases:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

- **Precision:** is the proportion between correctly labeled data in a category and the total number of data assigned (correctly or not) to that category:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

- **Recall:** is the proportion between correctly labeled data in a category and the total number of elements that should be in that category:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

[1] https://opusproject.eu/
[2] https://colab.research.google.com/?hl=es

- **F-Measure:** is the harmonic mean of precision and recall:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

where *TP* means *True Positive*, *FP* means *False Positive*, *TN* means *True Negative* and *FN* means *False Negative*.

C. BALANCING DATA

As seen in section III-C, the number of tweets corresponding to each of the three polarity groups (positive, negative, and neutral) is highly imbalanced. Although, as can be seen, the neutral class is the majority, the number of positive and negative observations corresponds to 8.6% of the total. This situation can be a problem as machine learning algorithms are designed under the hypothesis of having an equal number of observations per class [37], worsening their predictive capabilities since the tendency will be to assume that all data belong to the majority class. Over-sampling techniques have been used to solve this issue. These techniques artificially increase the number of elements in minority classes [38] until they are balanced. Two of these methods have been applied in this study:

- **Random Over-Sampling [39]:** where data are randomly duplicated. In this method, there is no loss of information, but the dataset is more prone to suffer *overfitting* [40]
- **Synthetic Minority Over-sampling Technique (SMOTE) [41]:** this method uses the *k-nearest neighbor* [42] to create synthetic data from real data. In this way, artificial observations have been generated that increase the minority classes until the number of data in all of them is equal but avoiding the *k-nearest neighbor overfitting* problem.

D. FEATURES

One of our main objectives is to establish baselines for topic and polarity detection tasks. For this purpose, word embeddings extracted directly from the texts of the dataset have been generated using TF-IDF (Term Frequency - Inverse Document Frequency) [43]. TF-IDF is a numerical measure that expresses the relevance of a word in a document in a collection. It increases proportionally to the number of times a word appears in the document. However, it is offset by the word frequency in the document collection, allowing the system to handle that some words are generally more common than others. Its mathematical definition is defined in the following equations:

$$\mathrm{tf}(t, d) = \frac{\mathrm{f}(t, d)}{\max\{\mathrm{f}(t, d) : t \in d\}} \tag{9}$$

$$\mathrm{idf}(t, D) = \log \frac{\mid D \mid}{\mid \{d \in D : t \in d\} \mid} \tag{10}$$

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D) \tag{11}$$

where $f(t, d)$ is the number of occurrences (frequency) of the term t in the document d. $\mid D \mid$ is the number of documents in the collection D and $\{d \in D : t \in d\}$ is the number of documents with the term t appearing.

Before creating the embeddings, a pre-processing step was performed: punctuation marks, capital letters, emoticons, and other non-alphanumeric symbols were eliminated. Furthermore, words that appeared in the documents less than 10 were not considered. As a result, the vocabulary obtained contains 1727 independent terms, a fragment of which is shown in Figure 3 together with their IDF values:

```
Number of documents = 11014, Number of TF-IDF features = 1727

TF-IDF features, with scores:
  0. abierta:          5.36 |   1. abierto:          6.44 |
  4. abrirá:           7.91 |   5. absoluto:         7.67 |
  8. acabo:            7.74 |   9. academia:         7.42 |
 12. acced:            7.42 |  13. acceso:           6.59 |
 16. acerca:           7.47 |  17. acog:             7.53 |
 20. acreditacion:     7.91 |  21. acreditación:     7.47 |
 24. activo:           7.91 |  25. acto:             4.77 |
 28. actuaupm:         7.13 |  29. actúaupm:         7.09 |
 32. acústica:         7.53 |  33. adam:             7.67 |
 36. adrián:           7.67 |  37. aforo:            7.74 |
 40. agradecimiento:   7.91 |  41. agrónomo:         7.74 |
 44. ahorro:           6.97 |  45. ahí:              7.60 |
 48. alberto:          6.48 |  49. alejandro:        6.67 |
 52. algoritmo:        7.42 |  53. alguien:          6.78 |
 56. alimento:         7.91 |  57. allá:             7.60 |
 60. almudena:         7.91 |  61. alonso:           7.42 |
 64. alto:             7.74 |  65. alumna:           7.26 |
 68. ambient:          6.12 |  69. amigo:            6.59 |
 72. amp:              7.82 |  73. ampliación:       7.82 |
 76. analizar:         7.91 |  77. and:              7.26 |
 80. anterior:         7.67 |  81. antiguo:          5.99 |
 84. anual:            6.94 |  85. análisi:          5.10 |
```

FIGURE 3. Fragment of the vocabulary obtained.

TABLE 6. Overall results for the topic detection task of the CEIMaT2021 dataset.

Method	Accuracy	Precision	Recall	F-Measure
NB	0.73	0.70	0.73	0.69
SVM	0.77	0.76	0.77	0.75
XGB	0.74	0.73	0.74	0.72
SGPT2-M	0.82	0.72	0.66	0.68
SGPT2-W	0.82	0.82	0.82	0.82
BETO-M	0.84	0.75	0.68	0.71
BETO-W	**0.84**	**0.83**	**0.84**	**0.84**
MarIA-M	0.83	0.72	0.70	0.70
MarIA-W	0.83	0.85	0.83	0.83

V. RESULTS AND DISCUSSION

This section presents the results obtained for topic and polarity detection tasks.

A. TOPIC DETECTION TASK

Table 6 summarizes the results (in terms of Accuracy, Precision, Recall, and F-Measure) of all algorithms used to perform the experiments over CEIMaT2021 for topic detection tasks using TF-IDF embeddings to handle textual content.

Only the best results achieved for each pre-trained model are shown in Table 6, at both macro average (suffix M) and weighted average (suffix W) levels. The macro average shows

a fair view of the performance of each class independently of its size, while the weighted average gives weight to the classes based on the proportion of the number of its elements. The following lines explain the configurations associated with these results:

- SGPT-2: 3 epochs, batch size 16, and a learning rate 3e-5.
- BETO: 10 epochs, batch size 16, and a learning rate 3e-5.
- MarIA/RoBERTa-base: 5 epochs, batch size 32, and a learning rate 5e-5.

According to the set of results obtained in Table 6, the following can be highlighted:

- In a direct comparison of the different classic algorithms used in the experimentation, the authors observed that SVM is the best performer for all evaluation metrics. This improvement is between 8% and 5% for NB and around 4% for XGB.
- In a direct comparison of the different AI models, it can be seen that BETO is the best performer for both macro and weighted average, in terms of f-measure, with a difference between 4.41% for macro average to 2.44% for weighted average with SGPT2 and a difference between 1.43% and 1.20% with MarIA.
- The discrepancy between the macro and the weighted average suggests that the class imbalance may influence the performance of the models.
- For weighted average, AI models outperform traditional models in this task. However, when considering the macro average, traditional models demonstrate superior performance.
- The deployment of a feature extracted directly from the texts, such as the embeddings generated from the TF-IDF value, provides competent baseline results, which are an encouraging starting point for further research on these data.

B. POLARITY DETECTION TASK

Table 7 summarizes the results (in terms of Accuracy, Precision, Recall, and F-Measure) of all algorithms used to perform the experiments over CEIMaT2021 for polarity detection task using TF-IDF embeddings to handle textual content.

The configurations selected for this task were:

- SGPT-2: 2 epochs, batch size 16, and a learning rate 3e-5.
- BETO: 3 epochs, batch size 16, and a learning rate 1e-5.
- MarIA/RoBERTa-base: 5 epochs, batch size 32, and a learning rate of 1e-5.

According to the results in Table 7, the following ones can stand out:

- In a direct comparison of the different algorithms used in the experimentation, it can be seen that XGB is the best performer for all the evaluation metrics but very close to the other methods, 1% to SVM, $2\% - 1\%$ to NB.

TABLE 7. Overall results for the polarity detection task of the CEIMaT2021 dataset.

Method	Accuracy	Precision	Recall	F-Measure
NB	0.89	**0.92**	0.88	**0.92**
SVM	0.89	**0.92**	0.89	**0.92**
XGB	0.90	0.92	0.90	0.92
SGPT2-M	0.94	0.67	0.48	0.51
SGPT2-W	0.94	0.92	0.94	0.92
BETO-M	0.94	0.79	0.51	0.57
BETO-W	0.94	0.93	0.94	0.93
MarIA-M	0.94	0.68	0.68	0.68
MarIA-W	**0.94**	**0.94**	**0.94**	**0.94**

TABLE 8. F-Score for polarity detection task applying over-sampling techniques.

Method	F-Score		
	Without Over-Sampling	Random Over-Sampling	SMOTE
NB	0.92	0.81	0.82
SVM	0.92	0.85	0.83
XGB	0.92	0.86	0.89
SGPT2-M	0.51	0.54	0.52
SGPT2-W	0.92	0.92	0.93
BETO-M	0.57	0.61	0.68
BETO-W	0.93	**0.93**	**0.94**
MarIA-M	0.68	0.59	0.68
MarIA-W	**0.94**	**0.93**	**0.94**

- Comparing the different AI models, the authors observed that MarIA achieves the best macro and weighted average performance in f-measure. The difference goes from 19.30% for the macro average to 1.07% for the weighted average with BETO and from 33.33% for the macro average to 2.17% for the weighted average with SGPT2.
- As observed, the disparity between the macro and weighted averages is more significant in this particular task compared to the topic detection task. This suggests that class imbalance directly impacts the performance of the AI models.
- The AI models outperform the classic models for a weighted average, but the latter achieved very close results.
- The same performance in the three methods may be due to the existing decompensation between classes. The models are limited to assuming that the data belong to the majority class, neutral polarity, so these values would not reflect the predictive capacity of the methods used.

Table 8 shows how the presence of balanced classes influences when working on a classification task. According to the results shown, the following can be outlined:

- In a direct comparison of the different algorithms used in the experimentation, the authors observed that MarIA

and BETO are the best performers for all the evaluation metrics.
- The loss of efficiency, in quantitative terms (between 6% and 1%), when using over-sampling methods only confirms that the imbalance between classes is a problem that must be addressed for this type of task. On the other hand, there are two AI models, BETO and DGPT2, whose macro average results are increased using over-sampling methods, which confirms the imbalance between classes.
- The deployment of a feature extracted directly from the texts provides strong baseline results.

VI. THREATS TO VALIDITY

Despite our thorough investigation and extensive experimentation with the data presented, we acknowledge certain limitations. These limitations could be classified into internal and external threats:

A. THREATS TO INTERNAL VALIDITY:

- *Subjective interpretations by annotators:* despite the iterative refinement of the annotation guidelines, the potential for subjective interpretations among annotators may introduce variability in the annotations.
- *Resource constraints and model limitations:* constraints related to resource availability, such as relying on the free GPU from Google Colab, limited the utilization of more robust baseline models, affecting the scalability and generalizability of the proposed approach.

B. THREATS TO EXTERNAL VALIDITY

- *Contextual specificity of the findings:* the study's findings are specific to the context of ETSII-UPM on Twitter and may not be directly extrapolated to other entities or social networks.
- *Limited generalizability:* resource constraints prevented the utilization of a larger Spanish GPT-2 model, potentially limiting the generalizability of the proposed model beyond the community setting studied.
- *Scope of the community setting:* The focus of the study on a specific community setting may restrict the external validity of the findings to similar contexts.

Despite these challenges, it is essential to note that the study serves as a foundational step in understanding the classification of the text within a particular community setting, acknowledging its specific limitations and contextual boundaries.

VII. CONCLUSION

Our main goal in this study was to investigate how Social Networks influence the university environment, and we highlight the following conclusions:

a) An original dataset (CEIMaT2021) has been built based on tweets mentioning the ETSII, which compresses two of the main tasks in online reputation: i) the detection of topics about the entity; ii) the polarity that these topics have on the entity.
b) In addition, this dataset has been tested with different state-of-the-art machine learning algorithms and AI models to serve as a baseline for future research.
c) According to the results shown in section V, features extracted from the texts provide strong baseline results and are an encouraging starting point for future projects.

In future work, we would i) add multi-label classification task for topics; ii) maintain and expand this dataset by adding information from other schools belonging to UPM University, such as the Technical School of Mining and Energy Engineering, Technical School of Architecture, and Technical School of Computer Engineering, among others; iii) expand the dataset by adding new information from other social networks, for example, Facebook, where there are no restrictions on the length of the text, to perform a comparative study between social networks and iv) continue with other main tasks in the field of online reputation monitoring which, when combined with the tasks outlined in this study, assist entities in making crucial decisions based on the people's opinions: a) the generation of automatic summaries and b) the generation of reputation reports.

REFERENCES

[1] J. Rodríguez-Vidal, J. Gonzalo, L. Plaza, and H. A. Sánchez, "Automatic detection of influencers in social networks: Authority versus domain signals," *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 7, pp. 675–684, Jul. 2019.
[2] L. Shi, G. Cheng, S.-R. Xie, and G. Xie, "A word embedding topic model for topic detection and summary in social networks," *Meas. Control*, vol. 52, nos. 9–10, pp. 1289–1298, Nov. 2019.
[3] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM," in *Proc. 8th Int. Conf. Comput. Sci. Educ.*, Apr. 2013, pp. 287–291.
[4] M. Yuan, J. Zobel, and P. Lin, "Measurement of clustering effectiveness for document collections," *Inf. Retr. J.*, vol. 25, no. 3, pp. 239–268, Sep. 2022.
[5] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*, A. Mechelli and S. Vieira, Eds. Cambridge MA, USA: Academic Press, 2020, ch. 6, pp. 101–121.
[6] L. Kotlerman, I. Dagan, and O. Kurland, "Clustering small-sized collections of short texts," *Inf. Retr. J.*, vol. 21, no. 4, pp. 273–306, Aug. 2018.
[7] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.
[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[9] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 2, p. e3767, Feb. 2020.
[10] K. M. Leung, "Naive Bayesian classifier," *Polytech. Univ. Dept. Comput. Sci./Finance Risk Eng.*, vol. 2007, pp. 123–156, Nov. 2007.
[11] D. Li, Z.-M. Xu, N. Chakraborty, A. Gupta, K. Sycara, and S. Li, "Polarity related influence maximization in signed social networks," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102199.
[12] S. Abas, M. Addou, and Z. Rachik, "Polarity switch within social networks," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 5, no. 6, pp. 817–820, Nov. 2020.
[13] J. Varga, O. B. P. Lezama, and K. Payares, "Machine learning techniques to determine the polarity of messages on social networks," in *Proc. Int. Conf. Intell. Comput., Inf. Control Syst.* Cham, Switzerland: Springer, 2021, pp. 117–123.
[14] J. S. Yang, M.-S. Ko, and K. S. Chung, "Social emotional opinion decision with newly coined words and emoticon polarity of social networks services," *Future Internet*, vol. 11, no. 8, p. 165, Jul. 2019.

[15] F. Kou, J. Du, Z. Lin, M. Liang, H. Li, L. Shi, and C. Yang, "A semantic modeling method for social network short text based on spatial and temporal characteristics," *J. Comput. Sci.*, vol. 28, pp. 281–293, Sep. 2018.
[16] L. Shi, J. Du, M. Liang, and F. Kou, "Dynamic topic modeling via self-aggregation for short text streams," *Peer-Peer Netw. Appl.*, vol. 12, no. 5, pp. 1403–1417, Sep. 2019.
[17] L. Shi, G. Song, G. Cheng, and X. Liu, "A user-based aggregation topic model for understanding user's preference and intention in social network," *Neurocomputing*, vol. 413, pp. 1–13, Nov. 2020.
[18] E. Amigó, J. C. de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín, E. Meij, M. D. Rijke, and D. Spina, "Overview of replab 2013: Evaluating online reputation monitoring systems," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2013, pp. 333–352.
[19] E. Amigó, J. C.-D. Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. D. Rijke, and D. Spina, "Overview of replab 2014: Author profiling and reputation dimensions for online reputation management," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2014, pp. 307–322.
[20] L. Abednego, C. E. Nugraheni, and A. Fedora, "Forex sentiment analysis with Python," *Int. J. Adv. Res. Econ. Finance*, vol. 4, no. 1, pp. 46–55, 2022.
[21] T. A. Lampert, A. Stumpf, and P. Gançarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2557–2572, Jun. 2016.
[22] J. Cerda L and L. Villarroel Del P, "Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa," *Revista Chilena de Pediatria*, vol. 79, no. 1, pp. 54–58, Feb. 2008.
[23] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The Kappa statistic," *Family Med.*, vol. 37, no. 5, pp. 360–363, 2005.
[24] J. Carrillo-de-Albornoz, J. Rodríguez Vidal, and L. Plaza, "Feature engineering for sentiment analysis in e-health forums," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0207996.
[25] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, pp. 41–46, 2001.
[26] S. J. Hickey, "Naive Bayes classification of public health data with greedy feature selection," *Commun. IIMA*, vol. 13, no. 2, p. 7, Jun. 2014.
[27] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with naive Bayes for text classification," *Appl. Soft Comput.*, vol. 69, pp. 344–356, Jan. 2018.
[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[29] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature Biotechnol.*, vol. 26, no. 9, pp. 1011–1013, Sep. 2008.
[30] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[31] C. Schaffer, "Selecting a classification method by cross-validation," *Mach. Learn.*, vol. 13, no. 1, pp. 135–143, Oct. 1993.
[32] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1. Oxford, U.K.: Academic, 2018, pp. 542–545. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/B978012809633820349X?via%3Dihub
[33] B. Josúe and O. Carrer, "Datificate GPT2 small Spanish model," 2022.
[34] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained BERT model and evaluation data," in *Proc. PML4DC ICLR*, 2020, pp. 1–9.
[35] A. Gutierrez-Fandino, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas, "Maria: Spanish language models," 2021, *arXiv:2107.07253*.
[36] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Exp. Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016.
[37] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. AAAI Workshop Imbalanced Data Sets*, vol. 68, 2000, pp. 1–3.
[38] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Comput. Inform.*, vol. 34, no. 5, pp. 1017–1037, 2016.
[39] A. Ghazikhani, H. S. Yazdi, and R. Monsefi, "Class imbalance handling using wrapper-based random oversampling," in *Proc. 20th Iranian Conf. Electr. Eng. (ICEE)*, May 2012, pp. 611–616.
[40] X. Ying, "An overview of overfitting and its solutions," *J. Phys., Conf.*, vol. 1168, Feb. 2019, Art. no. 022022.
[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[42] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[43] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Instructional Conf. Mach. Learn.*, 2003, vol. 242, no. 1, pp. 29–48.

ALMUDENA SÁNCHEZ RUÍZ received the degree in industrial engineering with Universidad Politécnica de Madrid, Spain, and the master's degree, in February 2022, with a focus on automation and electronics. She is currently a Railway Engineer with CAF Signalling. She is a Tester Engineer of a signaling component of the European Train Control System (ETCS). During her last academic year, she participated an Exchange Program with Aalto University, Helsinki, Finland. For her master's thesis, she focused on research about natural language processing and machine learning techniques.

DANIEL GALAN received the M.Sc. degree in automation and robotics from the Polytechnic University of Madrid, in 2013, and the Ph.D. degree in computer engineering and automatic control from Universidad Nacional de Educación a Distancia (UNED), in 2017. Currently, he is with the Centre of Automation and Robotics, Universidad Politécnica de Madrid (UPM)–CSIC. He is an Assistant Professor with UPM. His research interests include social robotics, virtual and remote laboratories, and intelligent control.

ÁNGEL GARCÍA-BELTRÁN has been taught many programming courses with ETSII–UPM, since 1992. He has been a part of AulaWeb development, since 1998. He is currently an Associate Professor with UPM. His research interests include computer, multimedia, and web-based educational systems, and numerical models for engineering applications.

JAVIER RODRÍGUEZ-VIDAL received the Ph.D. degree in intelligent systems from Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, in October 2019. He is currently an Assistant Professor with Universidad Politécnica de Madrid (UPM). His main research interests include the fields of social media, machine learning, automatic summarization, and laser additive construction.