International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Classification Model Based on Supervised Learning in the Constituent Context of the Ayacucho Region, Peru.

Yordan Sullca-Palomino[1], Yudi Guzmán-Monteza[2]

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: Using data mining techniques, data from platforms such as Twitter (now called X) represent a valuable opportunity to analyze preferences, specifically in discussing political and social issues. In this study, a text classification model designed to categorize content related to the constituent process in Ayacucho was developed. Using data collected from Twitter, we sought to classify text as 'constituent' or 'non-constituent'. Supervised learning techniques (SVM, RF, and NB) were applied along with three vectorization methods (BOW, TF-IDF, and W2V). An annotation process was established to label classes, ensuring data reliability with a Kappa coefficient 0.72. The data were divided into training, test, and validation sets. Data Augmentation strategies were explored to address data imbalance. Experimental results on the validation dataset revealed that the SVM classification model obtained the highest F1 score, reaching a value of 0.74, outperforming other evaluated models. The findings of this study offer valuable insights for other researchers facing similar challenges in niche-specific text classification. Both the annotation methodology employed and the effectiveness of the classification techniques, together with an approach focused on continuous improvement, lay a solid foundation for future projects in this field.

Keywords: *Supervised learning, Data mining, Constituent process, Annotation Rules, Data mining.*

1. Introduction

In recent years, a growing popular discontent in Peruvian politics has generated the need for constitutional reforms that include citizen participation. According to the Americas Barometer 2018/2019 report, there is a reduction in public satisfaction with the functioning of democracy, while there is an increase in the acceptance of actions by the executive branch, such as the closing of Congress. There was an increase between 2017 and early 2019 in the proportion of public opinion in Peru that considers it acceptable for the president to dissolve Congress, a change in attitude that anticipated the current crisis. Political legitimacy, measured by public trust and respect for the country's fundamental institutions and processes, remains low on average in the United States. [1]

Constitutional reform demands the population's participation and can be executed through referendums and other mechanisms. According to a 2018 JNE report, a national referendum constitutes a citizen participation mechanism in which the approval or rejection of a parliamentary bill that partially modifies the constitution is submitted to the citizens' consideration [2].

This reform requires popular participation and can be carried out through referendums and other means. According to a 2018 JNE report, a national referendum is a citizen participation mechanism through which citizens are consulted on the approval or rejection of a parliamentary bill that partially reforms the constitution [2]. The popular consultation process demands significant resources and requires an extensive period for its organization and execution, and it is essential for the country to find viable solutions to carry it out efficiently.

The growing popularity of Online Social Networks (OSNs) has facilitated the prediction of user characteristics, allowing researchers to access previously inaccessible data. Analyzing digital records of human behavior in these networks has become a valuable source of information about individuals and social dynamics [4]. The objective is to identify user characteristic patterns through the interactions collected on these platforms [5].

Twitter, now called X, will be referred to as Twitter throughout this article to facilitate understanding of the research. This decision is based on the data collection period from 2021 to mid-2023, during which the platform was still called Twitter. The X network has approximately 354 million active users worldwide and is widely used for microblogging (2023). Users can share thoughts and preferences through real-time Tweets. The appeal as a research source lies in the public nature of most accounts and the limited restrictions on the data collection platform, making it a safe option for academic exploration while respecting the user's privacy. The extraction of data from public accounts on Twitter stands out for its efficiency, thanks to the availability of its website for developers and its APIs, which represents an advantage over other platforms where data collection is usually more complex. In addition, it is important to note that Twitter presents a

[1] National University of San Marcos, PERU
Email: yordan.sullca@unmsm.edu.pe.
ORCID ID: 0009-0003-0809-1285
[2] National University of San Marcos, PERU
ORCID ID: 0000-0001-5306-5295
Email: yudi.guzman@unmsm.edu.pe.

remarkable politicization [6], with a higher probability of its users expressing political opinions than other platforms. It is important to note that the data collected for this research covers the period from 2021 to mid-2023, before the platform's name change. This study represents a significant contribution to data mining applied to text classification in specific political and social environments in the Spanish language. The main contribution lies in a meticulous data annotation and labelling process, followed by implementing and evaluating supervised learning models. Algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) are explored to carry out the classification of texts related to the constituent process in the region of Ayacucho, using data extracted from Twitter as a primary source.

The paper is organized as follows: Section II represents the current state of the field and previous work addressing data mining techniques, supervised learning, and annotation process. Section III details the methodologies employed for constructing the dataset, including the annotation process and data vectorization. Section IV describes the adjustments made to the dataset for model experimentation, while Section V represents the performance and results obtained by the classification models. Section VI is devoted to the discussion, and Section VII concludes the paper.

2. Related Work

Classifying texts related to the constituent process has been little researched despite the emphasis on argument mining in various fields. [14]. This lack of specific focus offers opportunities for research analysing social network texts and other relevant sources. The concept of “constituent process” was proposed by Roberto Arroyo (2021) during the implementation of the web platform at the service of citizens for the constituent process of the people of Peru.

Different annotation approaches have been proposed in Spanish text notation to ensure accuracy in classifying classes of tweets. For example, [23] conducts manual annotation on hate-related tweets. [14] presents a robust method for annotating argumentative text in Spanish on Twitter, employing Cohen Kappa to measure reliability, with values between 0.52 and 0.75, and emphasizing discursive markers. In addition, [13] proposes guidelines for annotating Spanish texts on Twitter, achieving a concordance index of 0.84 and using labeling with classification models. This study highlights the importance of the Kappa coefficient and the construction guideline in annotation.

The evaluation of binary classification models has been addressed by using different metrics, [19], [20], and [21] suggest the use of Accuracy and F1-Score for binary classification, especially in cases of class imbalance.

Several studies have explored text classification in social networks through data mining and supervised learning techniques, offering significant contributions to analyzing political trends and opinions. These research, such as those of [7] on political trends in Quito, [8] on diverse social platforms, and [10] on political messages on Twitter in Spanish, have employed combined qualitative and quantitative analysis methods to classify and understand users' political trends.

Other approaches, such as that of [9] in analyzing trends about Eurovision on Twitter using adaptive data mining techniques, highlight the complexity of predicting outcomes based on the mentioned frequencies alone. In [12] they presented a monitoring and early warning system for public opinions, employing improved algorithms to detect frequent patterns and classify text. The study [25] focuses on classifying tweets related to positions on opening or closing the COVID-19 pandemic. Data collected from Twitter and a technique called Easy Data Augmentation (EDA) are used. This technique helped to balance the data samples by increasing their volume through operations such as synonym replacement, insertion, swapping, and random deletion. This approach improved the performance of the model used in the study.

For preprocessing and feature vectorization, different studies employ varied approaches. In [17], tweets are classified as ideas or not using BoW with TF-IDF, BERT, supervised and semi-supervised models. In [15], fake news is detected using a count-vectorizer, n-gram model, TF-IDF with SVM, Naive Bayes, Random Forest, and Logistic Regression. [22] classifies tweets into news, conversations, questions, or wishes with TF-IDF and Word2Vec, adapting preprocessing for Arabic texts and SMOTE techniques. [16] focuses on binary and categorical classification using geocoding and counting vectors and TF-IDF with supervised models. [22] focuses on building a dataset for news detection using count vectors and TF-IDF with supervised models.

Various algorithms have been explored in the context of the performance evaluation of Twitter classification models. [11] performed a supervised analysis with Twitter tools, evaluating algorithms with a set of 7,000 manually tagged tweets in categories, where the Multinomial NB algorithm obtained the highest score with 75.81%, followed by LSVM (72.76%) and Logistic Regression (72.25%). [21] uses RF, SVM, NB and DT for binary classification of tweets about environmental hazards with an accuracy of 85.13% and F1-Score of 85% with RF. In [19], TF-IDF, W2Vec and GloVe with Logistic Regression, SVM and NB are used to classify gender, obtaining an accuracy of 57.14% with LR and W2Vec. [24] uses AdaBoost, XGBoost, SVM and NB to classify user gender on Instagram, achieving an accuracy of 78.64% with NB. 20] classifies political users with SVM, LR, SGD, RF and XGBoost, highlighting LR with accuracy of 0.93% and F1-Score of 0.92% after hyperparameter adjustment.

These studies evidence the effectiveness of data mining and supervised learning in Twitter text classification, providing valuable insights. Together, they establish a comprehensive framework for future research in this area.

3. Materials and Methods

This section describes the materials used in the research and details the methods used to carry out the study. A list of the resources used is provided, followed by a detailed explanation of the procedures and techniques used in the research.

3.1. Methodology

This section will detail the methods used, taking as a guide the scheme represented in Figure 1. The procedures used in different phases, from the dataset's creation to the model's interpretation, will be described.

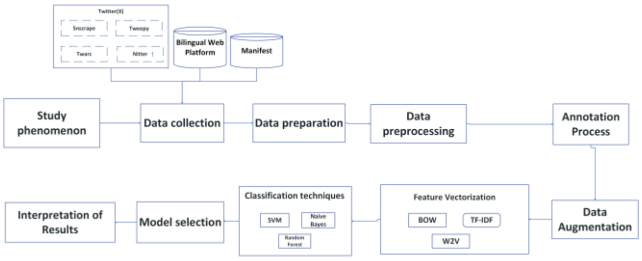

Fig. 1. High-level framework for binary text classification models.

3.2. Dataset construction

The construction of the dataset involved the selective extraction of tweets published by users from the province of Ayacucho through various Twitter APIs, including Snscrape, Tweepy, Twarc, and Nitter, as shown in Figure 1. It began with the definition of 153 keywords, based on an exhaustive review of the constitution, relevant documents, and digital media related to political and social issues linked to the constituent process, to guide the collection of these comments. This was complemented by including records from a dataset of a bilingual web platform for citizen participation [18] and a document produced by the social-political movement "Nueva Republica" in September 2021.

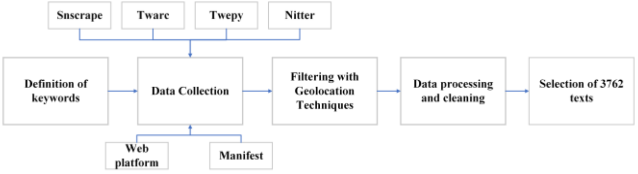

Fig. 2. Dataset Construction Diagram

The search was geographically delimited using geocoding parameters and the location of users who posted tweets in the Ayacucho region. The collection was carried out through Snscrape, a library that performs web scraping on Twitter searches to collect public domain publications. A total of 6600 records from the last five years were obtained through this API.

The Tweepy and Twarc tools required the official Twitter API, from which data was collected exclusively from 2021, totaling 19786 records. In addition, Nitter, an alternative Twitter platform with an API for data extraction, was used, obtaining information from 2021 and collecting 1047 records.

This set was complemented by including 84 records from a citizen service platform for the constituent process of the peoples of Peru, collected between 2020 and 2021.

The attributes shown in Table 1 were considered in the collection of Twitter data.

Table 1. Attributes

Attribute	*Description*
Id	Unique identifier of each tweet
Handles	Unique usernames used on the platform to identify each user
FechaPost	The date and time when the tweet were posted.
Tweet	The content of the tweet
authorLocation	Location of user who posted the tweet

3.3. Preparation, Preprocessing and Vectorization of data.

After data preparation, a significant reduction in the size of the dataset was observed due to the filter applied to select only records related to the Ayacucho region. This filter was implemented both manually and by using regular expressions (regex) to identify and select comments that originated specifically in the Ayacucho region. The location attributes (authorLocation) in the users' profiles were reviewed and a manual inspection of their descriptions and information was carried out to confirm their connection to the Ayacucho region. In addition, tweets that did not contain relevant information were purged. As a result of this process, a more refined dataset was obtained, with a total of 3578 records that meet the established criteria. These records are detailed in Table 2 for reference.

Table 1. Pre-processed Data Set

Sources used	*Original Dataset*	*Pre-processed Dataset*
Snscrape	6600	2501
Twarc	5625	388

Twepy	14161	345
Nitter	1047	344
Web Platform	84	84
Manifiesto	100	100

The preprocessing phase was divided into two stages: the first focused on preparing the data for annotation, while the second completed the post-annotation processing.

The first part sought clean text by removing null values, checking data integrity, deleting duplicates, and manual spell-checking. Initial text cleaning was also carried out, eliminating mentions (@), links (http, www), hashtags (#), special characters, escape characters and emojis.

The second stage was dedicated to finalizing the preprocessing by tokenization, digit, and empty word removal. NLTK was used for stopwords, with its initial set of empty words in Spanish.

Subsequently, feature vectorization was performed using techniques such as Bag of Words (BoW), TF-IDF and Word2Vec. These techniques allowed the transformation of processed texts into numerical representations, facilitating their use in machine learning algorithms.

3.4. Annotation Methodology

In this phase, the annotation process was carried out, as illustrated in Figure 2, with the purpose of labeling a corpus into two classes: 'constituent' or 'non-constituent'. The annotators performed the annotation following a double-blind procedure, a practice in which the parties involved are unaware of the previous work done by the other annotator. There were two annotators, one of whom played the role of supervisor. Communication was absent during the annotation process in order to maintain levels of impartiality. In addition, for the interpretation of the Kappa Coefficient, we focused on the table in the paper [13].

The annotation process led to the iterative development of an annotation guide that reached its final version in 5.0. This guide is documented in detail in Table 3.

Table 2. Annotation Rules v5.0

Class	*Id*	*Rule*
Constituent	1	Text demand or demand or a claim regarding issues related to the constituent process.
	2	Text expressing a perspective, vision, or expectation of change regarding issues of the constituent process.
	3	The text is intended to express a denunciation or malpractice of situations, illegal acts, or improper practices related to constituent issues.
	4	The text expresses criticism or denunciation regarding constituent issues.
	5	Text expressing the need or concern for specific policies and actions related to the constituent theme.
	6	The text expresses a demand, concern, or claim related to protecting, guaranteeing, or promoting rights.
	7	The text includes a news item on the subject of the constituent issue.
	8	The text informs about events, debates, proposals, or any kind of aspect relevant to constituent issues.
	9	Text that includes premises referring to constituent subjects or the above rules.
No Constituent	1	An informative text is presented that offers a news item, but it is not related to the constituent issue.
	2	The text announces a product, service, event, or initiative that is unrelated to the constituent theme.
	3	Text that expresses the intention to carry out an action or carry out a project unrelated to the constituent topic.
	4	This text contains an opinion, mention, or comment of a personal nature.
	5	Text that expresses a feeling without any connection or reference to the constituent topic.
	6	Text that does not express any content.
	7	Text that expresses a statement, complaint, or denunciation of a specific situation not related to the constituent topic.
	8	Text that expresses informative text not related to the constituent topic.

During the last annotation process, a Kappa coefficient of 0.72 was reached, exceeding the threshold established in [13] and [14], set at 0.61 to be considered substantial. This significant consistency between annotators highlights the reliability of the annotated data, validating the effective application of the annotation rules and supporting the acceptability of the results obtained.

To understand some of the annotation rules, some illustrative examples will be presented:

- Rule 2 of constituent class: “investing in education an alternative solution, it is not possible for a police officer to have a higher remuneration than a teacher”, this rule complies with text that expresses an expectation of change regarding issues of the constituent process.

- Rule 7 of the constituent class: “first you murder and then you sign agreements, agreements are signed in vain for the photo. What of the 1993 constitutional agreement in its art 1 indicates the supreme purpose of the state is the person and their right to a dignified life you murdered to more than 60 innocent Peruvians”

- Rule 1 of non-constituent class: “private and access to markets. The ceremony had the participation of authorities, officials and artisans of the region who met to make agreements from the artisan sector close to Holy Week. This event took place on the 24th. ", text reports on a news item unrelated to the constituent issue.

- Rule 3 of non-constituent class: “a long time ago we decided that all the proceeds from the trekking outings of this year 2018 would be used to carry out social work and that is why we decided together with Erika Hugo and Sergio…”, the text expresses the intention of Carrying out social work is not a topic that discusses any issue of the constituent process.

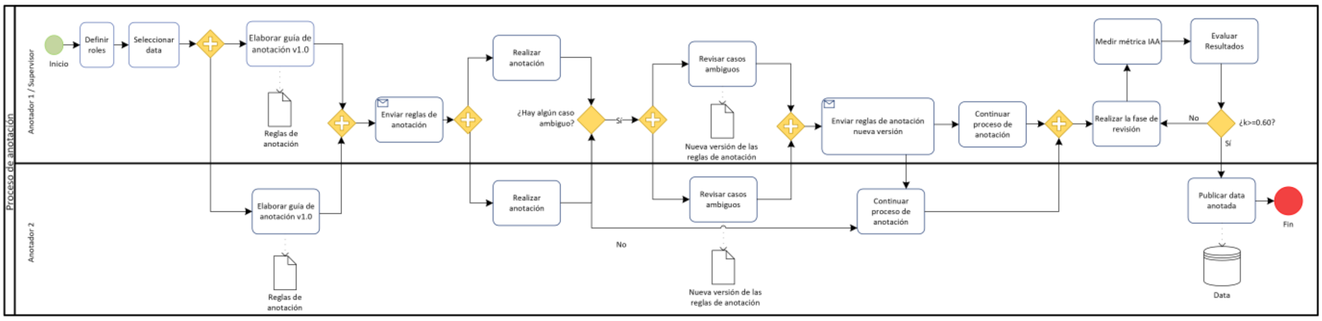

Fig. 3. Annotation Process

4. Experiments

In this section, the model's experimental setup is described. It was evaluated using a data set partitioned into three subsets: training, testing, and validation. The adjustment of hyperparameters, the distribution of data in three different scenarios, and the implementation of data augmentation techniques were included.

4.1. Hyper Fitting Parameters

We optimize the performance of each classifier by adjusting its hyperparameters based on the specific vectorization technique used through Grid Search using the scikit-learn library. Table 4 details the classifiers and the hyperparameters adjusted for each vectorization technique in our study with their respective model.

Table 3. Hyper fitting Parameters

Model	*Hyperparameters*	*Tuned value*		
		BOW	*TF-IDF*	*W2V*
SVM	kernel	sigmoid	rbf	sigmoid
	gamma	0.001	0.01	0.001
	C	50	25	1000
RF	n_estimators	200	100	200
	max_depth	None	40	None
	min_samples_split	5	5	10
	min_samples_leaf	1	2	4
GNB	priors	[0.1, 0.9]	None	[0.1, 0.9]
	var_smoothing	1,00E-06	1,00E-09	1,00E-09

4.2. Data Distribution

There were three sets of training, testing and validation data. The base data set was split using a random approach using the Scikit-learn library. 80% of the base data set was allocated to the training set, while the remaining 20% was allocated to the test set. Additionally, 50 new samples were added for each class in the validation data set with a total of 100 instances for the validation set. This validation data was not used in training or testing the model, ensuring an unbiased evaluation.

The experimentation was divided into three scenarios. In the first scenario, no data augmentation techniques were applied to the sets. In the second scenario, data augmentation was used only on the training and test data to address imbalances, leaving the validation set unaltered. The third scenario implemented data augmentations on all sets, including validation data, to analyze their influence on model performance. The data distribution for these scenarios is presented in Table 5.

Table 4. Data Distribution Table

Scenario	*Training*	*Test*	*Validation*	*Data Augmentation (DA)*
1	766	192	100	DA was not applied
2	1566	392	100	DA was applied to the training and test set.
3	1566	392	200	DA was applied to the training, testing and validation set.

5. Results

This section presents the results obtained for each scenario, showing the performance metrics of the models and vectorization methods used. From now on, the term 'Class 1' will refer to the 'constituent' category, while 'Class 0' will refer to the 'non-constituent' category to simplify the presentation of results and avoid unnecessary repetitions.

5.1. Scenario 1

The original data were maintained without applying augmentation techniques.

5.1.1. Training and Test Data Results

For the SVM model, it was observed that vectorization with W2V achieves high performance in class 0 with an F1-score of 85.29%, while in class 1, vectorization with W2V vectorization has an F1-score of 83.3. %. On the other hand, the RF model shows outstanding performance in class 0 with TF-IDF vectorization, reaching an F1-score of 82.67%, while in class 1, vectorization with W2V obtains an F1-score of 77.77%. On the other hand, the Naive Bayes (NB) model performs better in terms of F1-score when using W2V vectorization for both classes, obtaining 78.50% for class 0 and 72.94% for class 1.

The results in Table 6 suggest that the SVM model achieves the best F1 Score in class 0 with W2V vectorization for both classes.

5.1.2. Training and Validation Data Results

For the SVM model, it was observed that vectorization with BOW achieves high performance in class 0 with an F1-score of 74.28%, while in class 1, vectorization with BOW has an F1-score of 71.58%. On the other hand, the RF model shows outstanding performance in class 0 with BOW vectorization, reaching an F1-score of 67.96%, while in class 1, vectorization with BOW obtains an F1-score of 65.98%. On the other hand, the NB model performs better for class 0 in terms of an F1-score of 62.30% with W2V vectorization, and for class 1, an F1-score of 66.67% is obtained for TF-IDF vectorization.

The results in Table 7 suggest that the SVM model achieves the best F1 Score for both classes.

5.2. Scenario 2

Data augmentation was applied using NLTK and NLPAug to generate 1000 additional instances: 500 for 'constituent' and 500 for 'non-constituent' in training and testing. Validation data was left unchanged to evaluate model performance.

5.2.1. Training and Test Data Results

The results observed by applying the data augmentation technique in the different models are: the SVM model, it is observed that vectorization with TF-IDF achieves high performance in class 0 with an F1-score of 95.43%, while in class 1, vectorization with TF-IDF has an F1-score of 95.38%. On the other hand, the RF model shows outstanding performance in class 0 with TF-IDF vectorization, reaching an F1-score of 94.76%, while in class 1, vectorization with TF-IDF obtains an F1-score of 94.51%. On the other hand, the NB model performs better in terms of F1-score when using TF-IDF vectorization for both classes, obtaining 88.53% for class 0 and 89.48% for class 1.

The results in Table 8 suggest that the SVM model with TF-IDF vectorization achieves the best F1 Score for both classes.

5.2.2. Training and Validation Data Results

The results with the validation data without applying data augmentation are the following: the SVM model, it is observed that the vectorization with W2V achieves better performance in class 0 with an F1-score of 65.71%, while in class 1, the vectorization with BOW has an F1-score of 67.85%. On the other hand, the RF model shows outstanding performance in class 0 with TF-IDF vectorization, reaching an F1-score of 68.05%, while in class 1, vectorization with BOW obtains an F1-score of 57.40%. . On the other hand, the NB model performs better in terms of F1-score when using BOW and W2V vectorization for classes 0 and 1, obtaining 88.53% for class 0 and 89.48% for class 1.

The results in Table 9 suggest that better results are obtained for class 1 using SVM with BOW vectorization, while better results are obtained for class 0 with the RF model with TF-

IDF vectorization.

5.3. Scenario 3

Data augmentation was applied using a synonym augmenter to generate an additional 1100 data points: 500 for ‘constituent’ and ‘500’ for ‘non-constituent’ in the training and test sets. Additionally, 50 instances were added for each class in the validation data, with a total of 100 instances per class.

5.3.1. Training and Test Data Results

The results applied with the increase in validation data are the following:

The SVM model shows that the vectorization with TF-IDF achieves better performance in class 0 with an F1-score of 95.43%, while in class 1, the vectorization with TF-IDF has an F1-score of 95.38%. On the other hand, the RF model shows outstanding performance in class 0 with BOW vectorization, reaching an F1-score of 94.14%, while in class 1, vectorization with BOW obtains an F1-score of 94.11%. On the other hand, the NB model performs better in terms of F1-score when using TF-IDF vectorization for both classes, obtaining 89.12% for class 0 and 59.54% for class 1.

The results in Table 10 suggest that better results are obtained in the SVM model for both classes with TF-IDF vectorization.

5.3.2. Training and Validation Data Results

The results applied with the increase in validation data are the following:

In the SVM model, it is observed that vectorization with TF-IDF achieves better performance in class 0 with an F1-score of 66.67%, while in class 1, vectorization with BOW has an F1-score of 70.48%. On the other hand, the RF model shows outstanding performance in class 0 with TF-IDF vectorization, reaching an F1-score of 67.15%, while in class 1, vectorization with BOW obtains an F1-score of 60.67%. On the other hand, the NB model performs better in terms of F1-score when using TF-IDF vectorization for class 1 and W2V for class 0, obtaining 61.36% for class 0 and 64% for class 1.

The results in Table 11 suggest that better results are obtained in the SVM and RF models for classes 1 and 0 with BOW and TF-IDF vectorization.

5.4. Selection of the best model

Attention is focused on the sub-scenario that encompasses training and validation data to evaluate the performance of the models in the three scenarios. The best results obtained are presented in Table 12, highlighting the most outstanding model for Classes 1 and 0, the SVM (Support Vector Machine) of Scenario 1 using BOW vectorization.

Using BOW vectorization in Scenario 1, the SVM model achieves an F1-Score of 0.712857 and an accuracy of 73%. These values, detailed in Table 12, highlight the model's ability to make accurate predictions for Class 1.

Table 6. Table of training and test data results in Scenario 1.

Model	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*
SVM	BOW	0	0.776042	0.794393	0.801887	0.798122
	BOW	1	0.776042	0.752941	0.744186	0.748538
	TF-IDF	0	0.791667	0.875000	0.726415	0.793814
	TF-IDF	1	0.791667	0.721154	0.872093	0.789474
	W2V	0	0.843750	0.887755	0.820755	0.852941
	W2V	1	0.843750	0.797872	0.872093	0.833333
RF	BOW	0	0.776042	0.811881	0.773585	0.792271
	BOW	1	0.776042	0.736264	0.779070	0.757062
	TF-IDF	0	0.796875	0.781513	0.877358	0.826667
	TF-IDF	1	0.796875	0.821918	0.697674	0.754717
	W2V	0	0.760417	0.884615	0.650943	0.750000
	W2V	1	0.760417	0.675439	0.895349	0.770000
NB	BOW	0	0.682292	0.777778	0.594340	0.673797
	BOW	1	0.682292	0.612613	0.790698	0.690355
	TF-IDF	0	0.604167	0.714286	0.471698	0.568182
	TF-IDF	1	0.604167	0.540984	0.767442	0.634615
	W2V	0	0.760417	0.777778	0.792453	0.785047
	W2V	1	0.760417	0.738095	0.720930	0.729412

Table 5. Table of training and validation data results in Scenario 1.

Model	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*

SVM	BOW	0	0.73	0.709091	0.78	0.742857
	BOW	1	0.73	0.755556	0.68	0.715789
	TF-IDF	0	0.67	0.680851	0.64	0.659794
	TF-IDF	1	0.67	0.660377	0.70	0.679612
	W2V	0	0.54	0.522727	0.92	0.666667
	W2V	1	0.54	0.666667	0.16	0.258065
RF	BOW	0	0.67	0.660377	0.70	0.679612
	BOW	1	0.67	0.680851	0.64	0.659794
	TF-IDF	0	0.52	0.513158	0.78	0.619048
	TF-IDF	1	0.52	0.541667	0.26	0.351351
	W2V	0	0.48	0.487500	0.78	0.600000
	W2V	1	0.48	0.450000	0.18	0.257143
NB	BOW	0	0.63	0.644444	0.58	0.610526
	BOW	1	0.63	0.618182	0.68	0.647619
	TF-IDF	0	0.62	0.666667	0.48	0.558140
	TF-IDF	1	0.62	0.593750	0.76	0.666667
	W2V	0	0.48	0.488636	0.86	0.623188
	W2V	1	0.48	0.416667	0.10	0.161290

Table 6. Table of training and test data results in Scenario 2.

Model	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*
SVM	BOW	0	0.943878	0.934343	0.953608	0.943878
	BOW	1	0.943878	0.953608	0.934343	0.943878
	TF-IDF	0	0.954082	0.940000	0.969072	0.954315
	TF-IDF	1	0.954082	0.968750	0.939394	0.953846
	W2V	0	0.903061	0.878641	0.932990	0.905000
	W2V	1	0.903061	0.930108	0.873737	0.901042
RF	BOW	0	0.941327	0.934010	0.948454	0.941176
	BOW	1	0.941327	0.948718	0.934343	0.941476
	TF-IDF	0	0.946429	0.917874	0.979381	0.947631
	TF-IDF	1	0.946429	0.978378	0.914141	0.945170
	W2V	0	0.926020	0.927461	0.922680	0.925065
	W2V	1	0.926020	0.924623	0.929293	0.926952
NB	BOW	0	0.831633	0.783186	0.912371	0.842857
	BOW	1	0.831633	0.897590	0.752525	0.818681
	TF-IDF	0	0.890306	0.917127	0.855670	0.885333
	TF-IDF	1	0.890306	0.867299	0.924242	0.894866
	W2V	0	0.742347	0.778443	0.670103	0.720222
	W2V	1	0.742347	0.715556	0.813131	0.761229

Table 7. Table of training and validation data results in Scenario 2.

Model	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*
SVM	BOW	0	0.64	0.684211	0.52	0.590909

	BOW	1	0.64	0.612903	0.76	0.678571
	TF-IDF	0	0.60	0.580645	0.72	0.642857
	TF-IDF	1	0.60	0.631579	0.48	0.545455
	W2V	0	0.52	0.511111	0.92	0.657143
	W2V	1	0.52	0.600000	0.12	0.200000
RF	BOW	0	0.54	0.547619	0.46	0.500000
	BOW	1	0.54	0.534483	0.62	0.574074
	TF-IDF	0	0.54	0.521277	0.98	0.680556
	TF-IDF	1	0.54	0.833333	0.10	0.178571
	W2V	0	0.51	0.505376	0.94	0.657343
	W2V	1	0.51	0.571429	0.08	0.140351
NB	BOW	0	0.63	0.638298	0.60	0.618557
	BOW	1	0.63	0.622642	0.66	0.640777
	TF-IDF	0	0.58	0.611111	0.44	0.511628
	TF-IDF	1	0.58	0.562500	0.72	0.631579
	W2V	0	0.50	0.500000	0.92	0.647887
	W2V	1	0.50	0.500000	0.08	0.137931

Table 8. Table of training and test data results in Scenario 3.

Model	***Vectorization***	***Class***	***Accuracy***	***Precision***	***Recall***	***F1-score***
SVM	BOW	0	0.64	0.684211	0.52	0.590909
	BOW	1	0.64	0.612903	0.76	0.678571
	TF-IDF	0	0.60	0.580645	0.72	0.642857
	TF-IDF	1	0.60	0.631579	0.48	0.545455
	W2V	0	0.52	0.511111	0.92	0.657143
	W2V	1	0.52	0.600000	0.12	0.200000
RF	BOW	0	0.54	0.547619	0.46	0.500000
	BOW	1	0.54	0.534483	0.62	0.574074
	TF-IDF	0	0.54	0.521277	0.98	0.680556
	TF-IDF	1	0.54	0.833333	0.10	0.178571
	W2V	0	0.51	0.505376	0.94	0.657343
	W2V	1	0.51	0.571429	0.08	0.140351
NB	BOW	0	0.63	0.638298	0.60	0.618557
	BOW	1	0.63	0.622642	0.66	0.640777
	TF-IDF	0	0.58	0.611111	0.44	0.511628
	TF-IDF	1	0.58	0.562500	0.72	0.631579

	W2V	0	0.50	0.500000	0.92	0.647887
	W2V	1	0.50	0.500000	0.08	0.137931

Table 9. Table of training and validation data results in Scenario 3.

Model	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*
SVM	BOW	0	0.665	0.726027	0.53	0.612717
	BOW	1	0.665	0.629921	0.80	0.704846
	TF-IDF	0	0.680	0.695652	0.64	0.666667
	TF-IDF	1	0.680	0.666667	0.72	0.692308
	W2V	0	0.540	0.522989	0.91	0.664234
	W2V	1	0.540	0.653846	0.17	0.269841
RF	BOW	0	0.585	0.595506	0.53	0.560847
	BOW	1	0.585	0.576577	0.64	0.606635
	TF-IDF	0	0.555	0.532164	0.91	0.671587
	TF-IDF	1	0.555	0.689655	0.20	0.310078
	W2V	0	0.525	0.515152	0.85	0.641509
	W2V	1	0.525	0.571429	0.20	0.296296
NB	BOW	0	0.610	0.614583	0.59	0.602041
	BOW	1	0.610	0.605769	0.63	0.617647
	TF-IDF	0	0.595	0.626667	0.47	0.537143
	TF-IDF	1	0.595	0.576000	0.72	0.640000
	W2V	0	0.490	0.493902	0.81	0.613636
	W2V	1	0.490	0.472222	0.17	0.250000

Table 10. Table of the best models in three scenarios

Model	*Scenario*	*Vectorization*	*Class*	*Accuracy*	*Precision*	*Recall*	*F1-score*
SVM	1	BOW	0	0.73	0.709091	0.78	0.742857
RF	2	TF-IDF	0	0.54	0.521277	0.98	0.680556
RF	3	TF-IDF	0	0.555	0.532164	0.91	0.671587
SVM	1	BOW	1	0.73	0.755556	0.68	0.715789
SVM	2	BOW	1	0.64	0.612903	0.76	0.678571
SVM	3	BOW	1	0.665	0.629921	0.80	0.704846

6. Discussion

The results obtained validate the coherence with our initial objectives and highlight the effectiveness of our constituent text classification models. We have prepared a thorough comparison with related research, highlighting essential aspects such as the annotation rules detailed in Table 3, the evaluation of the IAA using the Cohen's Kappa metric described in Table 13, and the evaluation of the performance of binary classification model, presented in Table 14, considering key metrics such as Accuracy and F1-score.

Table 11. Comparison of Cohen's Kappa Metrics in Previous Studies

C	*Research Study*	*Metrics*
1	Urpay-Camasi, J [13]	K = 0.84
2	Guzmán-Monteza, Y [14]	K = (0.73, 0.52, 0.75)
3	Our Study	K = 0.72

The annotation methodology we implemented was based on schemes proposed by Urpay-Camasi [13] and Guzmán-Monteza, Y [14], whose research facilitated adapting our study phenomenon in order to label tweets as constituent or non-constituent. The study [13] achieved a Cohen's Kappa result of 0.84, indicating a level close to perfection in the annotation process. On the other hand, in [14], where three annotators were involved, the results of Cohen's Kappa were (0.73, 0.52, 0.75), suggesting substantial annotation as a whole.

When comparing our results with the study [13], we noticed a similarity with a difference of 0.12 in the findings. This discrepancy can be attributed to the inherent complexity in labeling texts related to the constituent process because it is a sociopolitical phenomenon. In our case, we sought to capture specific nuances rather than opting for simplistic labeling based solely on positivity, negativity, or neutrality. This choice may have contributed to the variability in Kappa results by addressing the diversity and complexity of nuances in the analysis.

On the other hand, when analyzing the study [14], a discrepancy is observed in the kappa values (-0.01, 0.20 and -0.03). The adoption of strategies similar to those used in said study, such as syntactic, semantic and morphological detection, could have contributed to improving our results. It is also important to consider that variations in annotation criteria and the number of annotators involved in each study could explain the differences in the observed kappa values.

In the discussion about our classification models, binary classification studies were considered due to the limitation of studies directly related to our object of study. For example, the work of Ozcan, S. [17] focused on extracting and detecting tweets to ideate innovations. He achieved better results with the BERT vectorization and the SVM model, they achieved an F1-Score of 56%. In contrast, our work showed a significant improvement of 15% over the F1-Score. Likewise, the study by Onikoyi, B. [13] was considered, which focused on classifying the user's gender based on their tweets, using W2Vec vectorization and a logistic regression model, obtaining an Accuracy of 57.14%. In our research, we experienced a significant 22% improvement over Accuracy.

These studies observe considerable improvement, attributable to the adoption of a rigorous methodology for annotation, data quality assessment, preprocessing, and manual review of the data. Including stopwords derived from the annotation process also contributed to improving the quality of the data set.

Another relevant aspect is the comparison with the work [17], which used SMOTE data augmentation techniques. In contrast, in our research, we apply Data Augmentation methods. Although our results were superior in the scenario without data augmentation, this superiority was reflected only in the validation sets. On the other hand, notable improvements were observed in the test and training sets when applying this technique.

Furthermore, optimizing hyperparameters in our models also played a fundamental role in improving the results, allowing a tighter adaptation to the inherent complexity of the data and thus improving the predictive capacity of the models.

Regarding the limitations of this study, it is important to note that the availability of data was a significant challenge due to the specific nature of the phenomenon studied in the Ayacucho region, which encompasses sensitivities, sociopolitical context and the complexity of the social construction around to the constituent process. Furthermore, in most cases, a significant imbalance was observed between classes, with a predominance of non-constituent texts compared to constituent texts.

The ambiguity inherent in the content of tweets is another limitation to consider. The subjective nature of the texts,

The size and informality of the tweets in some cases made interpreting the textual content difficult, evidencing that this task can be challenging depending on the study phenomenon being addressed.

Applying this study phenomenon can inspire researchers to explore the generalization of approaches and methodologies to other regions or study phenomena in the sociopolitical field given its sensitivity and complexity. Furthermore, diversifying data sets and developing advanced approaches to address ambiguity in text classification are promising areas for future research in this field.

Table 12. Comparison of binary classification models based on other research works.

C	*Research Study*	*Best Model*	*Featu res*	*Metrics*
1	Ozcan, S. ([17])	SVM	BERT	Acc = 77.3%, F1-Score = 56.4%
2	Onikoyi, B ([19])	LR	W2V ec	Acc = 57.14%

3	Our Study	SVM	BOW	Acc = 73% y F1 = 71%

7. Conclusions, Limitations, and Futures Works

In conclusion, this study focused on developing classification models to improve the constituent process. He faced three crucial challenges in his analysis and successfully overcame them. A robust data set containing texts discussing political and social aspects relevant to improving the constituent process was built. Additionally, an annotation process based on literature reviews was implemented, allowing the creation of an annotation guide to label constituent and non-constituent classes. Finally, the proposed model achieved the implementation of classification models, evidencing a significant improvement in performance and precision compared to other studies carried out in this context. These results are fundamental to understanding and improving the constituent process as a whole.

However, it is essential to point out some limitations in our study. The limited availability of data, due to the specificity of the phenomenon studied in the Ayacucho region, restricted the amount of labeled data available.

Future research suggests exploring the use of Large-Scale Language Models (LLMs) tools to improve the classification and understanding of texts related to the constituent process. Furthermore, it would be valuable to replicate the study in different geographic contexts, such as a city on the north coast in contrast to a region in the southern highlands of Peru, to examine possible variations in the discussion and perception of the constituent process in different areas of the country.

Author contributions

Yordan Sullca-Palomino: Data curation, Writing-Original draft preparation, Software, Validation, Research. **Yudi Guzmán-Monteza:** Conceptualization, Methodology, Research, Writing-Reviewing and Editing.

References

[1] E. J. Zechmeister y N. Lupu, “El Barómetro de las Américas 2018/19”. 2019.

[2] S. L. Shaw, M. H. Tsou, y X. Ye, “Editorial: human dynamics in the mobile and big data era”, International Journal of Geographical Information Science, vol. 30, núm. 9, pp. 1687–1693, sep. 2016, doi: 10.1080/13658816.2016.1164317.

[3] “Tolerancia a los ‘golpes de Estado’ ejecutivos en Perú - Red de Desarrollo Social de América Latina y el Caribe (ReDeSoc).” Accessed: Mar. 11, 2024. [Online]. Available: https://dds.cepal.org/redesoc/portal/publicaciones/ficha/?id=5081

[4] ONPE, “Pasos para llegar al referéndum nacional 2018,” 2018.

[5] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor, “The power of prediction with social media,” Internet Research, vol. 23, no. 5, pp. 528–543, Oct. 2013, doi: 10.1108/IntR-06-2013-0115.

[6] A. Jungherr, “Twitter Use in Election Campaigns: A Systematic Literature Review,” Journal of Information Technology & Politics, vol. 13, pp. 72–91, Mar. 2016, doi: 10.1080/19331681.2015.1132401.

[7] R. E. Proaño Arias, “Aplicación de la minería de datos para análisis de tendencias políticas en redes sociales,” masterThesis, Quito, 2019. Accessed: Mar. 11, 2024. [Online]. Available: http://repositorio.uisrael.edu.ec/handle/47000/2023

[8] E. León Pluas, E. Proaño Arias, V. Muirragui Irrazábal, y J. Cajamarca Yunga, «Minería de datos en el análisis de tendencias políticas en redes sociales», CD, vol. 3, n.º 3.4., pp. 91-103, sep. 2019.

[9] N. Roales González, “Detección de tendencias en twitter utilizando minería de datos adaptativa,” 2014.

[10] [1] D. V. Calvo, M. A. A. Pardo, and C. G. Rodrıguez, “Análisis de contenidos en Twitter: clasificación de mensajes e identificación de la tendencia política de los usuarios,” Jun. 2014.

[11] M. E. Gordon Pico, “Desarrollo de una herramienta de minería de datos para el análisis de influencia de cuentas automatizadas en temas de tendencia sobre la opinión de los usuarios de twitter en Ecuador,” Jun. 2018.

[12] Z. Zhang, X. Lin, and S. Shan, “Big data-assisted urban governance: An intelligent real-time monitoring and early warning system for public opinion in government hotline,” Future Generation Computer Systems, vol. 144, pp. 90–104, Jul. 2023, doi: 10.1016/j.future.2023.03.004.

[13] J. Urpay-Camasi, J. Garcia-Calderon, and P. Shiguihara, “A Method to Construct Guidelines for Spanish Comments Annotation for Sentiment Analysis,” in 2021 IEEE Sciences and Humanities International Research Conference (SHIRCON), 2021, pp. 1–4. doi: 10.1109/SHIRCON53068.2021.9652313.

[14] Y. Guzmán-Monteza, “Assessment of an annotation method for the detection of Spanish argumentative, non-argumentative, and their components,” Telematics and Informatics Reports, vol. 11, p. 100068, 2023, doi: https://doi.org/10.1016/j.teler.2023.100068.

[15] M. S. Raja and L. A. Raj, “Fake news detection on social networks using Machine learning techniques,” Materials Today: Proceedings, vol. 62, pp. 4821–4827, 2022, doi: https://doi.org/10.1016/j.matpr.2022.03.351.

[16] L. Liu, A. Guevara, and J. E. Sanchez-Galan, “Identification and classification of road traffic incidents in Panama City through the analysis of a social media stream and machine learning,” Intelligent Systems with Applications, vol. 16, p. 200158, 2022, doi: https://doi.org/10.1016/j.iswa.2022.200158.

[17] S. Ozcan, M. Suloglu, C. O. Sakar, and S. Chatufale, “Social media mining for ideation: Identification of sustainable solutions and opinions,” Technovation, vol. 107, p. 102322, Sep. 2021, doi: 10.1016/j.technovation.2021.102322.

[18] Y. Guzman, A. Tavara, R. Zevallos, and H. Vega, “Implementation of a Bilingual Participative Argumentation Web Platform for collection of Spanish Text and Quechua Speech,” in 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2021, pp. 1–6. doi: 10.1109/ICECCE52056.2021.9514251.

[19] B. Onikoyi, N. Nnamoko, and I. Korkontzelos, “Gender prediction with descriptive textual data using a Machine Learning approach,” Natural Language Processing Journal, vol. 4, p. 100018, 2023, doi: https://doi.org/10.1016/j.nlp.2023.100018.

[20] M. Cardaioli, P. Kaliyar, P. Capuozzo, M. Conti, G. Sartori, and M. Monaro, “Predicting Twitter Users’ Political Orientation: An Application to the Italian Political Scenario,” in 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 159–165. doi: 10.1109/ASONAM49781.2020.9381470.

[21] A. Alshehri, W. Isaacs, A. Addawood, M. Trotz, and S. Chellappan, “Predicting Community Engagement on Twitter on Environmental Health Hazards,” in 2019 IEEE 13th International Conference on Semantic Computing (ICSC), Los Alamitos, CA, USA: IEEE Computer Society, Feb. 2019, pp. 450–455. doi: 10.1109/ICOSC.2019.8665530.

[22] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, “Short text classification for Arabic social media tweets,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 6595–6604, 2022, doi: https://doi.org/10.1016/j.jksuci.2022.03.020.

[23] K. A. Qureshi and M. Sabih, “Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text,” IEEE Access, vol. 9, pp. 109465–109477, 2021, doi: 10.1109/ACCESS.2021.3101977.

[24] N. Reynaldo, Goenawan, W. Chanrico, D. Suhartono, and F. Purnomo, “Gender Demography Classification on Instagram based on User’s Comments Section,” Procedia Computer Science, vol. 157, pp. 64–71, 2019, doi: https://doi.org/10.1016/j.procs.2019.08.142.

[25] L. Li, Z. Ma, H. Lee, and S. Lee, “Can social media data be used to evaluate the risk of human interactions during the COVID-19 pandemic?,” International Journal of Disaster Risk Reduction, vol. 56, p. 102142, 2021, doi: https://doi.org/10.1016/j.ijdrr.2021.102142.