UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

A Zoom into Ecuadorian Politics: Manifesto Text Classification using NLP

Fernanda Emilia Barzallo Burbano
María Emilia Moscoso Montalvo
Margorie Fernanda Pérez Simba

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 20 de diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

A Zoom into Ecuadorian Politics: Manifesto Text Classification using NLP

Fernanda Emilia Barzallo Burbano

María Emilia Moscoso Montalvo

Margorie Fernanda Pérez Simba

Nombre del profesor, Título académico **María Gabriela Baldeón Calisto, Ph.D.**

Quito, 20 de diciembre de 2022

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Fernanda Emilia Barzallo Burbano

Código: 00205499

Cédula de identidad: 172737099

Nombres y apellidos: María Emilia Moscoso Montalvo

Código: 00205672

Cédula de identidad: 1719805762

Nombres y apellidos: Margorie Fernanda Pérez Simba

Código: 00205247

Cédula de identidad: 1725457558

Lugar y fecha: Quito, 20 de diciembre de 2022

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La investigación en ciencias políticas sobre programas de gobierno es a menudo requerida para comprender las estrategias y acciones propuestas durante las campañas electorales. Para lograr esto, los politólogos clasifican las oraciones y cuasi oraciones de los programas de gobierno en siete dominios establecidos por el Comparative Manifesto Project. Sin embargo, esta tarea requiere de mucho tiempo, mano de obra, y puede dar lugar a sesgos. En Ecuador, todos los partidos candidatos deben presentar sus programas de gobierno, pero los análisis políticos cuantitativos en este país aún son incipientes. Por lo tanto, en este trabajo se desarrolla una forma eficiente y objetiva de analizar los programas de gobierno ecuatorianos mediante el etiquetado automático de sus declaraciones utilizando redes de transformadores. Primero, se lleva a cabo un diseño experimental para determinar qué modelo de Transformador, tipo de preprocesamiento y datos de entrenamiento se deben usar para aumentar la exactitud del modelo de clasificación. Los resultados mostraron que los modelos de Transformadores DistilBERT y RoBERTa funcionaron consistentemente bien. Sin embargo, DistilBERT superó estadísticamente a RoBERTa. Además, no se necesita procesamiento previo, pero se puede aplicarlo para mejorar la eficiencia computacional. Finalmente, los programas de gobierno de México y Argentina aumentan la exactitud de la clasificación cuando se utilizan como datos de entrenamiento. Adicionalmente, se desarrolló un segundo diseño experimental para comprender el efecto del conjunto de datos de entrenamiento. Este experimento demostró que hay un gran aumento en la exactitud cuando los conjuntos de datos de entrenamiento y validación comparten las mismas fuentes.

Palabras clave: Comparative Manifesto Project, ciencias políticas, Procesamiento Natural del Lenguaje, Transformador, DistilBERT, RoBERTa, preprocesamiento, exactitud.

ABSTRACT

Political science research on party manifestos is often required to understand their strategies and proposed actions during electoral campaigns. To achieve this, political scientists classify the manifestos sentences and quasi-sentences into seven main domains established by the Comparative Manifesto Project. However, this task is time-consuming, labor-intensive and can lead to biases. In Ecuador all candidate parties must submit their manifestos, but quantitative political analyses in this country are still incipient. Therefore, an efficient and objective way of analyzing Ecuadorian manifestos by automatically labeling its statements using Transformer networks is developed in this work. First, an experimental design is conducted to determine which Transformer model, type of pre-processing and training data should be used to increase the accuracy of the classification model. The results showed that the DistilBERT and RoBERTa Transformer models performed consistently well. However, DistilBERT statistically outperformed RoBERTa. In addition, no pre-processing is needed, but it can be applied to improve computational efficiency. Finally, Mexico's and Argentina´s manifestos increase the classification accuracy when they are used as training data. Furthermore, a second experimental design was also developed to understand the training dataset effect. This experiment demonstrated that there is a high accuracy increase when the training and validation datasets share the same sources.

Key words: Comparative Manifesto Project, political sciences, Natural Language Processing, Transformer, DistilBERT, RoBERTa, preprocessing, accuracy.

TABLE OF CONTENTS

TABLE INDEX

FIGURES INDEX

INTRODUCTION

Natural Language Processing (hereinafter NLP) is a branch of Artificial Intelligence that enables computers to understand spoken and written texts. Although it cannot reach the same level of understanding as humans, state-of-the-art deep learning models have achieved considerable success in several tasks such as sentiment analysis, text summarization, automated question and answering, translation, and topic classification (Khurana et al., 2022). This last application becomes particularly useful when analyzing political documents as it allows to automatically categorize unstructured text into different topics and political positions, which is the first step for more detailed political analysis (Terechshenko et al., 2020).

Social scientists' study political manifestos during electoral campaigns since they are the authoritative source of a party´s policy. Manifestos contain the vision, mission, ideology, main objectives, and strategies of a party (Volkens et al., 2021). Moreover, they can be used to examine whether promises were fulfilled, compare competitor parties, and comprehend the basis for government coalition programs (Suiter & Farrell, 2011). However, topic classification of political manifestos is time-consuming and labor-intensive (Zirn et al., 2016). Due to a large number of categories, manual classification is prone to error (Bilbao-Jayo & Almeida, 2018a). Furthermore, it can lead to biases since political experts only agree 50% of the time (Rasov et al., 2020). Therefore, automated tools that reduce the aforementioned problems have been explored, being deep learning models the most successful (Glavaš et al., 2017; Wiedemann, 2018; Bilbao-Jayo & Almeida, 2018a).

In Ecuador, political manifestos are better known as government programs. The Ecuadorian Constitution and Electoral Law requires all parties to have manifestos to participate in any electoral process. Moreover, these documents must contain the actions to be executed

by the political party if the candidate is elected (Constitución de la República de Ecuador [Const], 2008). In the last 16 years, political science in Ecuador has positioned itself as an area of research. Pinta et al. (2021) explored a way to compare government plans of the two 2021 Ecuadorian Presidential Elections candidates through NLP techniques. Their study consisted of the application of two neural network models, Distributed Memory Model (DM) and Distributed Bag of Words (DBOW). The results showed that the DM model is better for comparing specific topics in small documents, while the BDOW model provides more reliable results in large documents.

Another relevant article in this area described the development of an application architecture to measure the closeness of a candidates' campaign on Twitter and their electoral manifesto. The results showed that none of the candidates used a similar language or addressed related topics in their manifestos and online campaign (Riofrío et al., 2021). In the context of social media analysis, Cumbicus-Pineda et al. (2018) focused on tweet sentiment analysis for Ecuadorian political figures and organizations. The authors applied six classifiers, Naive Bayes, Logistic Regression, K-Nearest Neighbors, Random Forest, and Sequential Minimal Optimization (SMO). The latter model achieved the highest performance. Although these works had promising results, political studies in Ecuador are still incipient and there is a large field to explore (Zamora, 2021).

In this paper, an efficient and objective way of analyzing Ecuadorian manifestos by automatically labeling its statements using Transformer networks is developed. The main objective of this work is to classify sentences and quasi-sentences from government programs in Ecuador, through state-of-the-art Natural Language Processing techniques to speed up and avoid biases in future political analysis. To achieve this, Montgomery's guidelines for designing an experiment and basic data analytics process were used to statistically determine

which models, types of preprocessing and types of training data improve the classification accuracy. This study aims to answer the following questions:

1. Which Transformer model (RoBERTa or DistilBERT) increases the classification accuracy?
2. Which type of preprocessing increases the classification accuracy?
3. Which government programs from Spanish-speaking countries used as training data increase the accuracy of the classification model?

DEVELOPMENT OF THE TOPIC

1. Background and Related Work

This section shows how NLP has been utilized for analysis in social and political texts and then focuses on the analysis of political manifestos.

1.1 Analysis of social and political texts

The availability of digital documents in recent years has raised the interest in quantitative text analysis, especially within the social sciences. Hallac et al. (2018) compared the performance of Convolutional Neural Networks, standard Multilayer Perceptron, and hybrid Bi-LSTM-CONV model to classify tweets into 5 categories. The Bi-LSTM-CONV model had the best accuracy, particularly after a series of weight tuning steps. Charalampakis et al. (2016) proposed a classification schema to find a relation between ironic tweets that refer to political parties in Greece and their actual election results. The authors explored a semi-supervised learning technique that allows the classification of partially labeled data, called collective classification. Several algorithms were tested, and the highest precision was achieved with Random Forest. Likewise, Kent & Krumbiegel (2021) classified socio-political events using the 25 subtypes of the Armed Conflict Location & Event Data Project database. The best model according to the weighted F1-score was RoBERTa. The text used for this model was not preprocessed since the preliminary experiments of the study showed that the removal of locations and time data does not influence the system. In fact, the study indicated that the RoBERTa embeddings benefit from the inclusion of this type of detailed information.

As proposed by Büyüköz et al. (2020), NLP can also be used to classify news articles. The authors developed a DistilBERT model that determined whether a document referred to a

political protest or not. The study evaluated cross-context performance by testing the model on news coming from a different country than the training data. Results showed that DistilBERT is better at using longer sequences than ELMo (Peters et al., 2018) and it generalizes better in the cross-context setting. Terechshenko et al. (2020) studied transfer learning by training a RoBERTa model with U.S. congressional bills and testing it on New York Time headlines. The findings suggested that RoBERTa is especially useful when few data labels are available. However, using training data for the actual task offers substantial improvements.

Natural language processing tools have also been adopted in political science to avoid biases when analyzing official texts. For example, Öztürk & Özcan (2022) recognized political sentences as liberal, conservative, and neutral from the Ideological Books Corpus, a dataset that used United States electoral speeches. The study did a benchmark of Transformer-based models and Machine Learning methods, from which ELECTRA (Clark et al., 2020) resulted in the best one in terms of the F1-score. Additionally, Lehmann & Zobel (2018) applied crowd coding to election programs and introduced a new set of comparative data on the immigration positions and salience of political parties from different countries. Schoonvelde et al. (2019) also carried out automated text analysis applications that measure topics, ideology, sentiments and personality within political science and political psychology in the European Union. They mainly used scaling models such as Wordscore and Wordfish from an EUSpeech database. Table 1 summarizes the studies mentioned above.

Table 1. Summary of related work for social and political text analysis

Author	Model	Dataset	Countries	Evaluation metric
Hallac et al. (2018)	Bi-LSTM-CONV	Bigailab-5news500K (news data) Bigailab-5tweet-35K (Twitter data)	-	Accuracy: 0.88
Charalampakis et al. (2016)	Random forest with collective classification	Greek Political Tweets	Greece	Precision: 0.831

Kent & Krumbiegel (2021)	RoBERTa	Armed Conflict Location & Event Data Project	-	Weighted F-score: 0.83
Büyüköz et al. (2020)	DistilBERT	A proposed setting within the Lab Protest News of the Conference and Labs of the Evaluation Forum (CLEF)	India and China	F-score: 0.768
Terechshenko et al. (2020)	RoBERTa	Congressional bills from the Comparative Agendas Project and New York Times headlines	United States	Accuracy: 0.6
Öztürk & Özcan (2022)	ELECTRA	Ideological Books Corpus	United Sates	F-score: 0.7019 Accuracy: 0.7024
Lehmann & Zobel (2018)	Crowd coding method	Time-series indexed and cross-sectional dataset	United States, Switzerland, Sweden, Spain, Norway, New Zealand, Netherlands, Ireland, Germany, Finland, Denmark, Canada, Austria, Australia	Alpha score: 0.667
Schoonvelde et al. (2019)	Scaling Models (Wordscore and Wordfish)	EUSpeech	EU countries	-

1.2 Political Manifestos domain classification

Recent studies have already identified the benefits of using Machine Learning and Artificial Neural Networks to classify political topics in electoral manifestos. Hand labeling political text can be expensive and prone to human errors. Thus, Koh et al. (2021) proposed a BERT-CNN model that classifies political manifestos from seven different countries into 7 policy domains and 57 policy preferences. The model manages reproducibility and scalability issues. Nonetheless, the authors concluded that the model requires fine-tuning to improve the fine-grained policy positions classification. Zirn et al. (2016) applied three independent sentence-level classifiers to predict the domains of political manifestos and detect domain shifts between adjacent sentences. As a second step, the authors combined the predictions of the three classifiers with a global Markov Logic-based optimization setting. Based on experimental results it was demonstrated that the global model outperforms the sentence-level topic classifiers. Glavaš et al. (2017) proposed a CNN classification model that can be applied to

cross-lingual electoral manifestos by generating joint multilingual semantic vector spaces. To achieve this, one language was set as a target embedding space and the vectors of words from other languages were translated by Mikolv et al. (2013) model. Their results showed that classifiers trained on multilingual data outperform monolingual topic classification.

Likewise, another study by Rasov et al. (2020) explored Machine Learning models such as support vector machine and gradient boosting and with different kernel sizes. Moreover, they proposed a new approach to overcome the problem of words falling outside the training vocabulary and concluded that a longer textual context is helpful for increasing the classification accuracy. Wiedemann (2018) evaluated the performance of a regression-based approach and an aggregating classifier method. In both cases, the uneven distribution of characteristic language structures impacted negatively the performance. Therefore, active learning was used to improve this issue. This approach alternates between Machine Learning and human coding, which led to a higher accuracy for proportional classification. A summary of the studies reviewed is shown in Table 2.

Table 2. Related work summary for political manifesto domain classification

Author	Model	Dataset	Countries	Evaluation metric
Kohn et al. (2021)	BERT-CNN	Manifesto Project Corpus	United States, Canada, Great Britain, Ireland, New Zealand, Australia and South Africa	F1-score: 0.591 Accuracy: 0.591
Zirn et al. (2016)	SVM model with bag-of-words-term-vector of the first and second sentence	Manifesto Project Corpus	United States	F1-score: 0.793
Glavaš et al. (2017)	CNN	Manifesto Project Corpus	-	F1-score: 0.86
Rasov et al. (2020)	Automatic Coding Algorithm	Manifesto Project Corpus	-	Accuracy English: 0.485 German: 0.436 Spanish: 0.461
Wiedemann (2018)	Active learning	Manifesto Project Corpus	Australia, Ireland, United Kingdom and United States	F1-score: 0.86

1.3 Political Manifestos left-right wing classification

Manifestos have also been used to analyze a party's position on the left-right political wing and to determine the party's position on different issues. Subramanian et al. (2017) proposed a joint model that does sentence-level thematic classification and document-level position quantification in manifestos from 13 countries. The model used multilingual embeddings and an Adam optimizer to conclude that the joint sentence-document model proposed perform better for document-level regression. In another study, Subramanian et al. (2018) classified manifestos in policy issue classes and scored them based on a policy-based left–right spectrum. They used a Bi-LSTM that consists of a two-level structured model, which captures information of the temporal dependencies and context within manifestos. NLP tools are also useful to test theoretical assumptions related to political parties. Such is the case of Bielik (2020), who did a keyword and sentiment analysis to interpret the content and way of communication of four political parties in the Visegrad Group. NLP demonstrated to be an efficient method for analyzing textual data; still, limitations were encountered since the manifestos were translated into English and only 17 documents were available. Table 3 summarizes the results of the three studies mentioned above.

Table 3. Related work summary for left-right wing classification

Author	Model	Dataset	Countries	Evaluation metric
Subramanian et al. (2017)	Hierarchical neural network	Manifesto Project Corpus	Austria, Australia, Denmark, Finland, France, Germany, Italy, Ireland, New Zealand, South Africa, Switzerland, United Kingdom and United States	MSE: 0.044
Subramanian et al. (2018)	Bi-LSTM PSL Model	Manifesto Project Corpus	12 European countries	F-score: 0.48
Bielik (2020)	Rapid Automatic Keyword Extraction, noun phrase analysis and frequency distribution	Manifesto Project Corpus	Czech Republic, Hungary, Poland and Slovakia	-

1.4 Political Manifestos as training data for other texts classification

Furthermore, using political manifestos as the training data for classifying other types of texts has become more frequent. Chatsiou (2020) constructed a sentence-level political discourse classifier showing that CNN classifiers combined with BERT Transformer outperform models with other embeddings like Word2Vec, Glove, or ELMo. Moreover, Abercrombie et al. (2019) have developed a way to automatically tag debate motions with codes from a pre-existing coding scheme for analysis of political party manifestos from the Manifesto Project Corpus using a series of classification models and resulting in the combination of BERT+CNN as the best of them. Other studies also highlight the importance of considering the context of the words or phrases in their classification. For instance, Bilbao-Jayo & Almeida (2018b) developed an automated analysis of political speeches in different languages based on multi-scale CNN enhanced with context information and have statistically confirmed that adding the previous sentence as an additional channel or in another structure of CNN improves the performance of the classifier. Similarly, Bilbao-Jayo & Almeida (2018a) in another of their articles presented a model that takes advantage of the context to classify political discourse in social media based on a convolutional neural network architecture, demonstrating its usefulness by analyzing the activity on Twitter of the main political parties during the Spanish general elections of 2015 and 2016. Table 4 summarizes the studies that have used political manifestos for training models that classify other types of political texts.

Table 4. Related work summary for political manifestos used as training data

Author	Model	Dataset	Countries	Evaluation metric
(Chatsiou, 2020)	BERT+CNN	Manifesto Project Corpus and Coronavirus (COVID-19) Press Briefings Corpus	UK (England, Scotland, Wales, Northern Ireland) and the World Health Organization (WHO)	F-score: 0.6458 Accuracy: 0.6865
(Bilbao-Jayo & Almeida, 2018b)	Multi-scale CNN with Word2Vec word embeddings	Manifesto Project Corpus	-	RILE Scale
(Bilbao-Jayo & Almeida, 2018a)	BERT+CNN	Manifesto Project Corpus	Spain	F-score: 0.7529 Accuracy: 0.8763

2. Methodology

The methodology applied in this study is comprised of 4 steps and based on Montgomery's guidelines for designing an experiment (Montgomery, 2013) and the basic data analytics process (Mehrishi, 2019). The steps are shown in Figure 1. First, the experimental design model is defined. For this, the statement of the problem must be set. Next, the response variable, the factors and levels, and the experimental design are selected. The second step is data collection, where the database is acquired, cleaned, and preprocessed. The third step is the execution of the experiments, in which the Transformer architectures are trained with different combinations of preprocessing operations and training datasets. Finally, in step four the results of the statistical analysis are studied, and the conclusions and recommendations drawn. In the next subsections, each of these steps are described in detail.

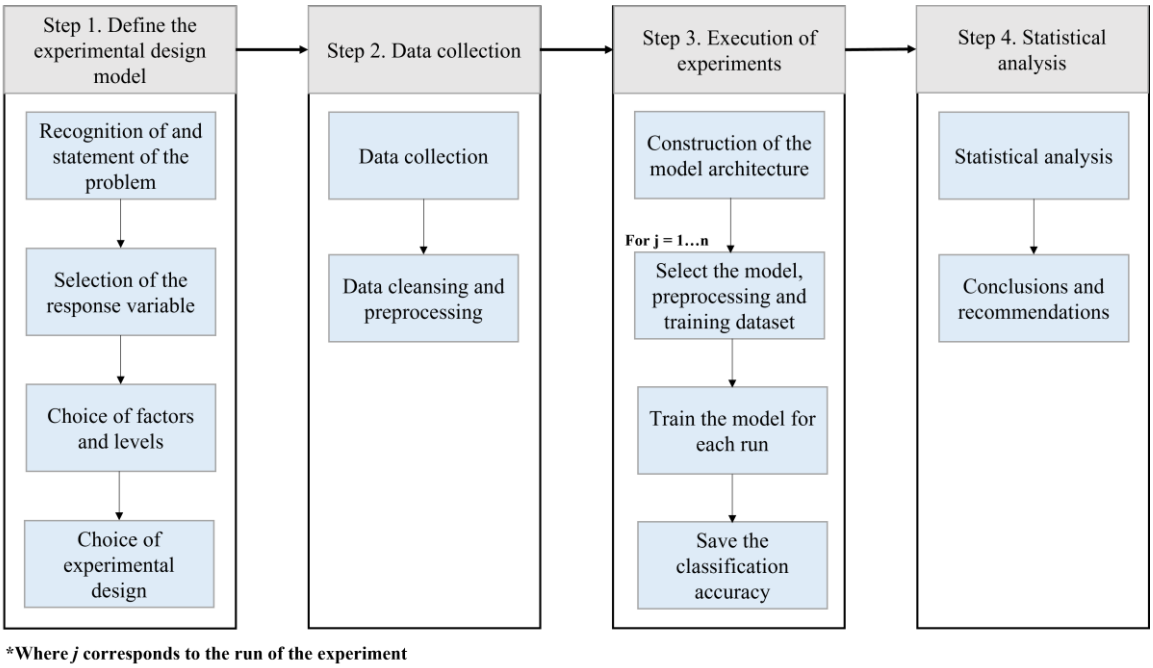

Figure 1. Methodology applied in the current study

2.1 Step 1. Define the experimental design model

2.1.1 Recognition and statement of the problem

As mentioned above, manually classifying sentences and quasi-sentences from manifestos is time-consuming, labor-intensive, and can lead to bias as political pundits only agree 50% of the time (Rasov et al., 2020). Therefore, automated tools have been used in this research to solve these problems. The objective then results in determining which combination of preprocessing operations, deep learning models and different training databases generate the highest accuracy in the classification of political text of the electoral programs of Ecuador. Thus, a design of experiments will be used so that the results are statistically significant.

2.1.2 Selection of the response variable

The response variable selected is the classification accuracy of topic labelling in Ecuadorian manifestos on the validation set. This metric has been chosen since it compares the model´s classification quality against human annotation. Specifically, the accuracy measures the ratio between the number of correct predictions and the total number of predictions performed by the model, as presented in equation 1 (Öztürk & Özcan, 2022).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

2.1.3 Choice of factors and levels

In this study, the statistical effect of 10 training factors on the response variable was analyzed. One factor defines the type of Transformer model to be trained, three factors consider the preprocessing operations that can be done to the training datasets, and 6 factors determine

the corpuses to be used for training. In Table 5, each of the 10 factors considered as well as the low and high level values tested are presented.

Table 5. Factors for experimental design

Factor	High level	Low level
Model	RoBERTa	DistilBERT
Stop Words	Yes	No
Stemming	Yes	No
Lemmatization	Yes	No
Argentina	Yes	No
Bolivia	Yes	No
Chile	Yes	No
Mexico	Yes	No
Uruguay	Yes	No
Spain	Yes	No

For the model type factor, the Transformer models considered are DistilBERT and RoBERTa (Sanh et al., 2020; Liu et al., 2019). These two networks are selected given their excellent performance in manifesto classification and cross-context training (Büyüköz et al., 2020; Kent & Krumbiegel, 2021; Terechshenko et al., 2020). Both networks are based on the Transformer architecture presented by Vaswani et al. (2017) and particularly based on the architecture of the BERT model (Devlin et al., 2019). The vanilla Transformer architecture has an encoder-decoder structure as shown in Figure 2. The encoder consists of six encoder layers, each one with two sub-layers. The first sub-layer applies a multiheaded self-attention mechanism that encodes the context of a given word in its vector to provide a better understanding of the semantic information of each sentence. The second sub-layer is a feed-forward neural network that further processes each output encoding individually (Terechshenko et al., 2020). The decoder, on the other hand, has 6 decoder layers. Each decoder layer has three major components that are the self-attention mechanism, attention mechanism over the encodings, and a feed-forward neural network (Devlin et al., 2019).

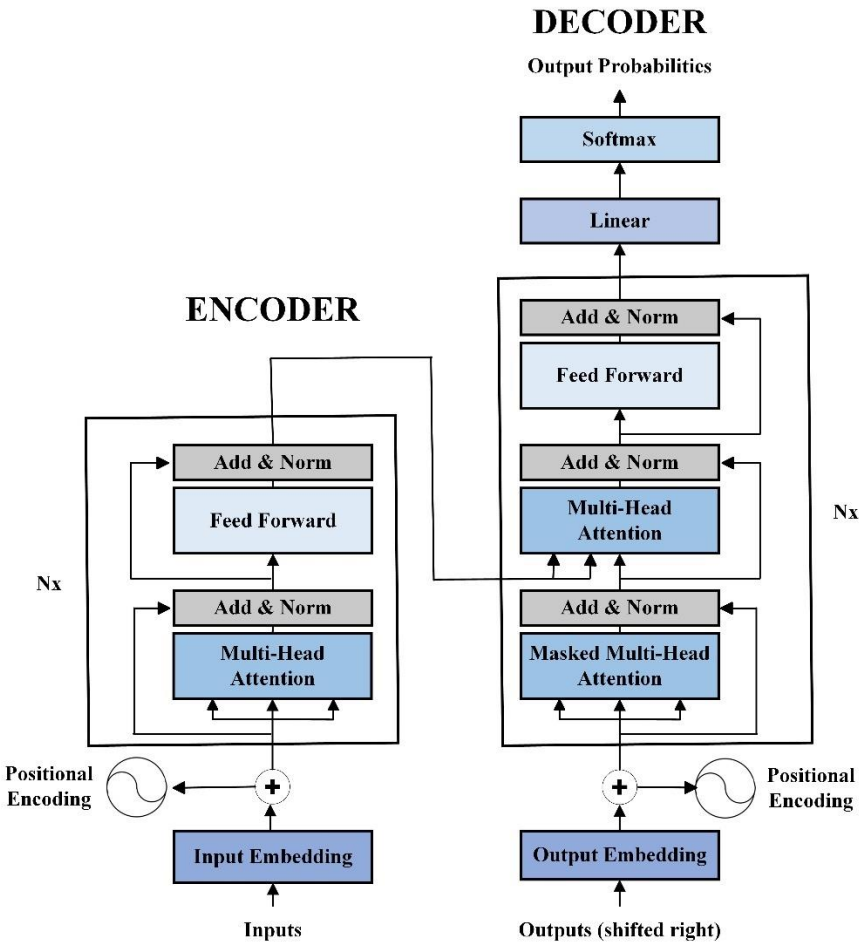

Figure 2. Transformer Language Model Architecture (Vaswani et al., 2017)

As previously mentioned, both models that are applied in the present study are derived from the BERT model developed by Devlin et al. (2019). BERT stands for Bidirectional Encoder Representations of Transformers and consists of the use of Next Sentence Prediction and Masked Language Model (MLM). The latter model's objective is to predict the original vocabulary ID, solely based on its context (left and right), of the random masked tokens from the input. Therefore, with one additional layer the pre-trained BERT model can be finetuned to develop models for eleven natural language processing tasks, one of which is text classification. BERT can be considered the first finetuning based representation model that accomplishes a state-of-the-art performance on a large range of tasks (Devlin et al., 2019). Additionally, BERT's training process involves two stages: pre-training on unlabeled data and training on labeled data (Koroteev, 2021).

Likewise, RoBERTa stands for Robustly Optimized BERT Approach. This model has the same general architecture as BERT, but it differs by its training procedure (Kent & Krumbiegel, 2021). RoBERTa introduces dynamic masking to avoid the same masked tokens in every epoch, which becomes particularly useful when pretraining with larger datasets. In addition, this model eliminates Next Sentence Prediction and uses a larger byte-level Byte-Pair Encoder. Finally, RoBERTa improves end task performance and becomes more efficient by using larger mini batches (Liu et al., 2019). On the other hand, DistilBERT is a small and fast Transformer model trained by distilling the BERT base (Büyüköz et al., 2020). All token-type embeddings and the pooler are removed from the architecture, while the number of layers is reduced by a factor of 2. In this way, the size of the model is reduced by 40% making it 60% faster (Sanh et al., 2020). It is pretrained using dynamic masked language and removing the Next Sentence Prediction objective. Additionally, a large number of operations used in this Transformer architecture are highly optimized, so that variations in the last dimension of the tensor have a lower impact on calculation efficiency (Sanh et al., 2020).

The following three factors taken into account are the type of preprocessing operations to be applied. The operations considered are stop word removal, stemming, and lemmatization. Stop word removal is the elimination of highly frequent words that do not provide additional information to the text, such as prepositions, articles, and conjuctions (Etaiwi & Nayma, 2017). Stemming refers to the process of removing affixes (prefixes and sufixes) from words. Therefore, reducing the word to its root based on grammatical rules (Kadhim, 2018). Lastly, lemmatization is the process of transforming the word to its base, called lemma, by removing or replacing its suffix (Tabassum & Patil, 2020). These operations have been selected as they perform textual feature reduction that has shown to increase the accuracy is some models (Orellana et al., 2018). These factors have levels of “Yes” and “No” that represent whether or not they will be implemented before training.

The remaining six factors are the Spanish speaking countries whose corpuses are included in the training dataset. The countries considered are Argentina, Bolivia, Chile, Mexico, Uruguay and Spain, which are all the countries that have manifestos in Spanish in The Comparative Manifesto Project. Each factor has the levels of "Yes" and "No" that represent whether they are part or not of the training dataset.

2.1.4 Choice of experimental design

Design of experiments allows to study the impact of different factors on a response variable (Montgomery, 2013). In the present study, 10 factors with 2 levels each are considered. This results in a 2^{10} factorial design in which all possible combinations of levels across all factors are tested. Since this experiment would require 1,024 runs, which is computationally expensive and time consuming, a fractional factorial design was chosen. Therefore, a resolution IV $2^{(10-5)}$ fractional factorial experiment was selected, which minimizes the number of runs to 32. This 32 combinations where obtained using the Design-Expert V.13 statistical software (State-Ease, 2022). This defines the random combinations indicating the order in which they must be executed to guarantee the independence of the data. In this way, the value of one observation will not influence or affect the value of other observations.

2.2 Step 2. Data collection

2.2.1 Data collection

This work uses two corpuses, one from the Comparative Manifesto Project to train the Transformer networks, and the other from the 2021 Ecuadorian presidential election for testing the model. The description of each dataset is provided next.

2.2.1.1 Comparative Manifesto Project Dataset

The Comparative Manifesto Project is a collection of party election documents from 50 countries around the world. Currently the database covers electoral programs from 1,000 parties in 40 languages dating back to 1,945. The documents have been manually annotated by political experts into statements or quasi-sentences and labelled into seven domains: External Relations, Freedom and Democracy, Political System, Economy, Welfare and Quality of Life, Fabric of Society and Social Groups. Moreover, there is an additional category that represents all statements that do not belong to any of the mentioned above (Volkens et al., 2021).

To obtain the most training sentences and quasi-sentences as possible, this study uses 126 manifestos from all native Spanish speaking countries that had their documents as digital texts in the Comparative Manifesto Project database. The oldest document dates from 1989 and the most recent one is from 2019. The training set is composed of 224,407 sentences and quasi-sentences. For the training set, all observations from the Comparative Manifesto Project country database were used without performing any partitioning. This is to increase the vocabulary to be learned by the network models and improve the accuracy of the classification (Bielik, 2020). Its distribution in the different countries and domains is shown in Table 6.

Table 6. Sentences and quasi-sentences from the Comparative Manifesto Project Dataset
Values in parentheses in the total columns represent the percentage of observations pertaining to each category

Country	Welfare and Quality of Life	Economy	Political System	Social Groups	Fabric of Society	Freedom Democracy	External Relations	Total
Spain	28,971	27,397	13,624	8,933	5,909	7,144	5,967	97,945 (43.6%)
Mexico	14,687	9,125	6,601	4,844	4,723	4,063	2,973	47,016 (21.0%)
Chile	13,582	9,671	4,112	3,115	3,877	3,423	1,865	39,645 (17.7%)
Argentina	3,632	3,967	3,109	1,412	1,965	1,875	1,175	17,135 (7.6%)
Uruguay	4,507	5,240	1,906	1,322	1,074	770	817	15,636 (7.0%)
Bolivia	1,351	2,458	1,110	483	733	528	367	7,030 (3.1%)
Total	66,730 (29.7%)	57,858 (25.8%)	30,462 (13.6%)	20,109 (9%)	18,281 (8.1%)	17,803 (7.9%)	13,164 (5.9%)	**224,407**

2.2.1.2 Ecuadorian Elections Dataset

The Ecuadorian corpus was obtained from the official sites of the 2021 Ecuadorian elections candidates. The candidates, Andrés Arauz and Guillermo Lasso, have different political ideologies. Hence, their political programs offer a higher variety of vocabulary for analysis. Both documents were pre-processed by removing all images and tables. Then, the documents were divided into paragraphs and labelled by a political expert (Pinta et al., 2021). Because the training data used in the experiments are sentences and quasi-sentences, the paragraphs of the corpus were further broken down into sentences following the Manifesto Coding Instructions 5th edition (Werner et al., 2021). The final dataset has 1,809 sentences and quasi-sentences as presented in Table 7, where 80% is used for training and 20% for validation (Abercrombie et al., 2019; Hallac et al., 2018; Subramanian et al., 2018). Particularly, this dataset is divided into training and validation, to evaluate the effect cross-domain training has over inference. The validation set is used to evaluate the model in the fractional factorial design. Furthermore, during the division it was guaranteed that the ratio of observations between policy domains has a similar distribution between training and validation set (Kohn et al., 2021).

Table 7. Sentences and quasi-sentences from the Ecuadorian Election Dataset
Values in parentheses in the total columns represent the percentage of observations pertaining to each category

Country	Partition (80/20)	External Relations	Freedom Democracy	Political System	Economy	Welfare and Quality of Life	Fabric of Society	Social Groups	Total
Ecuador	Train	52	58	276	289	374	256	142	1,447 (80%)
	Valid	13	14	69	72	94	64	36	362 (20%)
Total		65 (3.6%)	72 (4.0%)	345 (19.1%)	361 (20.0%)	468 (25.9%)	320 (17.7%)	178 (9.8%)	**1,809**

2.2.2 Data cleansing and preprocessing

Data cleansing and basic preprocessing was performed to remove incorrect, missing, or corrupted information. This causes the imbalance of the data which affects its quality since it

may violate the principles of validity, accuracy, consistency, completeness, and uniformity (Mehrishi, 2019). As mentioned above, manifestos from Spanish-speaking countries were selected, therefore Spain was part of this group. Nevertheless, Spain's dataset included manifestos in Catalan that had to be removed from the training dataset. Other steps of data cleansing included the removal of numbers, punctuation, special characters, headings, and null values, also all letters were lowercased (Babanejad et al., 2020; Işık & Dağ, 2020; Tabassum & Patil, 2020). According to Babanejad et al. (2020), removing numbers can reduce noise but it does not affect classification accuracy. On the other hand, the removal of punctuation can increase the performance of the classification model because punctuation is treated as an additional dimension in the feature set for each word (Işık & Dağ, 2020). Furthermore, punctuation can be considered noise within a dataset since the machine is not able to understand it (Tabassum & Patil, 2020). Additionally, lowercasing is considered a simple and effective step since variation in capitalization produces different results (the same word in uppercase and lowercase is recognized by the computer as two different words, and therefore two different word vectors are created) (Tabassum & Patil, 2020).

2.3 Step 3. Execution of experiments

2.3.1 Construction of the model architecture

The Transformer models (RoBERTa and DistilBERT) are implemented using the SMaBERTa library from the open-source software provided by Terechshenko et al. (2020). It is worth mentioning that the model uses the AdamW optimizer (Loshchilov & Hutter, 2019), a train batch size of 25 samples and an evaluation batch size of 50 samples. The number of epochs used was selected after testing the training for 25, 50, 75 and 100 epochs. In these runs there was no great difference in the accuracy; therefore, a value of 25 epochs was selected as it does not require excessive computational time. Moreover, for the RoBERTa model the

learning rate was set to 3e-5, as this value achieved state-of-the-art results on the General Language Understanding Evaluation (GLUE) task dataset (Liu et al., 2019), and for DistilBERT was set 5e-5 as it was suggested in several studies focused on multiclass classification tasks (Büyüköz et al., 2020; Muffo & Bertino, 2021).

2.3.2 Select the model, preprocessing and training data set for each run

The Design-Expert softare determined the 32 combinations that define the level at which the model, type of preprocessing, and countries present in the training set are set for each run.

2.3.3 Train the model for each run

In this step the experiments are executed. For each of the 32 combinations defined by the Design-Expert software, a model to classify the political statements was trained.

2.3.4 Save the classification for each run

Finally, once each combination is trained, the response variable (accuracy) is placed in the template of the Design-Expert software.

2.4 Step 4. Statistical Analysis

In this article two statistical models are applied: resolution IV $2^{(10-5)}$ fractional factorial and a factorial 2^3 experimental design. The factorial designs allow to study the joint effect of the factor on a response variable; however to meet this goal the assumptions of normality, independence and homogeneity of variance must be met (Montgomery, 2013). Fractional factorial designs are designs in which the number of experiments is a fraction of the number of experiments for the same full factorial design (Antony, 2014). Although this approach helps to understand the impact of the main factors, the interactions of two factors can become confused

forming an alias structure. This describes a pattern of confounding that occurs in fractional factorial designs because the design does not include all combinations of factor levels (Antony, 2014). Nonetheless, this is not a problem since the sparsity of effects principle states that when a system has several variables it is more likely to be driven by the main effects and low order interactions (Montgomery, 2013).

To identify the important factors and interactions, the software State-Ease provides a half-normal probability plot. This graph represents the absolute value of the effect's estimates compared to their cumulative normal probabilities (Montgomery, 2013). Therefore, this statistical tool allows to identify the magnitude and importance of each effect. Non-significant effects exhibit a normal distribution centered near zero (Cuthbert, 1959). Finally, an Analysis of Variance (ANOVA) is performed to statistically determine the significance of each factor, for this the p-value must be less than the level of significance established in 0.05 (confidence level of 95%) (Pontes et al., 2016).

3. Results

3.1 Design of Experiments

The half-normal probability plot, which presents the experimental design results, is shown in Figure 3. The transformer model, Mexico´s manifestos, Chile´s manifestos, and the interaction between Argentina´s manifestos with stop words removal preprocessing have a statistically significant effect over the response variable. These terms are further analyzed with an ANOVA test. The rest of the factors are considered as part of the error of the model since the experimental design was unreplicated.

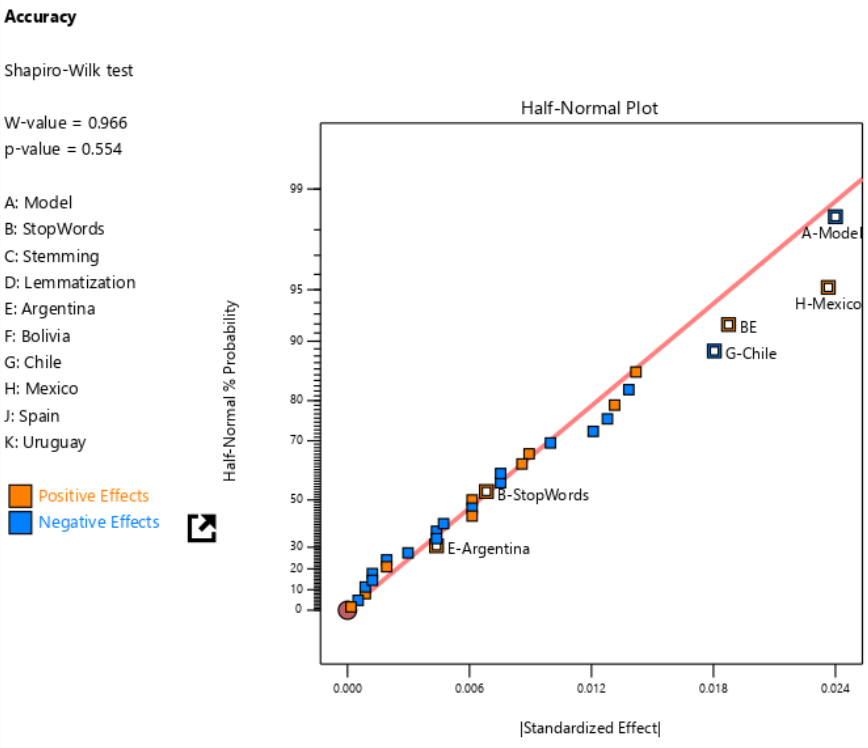

Figure 3. Half Normal Plot

The ANOVA was performed by using the significant factors mentioned above and the interaction with its hierarchy terms. Results are presented in Table 8 and show that the use of DistilBERT Transformer model, Mexico´s manifesto and Argentina´s manifestos with stop word removal preprocessing have a positive effect on the accuracy. However, the use of Chile´s manifestos in training set negatively affect the classification accuracy. All other factors are insignificant and do not have any impact on the response variable.

Table 8. ANOVA Results

Source	Sum of Squares	df	Mean Square	F-value	p-value
Model	0.0144	6	0.0024	5.45	**0.0010**
A-Model	0.0044	1	0.0044	10.02	**0.0040**
B-Stop words	0.0004	1	0.0004	0.8123	0.3760
E-Argentina	0.0001	1	0.0001	0.3338	0.5686
G-Chile	0.0025	1	0.0025	5.67	**0.0252**
H-Mexico	0.0043	1	0.0043	9.73	**0.0045**
BE	0.0027	1	0.0027	6.11	**0.0206**
Residual	0.0110	25	0.0004		
Cor Total	0.0255	31			

3.2 Training dataset effect

After carrying out the design of experiments explained above, some additional runs were executed with the DistilBERT model. A factorial 2^3 design was performed since the experimental design showed two countries (Mexico and Argentina with Stop Words preprocessing) positively affect the validation accuracy. In addition, Ecuadorian manifestos were considered as an additional factor to determine its impact on the classification model. In this way, the additional runs have three types of data sources in the training dataset:

1. The first group are experimental runs that use as training sets only manifestos of countries that improved the accuracy according to the experimental design.
2. The second group is only one run, that uses the 2021 Ecuadorian presidential election as training data. For that, the same partition (80/20) was applied to the data of the Ecuadorian elections of 2021 for the training set (1,447 sentences and quasi-sentences) and validation set (362 sentences and quasi-sentences) as shown previously in Table 7.
3. The third group of runs correspond to models that use a combination of the other groups training dataset (Ecuadorian, Mexican and Argentinian manifestos).

For these additional runs it was also necessary to determine the appropriate number of epochs that should be used to increase the classification accuracy. For this, some previous runs were carried out in which the number of epochs was increased by 25 following the same principle of the previous experimental design. In this way, it was determined that the number of suitable epochs for DistilBERT was 75 epochs and for RoBERTa was 25 epochs. This is because, if the number of epochs increases further, the validation accuracy begins to decrease, as shown in Figure 4 and Figure 5, respectively.

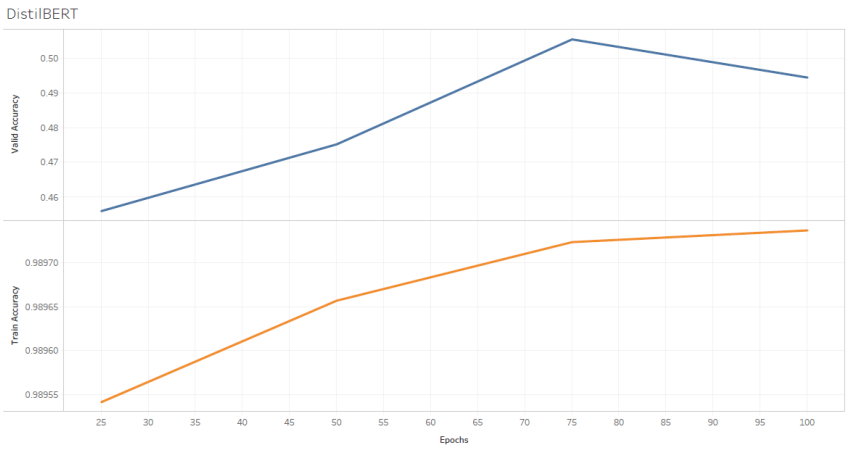

Figure 4. Learning curve for DistilBERT

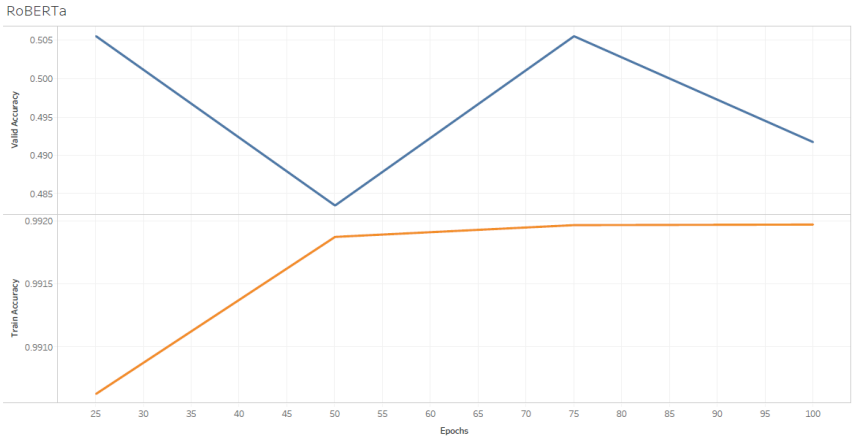

Figure 5. Learning curve for RoBERTa

In this way, the Design-Expert statistical software was used again to generate the template with the 8 combinations of the levels of the three factors used in this new experimental design. It is worth mentioning that since the experimental design has a combination in which none of these datasets are used, but the model must use at least some data, Bolivia´s manifestos were used in all runs since this was a non-significant factor. The results are shown in Table 9.

Table 9. Training dataset effect runs

Group	Training dataset	Accuracy
1	Bolivia, Mexico	0.4395
1	Bolivia and Argentina with stop words	0.4038
1	Bolivia, Mexico and Argentina with stop words	0.4423
2	Bolivia and Ecuador	0.8736
3	Bolivia, Mexico and Ecuador	0.8791

3	**Bolivia, Argentina with stop words and Ecuador**	**0.8901**
3	Bolivia, Mexico, Argentina with stop words and Ecuador	0.8791
NA	Bolivia	0.3736

These results were compared based on the accuracy of the validation set by using an Analysis of Variance (ANOVA) (Table 10). Even though the run that had the highest accuracy (0.8901) is the one that uses Argentinian and Ecuadorian manifestos, the only factor that is significant is Ecuador. It is evident that the use of Ecuadorian manifestos significantly impacts the classification accuracy. The runs that contain Ecuadorian manifestos as training data (group 2 & 3) have almost twice the accuracy of the runs that have only Mexico's and Argentina's manifestos (group 1).

Table 10. ANOVA Results

Source	Sum of Squares	df	Mean Square	F-value	p-value
Model	0.4352	3	0.1451	322.36	< 0.0001
Mexico	0.0012	1	0.0012	2.72	0.1746
Argentina - Stop Words	0.0003	1	0.0003	0.6792	0.4562
Ecuador	0.4337	1	0.4337	963.70	**< 0.0001**
Residual	0.0018	4	0.0005		
Cor Total	0.4370	4			

4. Discussion

In the present article, the sentences and quasi-sentences from Ecuadorian manifestos were automatically label into 7 policy domains using Transformer networks to facilitate more advance political analysis. Both model architectures, DistilBERT and RoBERTa, performed consistently well across the different combinations of the first experimental design. However, it was statistically proven that DistilBERT provides better results. This can be attributed to the fact that DistilBERT performs well with longer sequences as Büyüköz et al. (2020) study showed. This same study demonstrated that DistilBERT performs well under cross-context

settings, which can be useful in this specific case since the training data consists of manifestos from foreign countries and the model is evaluated on the classification of Ecuadorian government programs (Büyüköz et al., 2020). Additionally, the superior performance of DistilBERT may be due to overfitting, since large architectures, such as RoBERTa, tend to present this effect. This concept is related to the poor generalization capacity of the model. In other words, overfitting occurs when the model's performance in the training dataset is high but fails to perform well on unseen data (Salman & Liu, 2019).

In addition, studies such as Fusco et al. (2022), Baldeón & Lai-Yuen (2021), Jiao et al. (2020) and Baldeón & Lai-Yuen (2020) proposed models with smaller architectures that achieve competitive performances and significantly reduce the computational time. In fact, the article written by Fusco et al. (2022) mentions that the smaller architecture developed by the researchers (pNLP-Mixer) achieved a higher performance than RoBERTa. This is the same case as in the present study, since the smaller architecture DiltilBERT outperforms RoBERTa. Therefore, the use of shorter paths has demonstrated an improvement in the flow of information (Baldeón & Lai-Yuen, 2021). Furthermore, according to the study performed by Anggrainingsih et al. (2022) the difference of the results obtained from RoBERTa and DistilBERT were insignificant (between 1% and 3%). Therefore, other aspects such as training time, cost and computational complexity should be considered. Other studies such as Romell & Curman (2022), Cortiz (2021) and Shaheen et al. (2020) mentioned the same trade-off conclusions between computational time and small improvements in results.

On the other hand, the results related to the preprocessing operations showed that only the interaction of Argentina and stop words removal is statistically significant. Therefore, the combination of all the types of preprocessing is not necessary to achieve the highest classification performance. This is confirmed by several studies that state that the accuracy of

the NLP classification models does not improve by applying all preprocessing steps. On the contrary, the best preprocessing methods depend on each application (Işık & Dağ, 2020; Etaiwi & Nayma, 2017; Fernández Anta et al., 2013). Furthermore, according to Nayak et al. (2016) the process of stemming is highly dependent on the context, ergo the stemmed words do not have the required meaning to connect it with the context of the text. Conversely, the removal of stop words results in computational efficiency as these words can be considered noise for the models (Alshanik et al., 2020). Furthermore, in relation to the government programs from Spanish-speaking countries used as training data, the first experimental design demonstrated that Mexico and Argentina (with stop words removal) are statistically significant. This is due to the conservative nature of the dialect in Mexico and the Andean highlands (Guy, 2014). Additionally, Otheguy & Zentella (2012) classify Ecuadorian and Mexican dialects within the Mainlanders division due to the heterogeneity of the language.

Finally, DistilBERT and RoBERTa were tested on two types of training sets (Spanish-speaking countries' manifestos and Ecuadorian manifestos). In both datasets, the model achieved a performance comparable to text classification studies of political manifestos in Spanish (Dai & Radford, 2018; Rasov et al., 2020). It is worth mentioning that the highest levels of accuracy were obtained in the second experimental design, specifically in the combinations where the 2021 Ecuadorian Election Dataset was used as training data. The combination that generated the best accuracy (0.8901) was the one that had as training data the manifestos from Ecuador and Argentina. Nonetheless, the statistical analysis showed the only significant factor is the Ecuadorian government programs. This becomes evident by evaluating the same run without the Ecuadorian dataset in the training data, which yields an accuracy of less than half of the best run (0.4038). This is consistent with the study performed by AlBadawy et al. (2018), in which it is verified that the performance of the model is higher when the training set and validation set share the same origin.

CONCLUSIONS

Throughout this study, an efficient and objective way of analyzing the political programs of the Ecuadorian elections of 2021 was presented. This was done through automatic labeling of their sentences and quasi-sentences by Natural Language Processing (NLP) techniques using Transformer networks. In this way, it was possible to speed up and avoid biases in the political analyses given that the manual classification is time-consuming, labour-intensive, and can generate discrepancies by the political experts involved. Thus, the methodology used in this study was based on four main steps compiling Montgomery's guidelines to design an experiment and the steps of a basic data analysis process. It began by establishing the statement of the problem, selecting the response variable, the factors and levels, and determining the experimental design used, which in this case was a fractional factorial design 2^10-5 resolution IV. The factors of it include the type of Transformer model, the types of preprocessing, and the Spanish-speaking countries that are part of the training dataset. Next, the data to be used were collected from two data sources. The first corresponded to the Comparative Manifesto Project Dataset and the second corresponded to the data set of the Ecuadorian elections of 2021. These data were cleaned and preprocessed with basic techniques. Subsequently, the execution of the experiments was carried out considering the 32 combinations defined by the Design-Expert software and the statistical analysis of the results obtained was carried out.

In this way, it was possible to statistically answer the research questions raised for this study. The first question aims to determine the Transformer model that increases the accuracy of the classification of sentences and quasi-sentences of the electoral programs of Ecuador in 2021. It was found that DistilBERT is the model that presents the highest accuracy in the classification, showing a statistically significant difference with respect to the accuracy

presented by the RoBERTa model. The second research question aimed to determine what type of preprocessing increases the accuracy of the classification. It was found that none of the three types of preprocessing (stop words removal, stemming and lemmatization) improved the classification accuracy. This is because there was no statistically significant difference between using or not using them when preprocessing the data. However, the interaction of the stop words preprocessing with Argentina, that is one of the countries that are part of the training data set, was significant to improve the accuracy of the classification. The last research question aims to determine which government programs in Spanish-speaking countries used as training data increase the accuracy of the classification model. As mentioned, Argentina's interaction with Stop Words preprocessing helped improve classification accuracy. Additionally, the political programs of Mexico showed a statistically significant difference when including this data in the training set, improving the accuracy of the classification.

Once these results were obtained, it was decided to inquire about the training dataset effect when the Ecuadorian database is used as part of the training dataset. For this, three groups of experimental runs were defined, and the result was that the classification accuracy increased significantly when using the data from Argentina with the preprocessing of stop words removal and the data from Ecuadorian government programs as training data. Therefore, it was determined that the presence of the data from Ecuador in the training set has a statistically significant impact on the accuracy of the classification.

LIMITATIONS AND RECOMMENDATIONS

The present study used an unreplicated factorial design because time and computational resources were restricted. Therefore, the first experimental design may be fitting the model to noise. Even though it was assumed that third and higher-order interactions were negligible, and their mean square was combined to estimate an error, it should be noted that there is no estimate of the pure error. Another limitation is presented in the second experimental design since one of the eight combinations was not feasible and all runs had to include Bolivia´s manifestos.

To avoid the use of Bolivia´s manifestos in the second experimental design it is suggested to analyze the training dataset effect with a single factor experiment with four levels. The factor corresponds to the training dataset and the levels are four datasets composed of government programs from different countries as follows: 1) Ecuador, 2) Ecuador and Argentina, 3) Ecuador and Mexico, 4) Ecuador, Argentina and Mexico. It should be noted that Argentina´s manifestos should be preprocessed with the stop words technique. In addition, to analyze this experiment replicates should be executed. To achieve this a k-fold cross-validation should be performed. It is recommended to use a k=5 so the validation set has 20% of the total observations in the Ecuadorian manifestos. In this way, each replicate would use one-fold as the validation set and four-folds would serve as the training set.

In this paper all the available manifestos of Spanish-speaking countries were used as training data. However, further studies could analyze the impact of lexical evolution in the response variable by comparing recent and older manifestos in the training dataset. It is worth mentioning that similar studies could use the F1-score as the response variable since false positives and false negatives are considered to have an equal impact on the results.

BIBLIOGRAPHIC REFERENCES

Abercrombie, G., Nanni, F., Batista-Navarro, R., & Simone, P. (2019, November). Policy Preference Detection in Parliamentary Debate Motions. *Association for Computational Linguistics, Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 249–259. https://doi.org/10.18653/v1/K19-1024

AlBadawy, E. A., Saha, A., & Mazurowski, M. (2018). Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys, 45*(3), 1150-1158. https://doi.org/10.1002/mp.12752

Alshanik, F., Apon, A., Herzog, A., Safro, I., & Sybrandt, J. (2020). Accelerating Text Mining Using Domain-Specific Stop Word Lists. *International Workshop on Big Data Reduction 2020.* Clemson: IEEE. https://doi.org/10.1109/BigData50022.2020.9378226

Anggrainingsih, R., Mubashar Hassan, G., & Datta, A. (2022, May 05). Evaluating Pre-trained BERT-based Language Models for Detecting Misinformation. *Research Square*(1). https://doi.org/10.21203/rs.3.rs-1608574/v1

Antony, J. (2014). Fractional Factorial Designs. In J. Antony, *Design of Experiments for Engineers and Scientists* (pp. 87-112). Elsevier B.V. https://doi.org/https://doi.org/10.1016/B978-0-08-099417-8.00007-9

Babanejad, N., Agrawal, A., An, A., & Papagelis, M. (2020, July). A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks. *Association for Computational Linguistics, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5799–5810. https://doi.org/10.18653/v1/2020.acl-main.514

Baldeón, M., & Lai-Yuen, S. (2020). AdaEn-Net: An ensemble of adaptive 2D–3D Fully

Convolutional Networks for medical image segmentation. *Neural Networks, 126,* 76-94. https://doi.org/10.1016/j.neunet.2020.03.007

Baldeón, M., & Lai-Yuen, S. (2021). EMONAS-Net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3D medical image segmentation. *Artificial Intelligence In Medicine, 119*(102154). https://doi.org/10.1016/j.artmed.2021.102154

Bielik, I. (2020). Application of natural language processing to the electoral manifestos of social democratic parties in Central Eastern European countries. *The Journal of the Central European Political Science Association, 16*(1), 259-282. https://doi.org/10.2478/pce-2020-0012

Bilbao-Jayo, A., & Almeida, A. (2018a). Political discourse classification in social networks using context sensitive convolutional neural networks. *Association for Computational Linguistics,* 76-85. https://aclanthology.org/W18-3513.pdf

Bilbao-Jayo, A., & Almeida, A. (2018b). Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks, 14*(11). https://doi.org/10.1177/1550147718811827

Büyüköz, B., Hürriyetoglu, A., & Özgür, A. (2020). Analyzing ELMo and DistilBERT on Socio-political News Classification. https://aclanthology.org/2020.aespen-1.4.pdf

Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2016). A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence,* 50-57. https://doi.org/10.1016/j.engappai.2016.01.007

Chatsiou, K. (2020). Text Classification of Manifestos and COVID-19 Press Briefings using BERT and Convolutional Neural Networks. *arXiv: Computation and Language*. https://doi.org/10.48550/arXiv.2010.10267

Clark, K., Luong, M.-T., Le, Q., & Manning, C. (2020). ELECTRA: Pre-training Text Ecoders as Dsicriminators Rather than Generators. https://doi.org/10.48550/arXiv.2003.10555

Constitución de la República de Ecuador [Const]. (2008). *Artículo 109.*

Cortiz, D. (2021, April 05). Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. *Computer Science*. https://doi.org/10.48550/arXiv.2104.02041

Cumbicus-Pineda, O., Ordoñez-Ordoñez, P., Neyra-Romero, L., & Figueroa-Diaz, R. (2018). Automatic Categorization of Tweets on the Political Electoral Theme Using Supervised Classification Algorithms. (M. P.-P. Botto-Tobar, Ed.) *Technology Trends, 895*, 671-682. https://doi.org/10.1007/978-3-030-05532-5_51

Cuthbert, D. (1959, November). Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments. *Technometrics, 1*(4), 311-341. http://www.jstor.org/stable/1266715

Dai, Y., & Radford, B. (2018). Multilingual Word Embedding for Zero-Shot Text Classification. https://yaoyaodai.github.io/files/Dai_0BlinC.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Association for Computational Linguistics, Volume I*, 4171-4186. https://doi.org/arXiv:1810.04805

Etaiwi, W., & Nayma, G. (2017). The Impact of applying Different Preprocessing Steps on Review Spam Detection. *Procedia Computer Science* , 273–279. https://doi.org/10.1016/j.procs.2017.08.368

Fernández Anta, A., Núñez Chiroque, L., Morere, P., & Santos, A. (2013, March). Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. *Procesamiento del Lenguaje Natural*(50), 45-52.

https://www.redalyc.org/pdf/5157/515751576005.pdf

Fusco, F., Pascual, D., & Staar, P. (2022, February 09). pNLP-Mixer: an Efficient all-MLP Architecture for Language. *Computer Science*. https://doi.org/10.48550/arxiv.2202.04350

Glavaš, G., Nanni, F., & Ponzetto, S. (2017, August). Cross-Lingual Classification of Topics in Political Texts. *Association for Computational Linguistics, Proceedings of the Second Workshop on NLP and Computational Social Science*, 42–46. https://doi.org/10.18653/v1/W17-2906

Guy, G. R. (2014). Variation and change in Latin American Spanish and Portuguese. In P. Amaral, & A. M. Carvalho (Eds.), *Portuguese-Spanish Interfaces: Diachrony, synchrony, and contact* (pp. 443-464). John Benjamins Publishing Company. https://doi.org/10.1075/ihll.1

Hallac, I., Ay, B., & Aydin, G. (2018). Experiments on Fine Tuning Deep Learning Models With News Data For Tweet Classification. https://doi.org/10.1109/IDAP.2018.8620869

Işık, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering & Computer Sciences, 28*, 1405 – 1421. https://doi.org/10.3906/elk-1907-46

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., . . . Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. *Computer Science, 5*, 4163–4174. https://doi.org/10.48550/arXiv.1909.10351

Kadhim, A. (2018). An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security, 16*, 22-32.

Kent, S., & Krumbiegel, T. (2021). CASE 2021 Task 2 Socio-political Fine-grained Event Classification using Fine-tuned RoBERTa Document Embeddings.

Khurana, D., Koli, A., Khatter, K., & Sukhdev, S. (2022). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13428-4

Kohn, A., Boey, D., & Béchara, H. (2021, April). Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks. *SocArXiv*. https://doi.org/10.31235/osf.io/fjh4q

Koroteev, M. (2021). BERT: A Review of Applications in Natural Language Processing and Understanding. *arXiv*. arXiv:2103.11943v1

Lehmann, P., & Zobel, M. (2018). Positions and saliency of immigration in party manifestos: A novel dataset using crowd coding. . *European Journal of Political Research*, 1056-1083. http://hdl.handle.net/10419/247344

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *ICLR*. https://doi.org/1711.05101

Mehrishi, R. (2019). *INTOSAI Working Group on IT Audit.* https://www.intosaicommunity.net/document/knowledgecenter/WGITA_Data_Analytics_Guideline_Final_QAC.pdf

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *ArXiv*. https://arxiv.org/abs/1309.4168

Montgomery, D. (2013). *Design and Analysis of Experiments.* Arizona: John Wiley & Sons, Inc.

Muffo, M., & Bertino, E. (2021). BERTino: an Italian DistilBERT model. *Proceedings of the Seventh Italian Conference on Computational Linguistics, Volume 2769*, 317-322.

http://ceur-ws.org/Vol-2769/paper_09.pdf

Nayak, A., Kanive, A., Chandavekar, N., & Dr. Balasubraman, R. (2016, June). Survey on Pre-Processing Techniques for Text Mining. *International Journal Of Engineering And Computer Science, 5*(6), 16875-16879. https://www.researchgate.net/publication/327337746_Survey_on_Pre-Processing_Techniques_for_Text_Mining

Orellana, G., Arias, B., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018). A study on the impact of pre-processing techniques in Spanish and English text classification over short and large text documents. *International Conference on Information Systems and Computer Science (INCISCOS)*, 277-283. https://doi.org/10.1109/INCISCOS.2018.00047

Otheguy, R., & Zentella, A. C. (2012). Continuity, Language Contact, and Dialectal Leveling in Spanish in New York. In R. Otheguy, & A. C. Zentella, *Spanish in New York: Language Contact, Dialectal Leveling, and Structural Continuity* (pp. 2-24). Oxford Studies in Sociolinguistics. https://doi.org/10.1093/acprof:oso/9780199737406.003.0001

Öztürk, O., & Özcan, A. (2022). Ideology Detection Using Transformer-Based Machine Learning Models. 30-52. https://doi.org/10.13140/RG.2.2.12303.51362

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://doi.org/10.18653/v1/N18-1202

Pinta, M., Medina-Pérez, P., Riofrío, D., Pérez, N., Benítez, D., & Flores, R. (2021). Automatic Manifesto Comparison using NLP Techniques and The Manifesto Project Domains - Case Study: 2021 Ecuadorian Presidential Elections. *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, 1-7.

https://doi.org/10.1109/ETCM53643.2021.9590825

Pontes, F., Amorim, G., Balestrassi, P., Paiva, A., & Ferreira, J. (2016). Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing, 186*, 22-34. https://doi.org/https://doi.org/10.1016/j.neucom.2015.12.061

Rasov, A., Obabkov, I., Olbrich, E., & Yamshchikov, I. (2020). Text Classification for Monolingual Political Manifestos with Words Out of Vocabulary. 149-145. https://doi.org/10.5220/0009792101490154

Riofrío, D., Almeida, P., Dávalos, J., Flores, R., Pérez, N., Benítez, D., & Medina, P. (2021, February 26). Towards Automatic Comparison of Online Campaign Versus Electoral Manifestos. (A. L.-L.-G. Orjuela-Cañón, Ed.) *Springer Cham, 134*, 60-73. https://doi.org/10.1007/978-3-030-69774-7_5

Romell, A., & Curman, J. (2022, March 07). Multilingual Large Scale Text Classification for Automotive Troubleshooting Management. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9076713&fileOId=9076714

Salman, S., & Liu, X. (2019, January 19). Overfitting Mechanism and Avoidance in Deep Neural Networks. *Computer Science ArXiv*. https://doi.org/10.48550/arXiv.1901.06566

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

Schoonvelde, M., Schumacher, G., & Bakker, B. (2019). Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology. *Journal of Social and Political Psychology*, 124-143.

https://doi.org/10.5964/jspp.v7i1.964

Shaheen, Z., Wohlgenannt, G., & Filtz, E. (2020, October 24). Large Scale Legal Text Classification Using Transformer Models. *Computer Science*. https://doi.org/10.48550/arXiv.2010.12871

State-Ease. (2022). *Design-Expert VERSION 13*. Stat-Ease, Inc.: https://www.statease.com/software/design-expert/

Subramanian, S., Cohn, T., & Baldwin, T. (2018, June). Hierarchical Structured Model for Fine-to-coarse Manifesto Text Analysis. *Association for Computational Linguistics, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1964-1974. http://hdl.handle.net/11343/258659

Subramanian, S., Cohn, T., Baldwin, T., & Brooke, J. (2017, December). Joint Sentence–Document Model for Manifesto Text Analysis. *Proceedings of the Australasian Language Technology Association Workshop 2017*, 25-33. https://aclanthology.org/U17-1003

Suiter, J., & Farrell, D. (2011). The Parties' Manifestos. In J. Suiter, & D. Farrell, *How Ireland Voted 2011* (pp. 29-46). London: Palgrave Macmillan. https://doi.org/10.1057/9780230354005_2

Tabassum, A., & Patil, R. (2020, June). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology (IRJET), 7*(6), 4864-4867. https://www.irjet.net/archives/V7/i6/IRJET-V7I6913.pdf

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J., & Bonneau, R. (2020). A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics. *Social Science Research Network*.

https://doi.org/10.2139/ssrn.3724644

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention Is All You Need. *Conference on Neural Information Processing Systems*, 3-31.

Volkens, A., Burst, T., Krause, W., Lehmann, P., Theres, M., Regel, S., . . . Zehnter, L. (2021). *Manifesto Project Main Dataset (Party Preferences)*. Manifesto Project: https://doi.org/10.25522/manifesto.mpds.2021a

Werner, A., Lacewell, O., Volkens, A., Matthieß, T., Zehnter, L., & van Rinsum, L. (2021). *Manifesto Coding Instructions* . Manifesto Project Dataset: https://manifesto-project.wzb.eu/

Wiedemann, G. (2018). Proportional Classification Revisited: Automatic Content Analysis of Political Manifestos Using Active Learning. *Social Science Computer Review*. https://doi.org/10.1177/0894439318758389

Zamora, E. J. (2021). Political Science in Ecuador, 2005-2019. A discipline in search of institutionalization. *Íconos*. https://doi.org/10.17141/iconos.70.2021.4667

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., & Stuckenschmidt, H. (2016). Classifying Topics and Detecting Topic Shifts in Political Manifestos. *International Conference on the Advances in Computational Analysis of Political Text*, (pp. 88-93). Dubrovnik.