UNIVERSIDADE FEDERAL FLUMINENSE

JÉSSICA SOARES DOS SANTOS

A Dataset Ranking Approach for Transfer Learning to Support Sentiment Analysis in Electoral Scenarios

NITERÓI

2023

UNIVERSIDADE FEDERAL FLUMINENSE

JÉSSICA SOARES DOS SANTOS

A Dataset Ranking Approach for Transfer Learning to Support Sentiment Analysis in Electoral Scenarios

Thesis presented to the Computing Graduate Program of the Universidade Federal Fluminense as a partial fulfillment of the requirements for the degree of Doctor of Computing. Area: Computer Science.

Advisors:

Flavia Bernardini

Aline Paes

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

```
S237d   Santos, Jéssica Soares dos
          A Dataset Ranking Approach for Transfer Learning to Support
        Sentiment Analysis in Electoral Scenarios / Jéssica Soares
        dos Santos. - 2023.
          131 f.: il.

          Orientador: Flavia Cristina Bernardini.
          Coorientador: Aline Marins Paes Carvalho.
          Tese (doutorado)-Universidade Federal Fluminense, Instituto
        de Computação, Niterói, 2023.

          1. Mineração de opiniões (Computação). 2. Aprendizado de
        máquina. 3. Mídia social. 4. Eleição. 5. Produção
        intelectual. I. Bernardini, Flavia Cristina, orientadora. II.
        Carvalho, Aline Marins Paes, coorientadora. III. Universidade
        Federal Fluminense. Instituto de Computação.IV. Título.

                                              CDD - XXX
```

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

JÉSSICA SOARES DOS SANTOS

A Dataset Ranking Approach for Transfer Learning to Support Sentiment Analysis in Electoral Scenarios

Thesis presented to the Computing Graduate Program of the Universidade Federal Fluminense as a partial fulfillment of the requirements for the degree of Doctor of Computing. Area: Computer Science.

Approved in March, 2023.

EXAMINATION BOARD:

Prof. Dr. Flavia Cristina Bernardini - Advisor, UFF

Aline Marins Paes Carvalho

Prof. Dr. Aline Marins Paes Carvalho - Advisor, UFF

Prof. Dr. Alexandre Plastino de Carvalho, UFF

Prof. Dr. José Viterbo Filho, UFF

Rezende.

Prof. Dr. Solange Oliveira Rezende, USP

Prof. Dr. Edimara Mezzomo Luciano, PUCRS

Niterói

2023

Resumo

Pesquisas eleitorais tradicionais envolvem a realização periódica de entrevistas com pessoas de diferentes regiões geográficas, demandando tempo, recursos financeiros, e esforços humanos. A disponibilidade de uma enorme quantidade de opiniões na Web surgiu como uma alternativa às pesquisas eleitorais tradicionais, devido ao fato de que esse tipo de dado pode ser coletado automaticamente de forma mais rápida e barata. Nesta tese, propomos um método que pode ser útil para analisar as opiniões dos eleitores com base na análise de sentimentos de dados do Twitter. Para lidar com a falta de dados rotulados nesse domínio, utilizamos técnicas de aprendizado por transferência (*transfer learning*) e aproveitamos o conhecimento prévio obtido com datasets rotulados que podem pertencer a outros domínios e idiomas. Um método de seleção de datasets baseado em ranqueamento de acordo com a similaridade é apresentado para lidar com este problema. Os resultados de nossos experimentos sugerem que a análise da (dis)similaridade entre datasets pode ser útil para escolher o conjunto de dados mais apropriado para transferir conhecimento e obter melhores previsões de sentimentos em cenários eleitorais. Outra contribuição deste trabalho é a disponibilização de um dataset eleitoral coletado do Twitter manualmente rotulado de acordo com diferentes dimensões: análise de sentimentos, presença discurso ofensivo, e posicionamento a favor ou contra os candidatos, considerando dados do Twitter sobre as eleições presidenciais brasileiras de 2018. Levando em consideração o conjunto de dados rotulados manualmente, realizamos um estudo que destaca o alto grau de divergência de rotulagem em cenários eleitorais e algumas das características que tornam a análise de opiniões eleitorais em mídias sociais uma tarefa complexa. O método para seleção de datasets e o dataset eleitoral fornecido podem ser adotados para auxiliar na condução de análises de eleições futuras.

Palavras-chave: análise de eleições; mineração de dados; análise de sentimentos; ranqueamento de datasets; seleção de datasets; similaridade de datasets; classificação de textos curtos; transfer learning.

Abstract

Traditional election polls are based on conducting interviews periodically with people from different geographical regions, requiring time, financial resources and human efforts. The availability of a huge amount of data in the Web containing opinions from potential electors has arisen as an alternative to traditional polls, due to the fact that this kind of data can be gathered automatically in a faster and cheaper way. In this thesis, we are concerned with proposing a method that can be useful to analyze electorate opinions based on Twitter data and sentiment analysis methods. In order to deal with the lack of labeled data in this scenario, we use transfer learning techniques and take advantage of prior knowledge achieved with other datasets that may belong to other domains or languages. A dataset selection strategy based on similarity ranking is presented to tackle this problem. Our experimental results suggest that analyzing the (dis)similarity between different datasets may be useful to choose the most proper dataset for transfer learning and achieve better sentiment predictions in electoral scenarios. Another contribution of this work is to provide an electoral dataset manually labeled according to different dimensions: sentiment analysis, offensive speech presence, and candidate support, considering Twitter data related to the 2018 brazilian presidential election. Taking into account the manually labeled dataset, we conducted a study that highlights the high degree of labeling divergence in electoral scenarios and some of the characteristics that make the analysis of electoral social media opinions a complex task. The dataset selection method and the electoral dataset provided by this research may be adopted to aid the analysis of future election forecasts.

Keywords: election analysis; data mining; sentiment analysis; dataset ranking; dataset selection; dataset similarity; short text classification; transfer learning.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	:	Artificial Intelligence;
BoW	:	Bag-of-Words;
BERT	:	Bidirectional Encoder Representations from Transformers;
CB	:	Confidence Based Measures;
CBoW	:	Continuous Bag of Words;
CNN	:	Convolutional Neural Networks;
DT	:	Decision Trees;
ELMo	:	Embeddings from Language Models;
GloVe	:	Global Vectors from Word Representation;
KNN	:	K-Nearest Neighbor;
LDA	:	Latent Dirichlet Allocation;
LIWC	:	Linguistic Inquiry and Word Count;
LM	:	Language Modeling;
LR	:	Logistic Regression;
MLP	:	Multi-Layer Perceptron;
NB	:	Naive Bayes;
NLP	:	Natural Language Processing;
RCA	:	Reverse Classification Accuracy;
RF	:	Random Forest;
RNN	:	Recurrent Neural Networks;
SCL	:	Structural Correspondence Learning;
SFA	:	Spectral Feature Alignment;
SG	:	Sentiment Graphs;
SLR	:	Systematic Literature Review;
SVM	:	Support Vector Machine;
TVC	:	Target Vocabulary Covered;
TF-IDF	:	Term Frequency-Inverse Document Frequency;
ULMFit	:	Universal Language Model Fine-tuning;
USEM	:	Universal Sentence Encoder Multilingual;

WVV	:	Word Vector Variance;

Summary

Chapter 1

Introduction

Elections are fundamental components for democracy as they allow citizens to choose their next political chiefs. According to Przeworski et al. [105], an electoral system is classified as democratic when the four conditions hold: (i) the political executive chief is elected; (ii) the political legislature is elected; (iii) the election involves more than one political party; and (iv) alternation: the opposition must have the real possibility to win the next elections, under identical rules, and taking office.

Choosing a political candidate to vote in the elections can be a challenging issue since it means selecting proposals and policies advocated by several political parties that may or may not share the same ideals of the citizens [46]. This choice becomes harder when the party political leaning is not well-defined. In the brazilian electoral domain, for instance, there is a high number of parties and many of them are not truly aligned with the left or right wings [65]. On the other hand, the election process is also challenging to the candidates since they need to focus on the questions that matter most to them at the same time that they need to clearly and concisely to communicate with the citizens to gain their votes.

In this way, in democratic systems, *election polls* play an essential role. They can measure voting intention [39] and their results can affect election outcomes [46], by influencing people that have not decided yet in which candidate to vote [112]. Additionally, election polls can be used by the candidates and their parties to adjust their campaigns and better communicate proposals [39, 154].

1.1 Motivation

The traditional way of predicting election outcomes is based on opinion surveys that include face-to-face or phone interviews and questionnaires. These polls are conducted involving people from different regions that have different profiles, such as, people that live in urban or rural zones, people with different ethnicity, age and gender. However, traditional polls involve some drawbacks [145], [62]:

- they demand much time to be conducted;
- they demand high monetary costs;
- they requires extensive human efforts to collect data (opinions) across the nation/state possibly making people living in hard-to-access regions less represented.

In addition, although most of the time the traditional polls can correctly predict the results of elections, there are some cases when they were not successful. This is the case, for instance, of the 2015 UK general elections, as mentioned in [27, 133], and the 2016 US presidential election, mentioned in [20, 155].

Taking into account the disadvantages of the traditional election polls, a number of approaches in the literature have proposed to predict voting intention by applying machine learning to data collected from social media [87]. According to Bovet, Morone and Makse (2018) [19], forecasting opinion trends based on data collected from the internet is one of the main goals of Big Data [115]. In addition, the growth of social media users brings the virtual community closer to the real community [132], although there are still restrictions on access. For this reason, social media may be explored as a new way of collecting data that can be utilized to analyze future outcomes in the real world [7, 23]. In this context, there is an increasing number of approaches in recent years that use data from social media in order to predict political elections results [103, 120]. Most of them adopt sentiment analysis techniques to help on this task [88]. We present a summary about existing strategies for forecasting elections using social media and computational techniques in Chapter 3.

Sentiment analysis has been largely adopted to infer people opinions about different topics in several areas [97]. Basically, it consists in determining the polarity of sentences as, for example, positive, negative or neutral sentiment. In order to analyze electorate's opinions, or even to predict elections outcomes [88], the negative/positive sentiment is inferred to the candidates and, from that, it is possible to point out the one that seems to be the favorite among people.

There are two main approaches that can be adopted to analyze opinions and classify them according to sentiment [72, 78]. While *dictionary* approaches associate words with the polarity they denote and compute sentiment sentence based on that, *corpus-based* approaches depend on labeled datasets with examples of positive, negative and neutral sentences to train classification models that are able to predict sentiment of unlabeled sentences. For domain specific problems, the latter approach usually achieves better predictions [72], as the target domain may contain specific terms that do not appear in generic dictionaries (a discussion about this issue is presented in Chapter 4).

Analyzing social media electoral opinions is not a trivial task. In addition to the challenges that are inherently related to social media data such as (i) *loss of context* [32, 116] – restriction in the number of characters of the posts –, (ii) *spam* [141] – fake content posted by bot accounts –, and (iii) *sarcasm and irony* [44], the electoral domain presents another particularities such as: (iv) *vocabulary with specific terms* – as hashtags that combine support messages with candidate names or campaign slogans [119]; (v) *dynamic data* [53, 86] – vocabulary of terms used by people to express their opinions about elections can change over time according to sub-events such as debates, speeches in interviews or political scandals; (vi) *short time for labeling* electorate opinions [121] – there is no enough time to manually annotate electorate thousands of social media electoral opinions reliably, during the short period of campaigns. Previous studies [79, 149] show that items (ii) and (iii) are intensified in the electoral domain.

We conducted a literature survey to better understand this domain problem and noticed that existing approaches for predicting electoral trends/ outcomes based on social media still present some pitfalls and limitations, as we better discuss in Chapter 3. In short, the difficulty of collecting and labeling a large number of tweets during the short period of elections caused that many approaches choose to conduct a post-hoc analysis of electoral tweets, *i.e.* they only analyze tweets after the occurrence of the real elections [51]. Most of the approaches that try to predict election results do not consider information specific of the domain to assign polarities, relying only on generic lexical dictionaries [23], [139], [136]. Only a few go on other directions. For instance, Heredia, Prusa and Khoshgoftaar (2017) [55] and Prasetyo and Hauff (2015) [39], which use methods that label tweets automatically according to emoticons – this is also problematic as the emoticons may not have the same polarity as the text.

Therefore, being aware of all the challenges of analyzing social media electoral opinions, in this thesis we focus on the ones that are closely related to the electoral domain (iv – vi). One possibility to deal with the lack of labeled data in this domain that still has not been well explored in the literature is the usage of existing *sentiment analysis* datasets from other domains as a *starting point* to construct models to be applied in electoral scenarios, what would enable the analysis of prior or future elections. This task can be seen as an instance of domain transfer learning [95]. To mitigate problems that may occur due to high divergence between training data and target data, we propose to adopt similarity metrics that help the selection of similar datasets. Our assumption is that the target electoral dataset may share characteristics with similar datasets, what would entail better *sentiment analysis* predictions.

1.2 Research Questions

This thesis focuses on addressing the following Research Questions (RQs):

RQ1: How machine learning algorithms and natural language processing may aid electoral analysis using social media?

RQ2: What are the existing computational approaches to analyze elections using social media?

In order to answer our first two RQs, we conducted a Systematic Literature Review, from which we constructed a survey, where we indicated many future research lines. We could observe that one interesting research line was related to out RQ3:

RQ3: How to take advantage of existing labeled datasets (of other domains) to better analyze electoral opinions using computational techniques?

So, considering our survey and our RQ3, we present the following hypothesis of this thesis:

Hypothesis (H): *If there is a high degree of similarity between a source labeled sentiment analysis dataset and a target electoral dataset, then machine learning classifiers trained with this source dataset will achieve proper sentiment predictions for the target electoral dataset.*

In this context, the main *goal* of this thesis is to *propose a dataset selection method that aids electoral analysis using social media data*, by improving the predictions related to the *sentiment analysis* task in the electoral domain.

1.3 Methodology

This thesis is conducted addressing the following steps:

1. Survey the state of the art of approaches that forecast elections using data from social media;
2. Conduct a study about approaches in the literature that deal with the problem of lack of annotated data in a given domain [120];
3. Create and analyze a (manually) labeled dataset of tweets related to the 2018 Brazilian Presidential Elections in Portuguese [119];
4. Investigate if the similarity between datasets can help the task of choosing the most proper sentiment analysis dataset (among the ones found in the literature) aiming at reusing knowledge [121, 122].

The similarity investigation is done by selecting a set of metrics to measure dataset similarity. From that, we propose to create a unified similarity ranking that will point out to the user datasets that are likely to achieve satisfactory sentiment analysis predictions when labeled data in the target domain is not available. The validation is conducted by comparing the predictions achieved by the datasets selected by our method and the predictions achieved with labeled target data.

1.4 Contributions

The main contributions of this thesis are as follows:

- Provide an annotated dataset[1] with Portuguese tweets of the electoral domain;
- Provide the state of the art of existing methods to forecast elections with social media data;
- Present a study that measures the degree of divergence when labeling tweets in the electoral scenario using a manual annotation process;
    - This study also highlights a set of characteristics of electoral data that make the labeling process of this data a complex task (as described in Section 5.1);

[1] due to privacy issues the text of the tweets will not be provided but only their IDs.

- Present a method for dataset selection that can aid the sentiment analysis tasks when labeled data is not available, such as in electoral scenarios.
  - The adoption of multilingual embeddings – factor that distinguishes our work from related ones – allows us to take advantage of data from different languages, what is very interesting in domains such as the electoral one.
- Present a case study of the application of the proposed method in a real political election.

1.5 Thesis Proposal Structure

The general structure of this thesis proposal is organized as follows: In the **Background** (Chapter 2), we present the main concepts needed for understanding this thesis, including Natural Language Processing (NLP) and Machine Learning concepts, Transfer Learning approach, Representational Models for Texts, Distance Metrics, Sentiment Analysis, and so on. In the **Related Work** (Chapter 3) we present related approaches along two dimensions. The first one is related to (i) approaches that propose means to forecast elections based on social media. We present the state of the art of existing methods in the literature for forecasting elections based on social media, after conducting a systematic literature review. From that, we also identified the main limitations, open issues and lines for future research. In the second dimension, we conducted an ad hoc search to find (ii) approaches that deal with classification tasks and the lack of annotated data in a given target domain, as we observed that it is a critical problem for electoral data analysis. In the **Thesis Proposal** (Chapter 4) we present the proposal, by relating our hypotheses to the main study topics, and the proposed dataset selection method is detailed. In the **Experiments** (Chapter 5), we present the experimental results of this research, the steps adopted for building the manually labeled electoral dataset, and a divergence analysis about the labeling process. Finally, in the **Conclusions** (Chapter 6), we present our conclusions, point out limitations and threats to validity, lines for future research, and the publications, presentations and awards related to this thesis.

Chapter 2

Background

Natural Language Processing (NLP) is a research topic concerned with analyzing how computational techniques can be applied on natural language to manipulate and extract textual meaning in order to perform useful tasks. By *natural language*, we refer to the language used by humans to communicate with each other (text or speech). NLP combines different disciplines such as computer science, artificial intelligence and linguistics [30]. The applications of NLP include: machine translation, speech recognition, text summarization, text processing, and so on. This chapter outlines the main topics that will be used along of this thesis in order to help the understanding of this research. In general, they are topics used by NLP tasks. We begin by introducing opinion mining, sentiment analysis and offensive speech concepts. After that, we present machine learning basic concepts, as the ones that are used by NLP tasks. Other topics such as transfer learning, bag-of-words, word embeddings and language models that allow transference, are also presented. Finally, we introduce the distance metrics that can be used to measure distance between word vectors as this topic is explored in the experiments conducted in this thesis.

2.1 Opinion Mining and Sentiment Analysis

The popularization of social media has arisen as a new way to collect and analyze people opinions. This is because the amount of available opinions about different topics on the web is increasing more and more. Most of the approaches adopt the terms *opinion mining* and *sentiment analysis* interchangeable. Pang and Lee (2008) [97] and Liu (2012) [83] are examples of authors that argue that opinion mining and sentiment analysis refer to the same field of study.

Liu (2012) [83] defines this topic as follows:

> "Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes."

In this thesis, we use the term *opinion mining* to refer to the sub-field of study of NLP that is responsible for analyzing people opinions, and use the term *sentiment analysis* to refer to the sub-field of Opinion Mining that is responsible for detecting feelings in texts, computationally inferring the sentiment polarity of them. In this way, we state that *opinion mining* is a broader concept than *sentiment analysis* considering that detecting sentiment in a text is a step needed to capture the meaning of an opinion in textual format.

There are two main approaches to sentiment analysis [72, 78]:

1. *Dictionary- or Lexical-Based approach:* they are the approaches that use dictionaries that relate a word to a sentiment (positive, negative, neutral). In this way, the calculus of the polarity of a sentence is based on the semantic orientation of the words that belong to it. Therefore, each word is associated with a (positive or negative) score, and the sum of word scores belonging to a sentence results in its final score. Usually, those methods can be associated with a set of predefined rules that can change the score of the words in a sentence when combined. For instance, when a negation term ("not") precedes a word, its score can be discarded or considered as a negative one. An example of lexicon resource that can be used with this method is the Linguistic Inquiry and Word Count (LIWC) [98], which groups words according to different categories, including the *posemo* category, which stands for positive emotion, and *negsemo* category, which stands for negative emotion;

2. *Corpus-Based approach:* they are the approaches that build machine learning classifiers to assign the polarity of a sentence, constructed using datasets of examples, in which each training instance (sentence) is associated with a label denoting its polarity (e.g.: positive, negative or neutral). Therefore, they depend on labeled datasets for training a classification model that predict what is the sentiment of an input sentence. These methods use a set of features to distinguish sentences with different sentiments.

Different from the dictionary-based approach, the corpus-based approach can deal with specific domain/contexts. The study presented by Kharde and Sonawane [72] concludes that corpus-based approaches usually present better predictions than dictionary-based approaches for sentiment analysis tasks. Because of that, we decided to use the corpus-based approach in our proposal.

2.2 Hateful and Offensive Language

The anonymity afforded by social media is a factor that favors the breeding and spreading of hateful or offensive content on that kind of environment [157]. Statistics show an increasing of online hate speech and offensive language during elections or political events [93], [14], [101], [41], [110]. Gao and Huang (2017) [49] already have observed that it occurs mainly when the elections are polarized.

Following the definition presented by Fortuna and Nunes (2018) [47], hate speech is defined as follows:

> "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used."

The definition presented by the Cambridge dictionary[1] is subtly different:

> "Public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation (= the fact of being gay, etc.)."

In this way, the term *hate speech* refers to any communication that disparages a group or an individual based on a given characteristic (also called hate speech type) such as, gender, orientation, ethnicity, nationality, religion, etc. [146].

Based on the definition presented by Alakrot (2019) [4], we consider that the term *offensive language* refers to any text containing cursing, swearing, insulting, profanity, obscenity, rudeness, impoliteness, or hate speech.

Detecting offensive language of online content can be useful, for example, to analyze if a text is racist, sexist, or contains insults, and filter it before recommending content. In

[1] https://dictionary.cambridge.org/dictionary/english/hate-speech

addition, it can also be important to filter offensive content in another activities, as for example in the task of training AI chatbots with Twitter data, avoiding the creation of a chatbot that spreads hate speech [10] or cursing. Even for humans, detecting if a text should be classified as hateful or not is not a trivial task [47]. Automatic methods for online offensive language detection typically combines natural language processing and machine learning, by classifying textual content into offensive and non-offensive. Some approaches also look for specific words with negative connotation such as slurs or insults in the textual content to detect offensive content [157].

In this research, we will provide as a contribution a manually labeled electoral dataset from Twitter that may be used by the community to detect offensive language in this domain, allowing this type of analysis in electoral scenarios. Although the method proposed in this thesis focuses on the sentiment analysis task, we believe that annotate electoral data in the offensive speech dimension can help other interesting researches. For example, one can investigate if a high level of offensive content oriented to one candidate is related to a high level of candidate rejection. Another point related to offensive content that may be explored in future researches is that if it is true that people who support candidates whose posts are full of hateful content against a given minority also propagate posts with similar hateful content.

2.3 Machine Learning

This section presents machine learning basic concepts and an introduction to the transfer learning technique, which is a fundamental topic to this research as we are proposing an approach that takes advantage of existing knowledge to improve analyses related to electoral social media data.

2.3.1 Basic Concepts

The term *Machine learning* refers to methods for designing and developing algorithms that allow systems to learn from data examples or past experience. The basic idea is to infer knowledge from data. The learning step focuses on optimizing a performance criterion. Next, basic concepts of machine learning that are mentioned along of this thesis will be described.

Learning Categories

Learning algorithms are mainly categorized into four groups [25, 52], as follows.

- **Supervised Learning (or predictive)**: algorithms that use labeled training data. In this case, training data contains a list of input/output pairs of the form $\langle x_i, y_i \rangle$, where x_i is the data instance, and y_i is the correct label associated with it. Supervised learning algorithms try to learn the mapping from inputs to outputs. In this way, it is expected to learn a function f that can map input/output pairs seen and that can predict $f(x_i) = y_i$ for all i. When the output is discrete, the function f is called a classifier. On the other hand, if the output is continuous, it is called a regression function. Examples of supervised learning algorithms are: Linear Regression, Nearest Neighbor, Guassian Naive Bayes, Decision Trees, Support Vector Machine (SVM), Random Forest, etc. It is also expected that the classifier/regression function is able to correctly predict the outputs associated with inputs never seen before.

- **Unsupervised Learning (or descriptive)**: algorithms that use unlabeled training data. In this case, only the dataset of input examples is available. However, output data (i.e.: the correct labels) are not presented. Those algorithms try to find the data patterns that can be used to determine the correct output value for new data instances. The assumption of them is that there is a structure to the input space, such that certain patterns occur more often than others, and we want to see what generally happens and what does not. In statistics, this is called density estimation. These algorithms try to use techniques on the input data to detect patterns on data, mine rules, or even to summarize and group data instances. Some of the most popular examples of unsupervised learning are clustering algorithms (e.g.: K-Means) and association rule learning algorithms.

- **Semisupervised Learning**: algorithms that make use of both labeled and unlabeled training data.

- **Reinforcement Learning:** In this case, the algorithms involve agents that explore an environment (context) and try to optimize a decision process toward a goal. Agents are able to learn receiving a different feedback (a reward or a penalty) according to their behavior, for example. Examples of this category include: decision making for stock market investments, optimizing the behaviour of autonomous software agents, game strategy and so on.

In this work we focus only on supervised machine learning algorithms for building classifiers.

Validation

In order to evaluate machine learning classifiers, the original dataset is splitted in, at least, training and test datasets. The idea is that, when there is no intersection between them, the trained model can be evaluated with the results obtained with the test dataset. Machine learning algorithms have the assumption that the data distribution of the training dataset is equal to the data distribution of the test dataset, factor that in practice not necessarily is true. In what follows, the main techniques for validating machine learning classifiers are described.

- **Holdout:** The holdout consists in splitting the datasets into two mutually exclusive groups: (i) training dataset: subset of the original dataset used to learn from examples; (ii) test (or holdout) set – subset of the original dataset used to measure the quality of what was learned. Usually, 2/3 of the data is designed to be part of the training set and 1/3 to be part of the test set. The use of holdout is not advised when a large amount of data is not available. This is because, if the dataset is small, the error calculated in the prediction can vary a lot (the distribution of the test set may not represent the underlying distribution) [74].

- **K-fold cross validation:** The k-fold cross validation is a commonly adopted technique to evaluate how well the machine learning model learned from some training data is going to perform on unseen data. According to James et al. (2013) [64], it can be defined as:

  > "This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds."

  Therefore, the k-fold cross validation is be conducted as follows [74]:

  1. Randomly split the dataset into k equally-sized and mutually exclusive partitions (also called "folds"). Usually k is a small integer, such as 5 or 10;
  2. For each one of the partitions/folds:
     (a) Take it as the test set;
     (b) Fit the model to the rest of the data;

    (c) Evaluate the model prediction on the current test set;

  3. Average the performance across all test sets considered in step (2). It will be the cross-validated estimate of generalization error of the model.

- **Leave-one-out:** Alternatively to the k-fold cross validation, the leave-one-out technique can be adopted. It is a special case where k is equal to n, where n is the number of instances. Thus, models are always evaluated on one instance and trained on all others. A drawback of this approach is that it is computationally expensive, mainly when the dataset contains a big number of instances. Thus, this method is only advised for small datasets.

2.3.2 Transfer Learning

Transfer learning [95] consists in the idea that a problem can be adapted from a given task (e.g.: sentiment analysis or offensive speech detection) or domain (e.g.: product reviews, movie reviews or election opinions) to help to build models to other domains or tasks, by exploiting prior knowledge. In other words, the knowledge acquired when trying to solve one problem could be used to solve similar problems [123]. Figure 2.1 illustrates the general approach of transfer learning, where:

1. There is an existing machine learning model (source model) that was trained to solve a task in a given domain (source domain), represented by the blue rounded rectangle;

2. The aforementioned source model acquired a given knowledge during its training, represented by the blue rectangle;

3. The knowledge acquired by the source model is used as additional input for learning by a target machine learning model (target model) that was designed to solve a task, that can be different or the same, in the same or in other domain (target domain). The target model is represented by a red rounded rectangle.

4. In summary, the target model is trained (or adapted) using data for the target task, represented in Figure 2.1 by the colorful points, and using knowledge acquired by the source model.

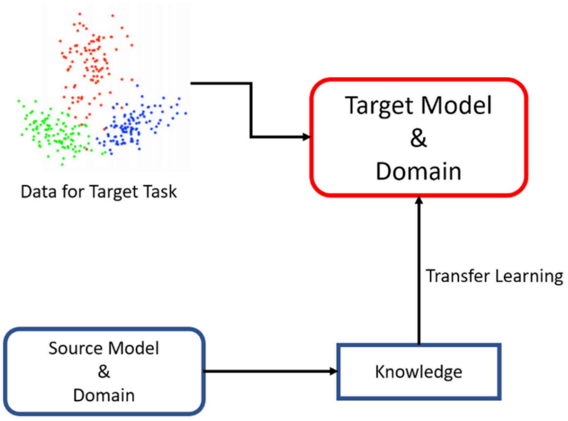


Figure 2.1: Transfer learning technique [123]

According to Rizoiu et al. [110], transfer learning can take advantage, for example, of features and model weights defined for one task/ domain (*source*) to solve another related task/ domain (*target*). Possible advantages of using transfer learning techniques include[2]:

- It can provide an improvement on the performance when modeling the target domain/task [135]. Regarding this advantage, it could be possible to achieve a better baseline performance (higher start) and/or even a better final performance [123];
- It can be useful to allow rapid progress (saving time) [135];

In summary, transfer learning techniques can be classified into three main categories [95]:

- Inductive transfer learning: source and target domains are the same but the task is different;
- Transductive transfer learning: source and target domains are different and tasks are similar or the same;
- Unsupervised transfer learning: both domains and tasks are different.

Figure 2.2 illustrates the transfer learning taxonomy proposed by Ruder (2019) [113]. Since in this research we try to improve predictions in a target domain using knowledge from models trained with different sources but to solve the same task, we are focusing on an instance of the *domain adaptation* case, which assumes that data are available in a source domain.

[2] Notice that one or more of these advantages are possible but not necessarily will occur.

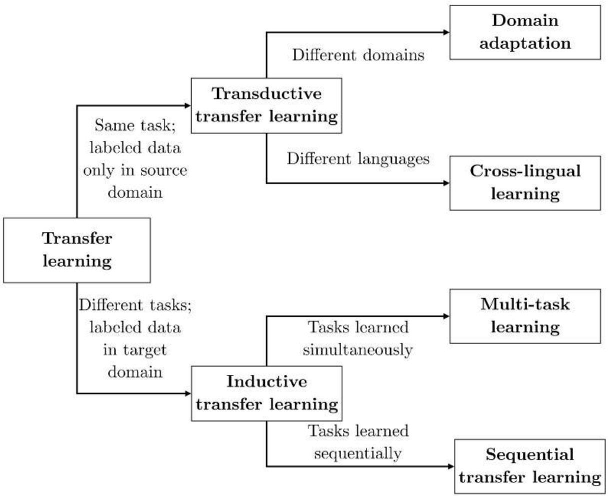

Transfer learning
Same task; labeled data only in source domain
Transductive transfer learning
Different domains
Domain adaptation
Different languages
Cross-lingual learning
Different tasks; labeled data in target domain
Inductive transfer learning
Tasks learned simultaneously
Multi-task learning
Tasks learned sequentially
Sequential transfer learning

Figure 2.2: Transfer learning taxonomy [113]

In some scenarios/settings, transfer knowledge can degrade the original performance of the target problem, resulting in worst predictions (*negative transfer learning*). This can occur, for instance, when knowledge obtained with a too specific dataset is adopted or even when the transfer method is not able to leverage well the relationship between the source and target problems [123]. The adoption of non-related domains (or labeled for non-related tasks) could also imply on undesirable results [100]. In the case of sentiment classifiers, for instance, the textual content involves semantics and the terms that denote sentiment can be different depending on the domain [150]. As an example, while the words "lengthy" and "boring" are usually adopted to express sentiment in a *book* or *movies* domain, both words will probably never appear in the reviews of the *electronics* domain. Furthermore, terms may have different meanings when employed in different domains. For instance, while the term "scary" indicates a negative sentiment when it is employed in most of the domains, it indicates a positive sentiment when employed in the domain of *horror movies* or *horror games*. Therefore, the success of the transfer learning method depends on some issues such as the degree of relatedness between the source and target tasks as well as the source and target domains [106].

2.4 Vectorization methods using Bag of Words

Considering that a word is a symbolic content and machine learning methods can only deal with numbers, vectorization methods are adopted to allow textual documents[3] to be processed. Vectorization techniques consist in transforming the list of documents in

[3] A document is the piece of text being processed.

numerical matrices. One of the first methods adopted for vectorizing texts is called *Bag-of-Words* (BoW) model. In such a model, documents are represented as a bag (multiset) of words. For this model, the order in which the words appear in the document does not matter but the number of times that they appear is very important [54]. It consists of:

- to build a vocabulary (words that appear in the document);
- to assign numerical values to each word with a valuation function.

2.4.1 Counting Word Occurrence

In the most basic method, the valuation function of the Bag-of-Words approach consists in simply counting the number of times each word of the vocabulary appears in each document.

In this way, a matrix of the form $\mathrm{M}_{[\mathrm{D\ x\ N}]}$ is built where D is the number of documents and N is the number of unique vocabulary terms that appear in the whole set of documents. Documents that belong to the corpus are represented as vectors. The value zero is assigned to words that do not appear in the document. On the other hand, when there are occurrences of the word in the document, the value assigned to the word is equal to the number of times it appears in the document.

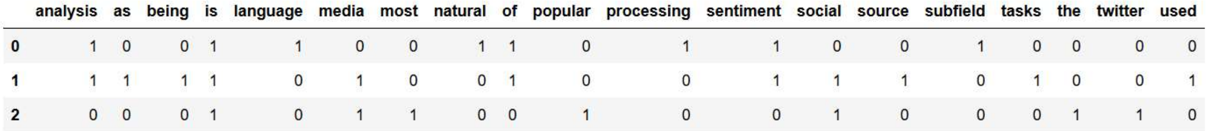
Let us consider as an example a corpus with three documents, as follows:

["Sentiment Analysis is a subfield of Natural Language Processing.",

"Social media is being used as source of Sentiment Analysis Tasks.",

"Twitter is the most popular social media."]

The respective Bag-of-Words matrix using counting word occurrence will be the one illustrated in Figure 2.3. The header of the matrix exhibits the vocabulary and each row a document (sentence).



	analysis	as	being	is	language	media	most	natural	of	popular	processing	sentiment	social	source	subfield	tasks	the	twitter	used
0	1	0	0	1	1	0	0	1	1	0	1	1	0	0	1	0	0	0	0
1	1	1	1	1	0	1	0	0	1	0	0	1	1	1	0	1	0	0	1
2	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	1	1	0

Figure 2.3: Bag of Words example using counting word occurrence.

The most simple approach of BoW uses *unigrams*, which means that it considers words one by one (each word is a column of the matrix, as in previous example in Figure 2.3).

Alternatively, groups of words can be taken in order to preserve a part of the sequence in which words appears in a given sentence, as occurs in the *bigrams* (where each column is composed of two adjacent words), *trigrams* (where each column is composed of three adjacent words), and so on.

2.4.2 Term Frequency-Inverse Document Frequency

The Term Frequency-Inverse Document Frequency (TF-IDF) is a method that can be used as the valuation function to assign values to the words represented with the Bag-of-Words approach. It computes values for each word that appears in a document based on the inverse proportion of the frequency that a word in a given document to the percentage of documents that this word appears in [107]. Therefore, the importance of a word increases proportionally to the number of times it appears in the document but is offset by the frequency of the word in the corpus. In this way, the TF-IDF is composed of two terms, namely, the Term Frequency (TF) – it corresponds to the number of times a word appears in a given document divided by the number of words in such a document; and the Inverse Document Frequency (IDF) – it corresponds to the logarithm of the number of the documents that belong to the corpus divided by the number of documents in which the given word appears. The TF-IDF formula is composed by the Equations (2.1) and (2.2). It can be used with *n-grams* with different size (unigram, bigram, trigram, etc.).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{2.1}$$

$$idf(w) = log(\frac{N}{df_t}) \tag{2.2}$$

2.5 Vectorization methods using Word Embeddings

Word embeddings [89] is a technique to map words from a corpus to n-dimensional dense vectors of real numbers, which have low dimensionality (when compared to traditional techniques such as TF-IDF Bag-of-Words). Therefore, in this technique every word in the vocabulary has its own numerical vector that can be used as input of machine learning algorithms.

This method has the ability to capture distributional semantics. For this reason, a subtraction between the numerical vectors assigned to the words *king* and *man*, followed by an addition to the *woman* vector, for instance, results in a numerical vector very close to the one that represents the word *queen* [5]. Similarly, vectors that refer to colors such as the ones of the words “green” and “blue” are located nearby. The same occurs for synonyms words (e.g.: “good” and “great”) or even words that denote concepts that are related in some way (e.g: “Paris” and “France”).

Word2Vec [89] and GloVe[4] [99] are popular examples of algorithms used to generate word embeddings, which will be described as follows.

Word2Vec: The Word2Vec [89] is a predictive-based method that uses a neural network for training words against other ones that neighbor them in the corpus to predict words. It relies on the hypothesis that words that appear in similar locations have similar meanings. This is an self-supervised method since it does not require labeled data. Words are represented in a continuous vector space preserving linear regularities, as for example, differences in syntax and semantics. Two algorithms have been proposed:

- Continuous Bag of Words (CBoW): it uses the context to predict the target word (as illustrated in Figure 2.4). Words that surround the target word are used as input (i.e., next or previous are analyzed according to the window size) and the target word is predicted using a softmax;

- Skip-Gram: it uses a word to predict the context (the words that surround it), as illustrated in Figure 2.5.

[4] https://nlp.stanford.edu/projects/glove/

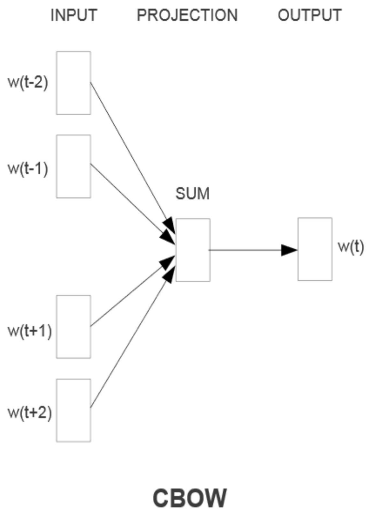


Figure 2.4: The Continuous Bag of Words (CBoW) model [89]

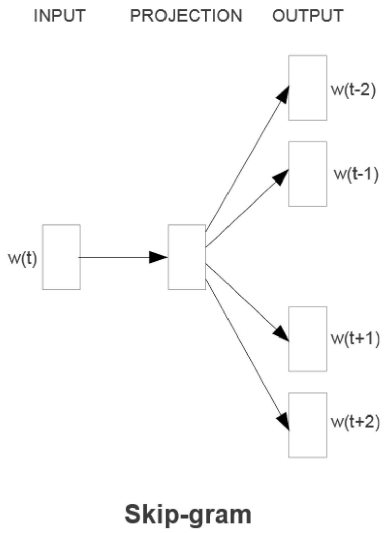


Figure 2.5: The Skip-Gram model [89]

GloVe: The Global Vectors from Word Representation (GloVe) [99] is a count-based method that leverages local and global co-occurrence statistics to learn word representations. Pennington et al. (2014) define it as a new global log-bilinear regression model that combines the advantages of the two major model families: global matrix factorization (e.g.: LSA method) and local context window methods (e.g.: skip-gram method).

Statistics about word co-occurrence are stored in a $V \times V$ matrix X, where V is the number of words in the corpus and each element $X_{i,j}$ represents how many times the word i co-occurred with j (i.e., they appear together within a fixed window). This method relies on the idea that co-occurrence ratios between two words in a given context are strongly related to meaning.

Both Word2Vec and GloVe do not allow polysemy, ignoring that the same word can have different meaning depending on the context. In this way, the same vector representation is provided to a word regardless of the context.

2.6 Models that allow transference

Language models can be used to deal with the next word prediction task, where the next word of a sentence is predicted given a sequence of past words. More formally, given a context, a language model predicts the probability of a word occurring in that context $P(W_i|W_1...W_{i-1})$. For masked language models, a certain % of words of a sentence is masked and the model is expected to predict them based on the other ones. Usually, it is an unsupervised learning problem because it only requires access to the raw text [114]. The model can be designed to address the word-level or the character level. The related practical applications include: email response suggestion, intelligent keyboards, spelling autocorrection, etc. According to Peters et al. (2018) [102], pretrained language models can be used for several NLP tasks. Recently, they are gaining importance given the possibility of knowledge transference. Howard and Ruder (2018) [59] point out that language modelling can be viewed as a counterpart of ImageNet for NLP. This is because they can be used to capture many characteristics of language that can be relevant for downstream tasks such as long-term dependencies and hierarchical relations. The general idea can be summarized as follows:

1. Train a model in the task that lead to word/sentence representations;
2. Release the pre-trained model;
3. Fine-tuning the pre-trained model on a target task.

Examples of some of the most popular models are briefly described in what follows.

ELMo [102]: Embeddings from Language Models (ELMo) uses the concatenation of independently trained left-to-right and right-to-left LSTM [58] to generate features for downstream tasks. Since ELMo representations are based on characters, the model can understand out-of-vocabulary tokens unseen during training. Transfer knowledge by using ELMo include the following steps: (i) train a Bi-directional Language Model in a very large corpus; (ii) freeze the encoders and put them in the lowest level in the model; (iii) replace the words with the associated word vectors; (iv) apply the encoders to the words and sum the hidden representation vector with the word vectors.

ULMFit [59]: ULMFit stands for Universal Language Model Fine-tuning. According to the authors, it can be viewed as a transfer learning method that uses key techniques for fine-tuning a language model and that can be applied to NLP tasks. It consists in three stages, which are as follows: (i) Train an AWD-LSTM language model[5], forward or backward or both, in a large general corpus to capture general features of the language in different layers; (ii) Fine-tune the model in the task dataset to learn task-specific features; (iii) Fine-tune the model on the target task using gradual unfreezing. In this way, low-level representations are preserved and high-level ones are adapted.

[5] AWD-LSTM (ASGD Weight-Dropped LSTM) is one of the most popular language models.

BERT [35]: The Bidirectional Encoder Representations from Transformers (BERT) uses an architecture based on a bidirectional Transformer, which encodes the important words (and dependencies between them and other words). It is a project released by the Google Research team. BERT representations are jointly conditioned on both left and right context in all layers. In other words, it is deeply bidirectional. To transfer knowledge for classification tasks by using BERT, one should: (i) train BERT with a large corpus; (ii) fine-tune the model to a task-specific dataset; (iii) and add one additional output layer (a softmax layer) at the top of the model for classification.

USEM [153]: The Universal Sentence Encoder Multilingual (USEM) was released by the Google Research team and is a pretrained model that was trained on 16 languages and is able to embed text from these different languages in a single vector space. To that, a multi-task trained dual encoder that learns tied representations using translation based bridge tasks [29] is adopted. The supported languages are as follows: arabic, chinese, chinese (Taiwan), dutch, english, german, french, italian, portuguese, spanish, japanese, korean, russian, polish, thai and turkish. Models are implemented in TensorFlow and are publicly available on TensorFlow Hub[6]. The multilingual embeddings provided by this model allow us to compare and use datasets that belong to different languages.

2.7 Distance Metrics

In this research, we investigate the usage of some distance metrics to measure similarity between textual datasets. Next, the distance metrics mentioned in this research will be introduced.

Jaccard Distance (d_J)

Jaccard coefficient is a statistic function used to compare different sets according to their similarity. It is calculated as the size of the intersection divided by the size of the union of the sample sets. When the two sets being compared are empty, the Jaccard coefficient is equal to 1. Analogously, the Jaccard distance, defined by Equation (2.3), measures the dissimilarity between sets and is computed as the complement of the Jaccard coefficient.

[6] https://www.tensorflow.org

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{2.3}$$

Cosine Distance ($\boldsymbol{d_{Cos}}$)

Given two vectors $\mathbf{x}$ and $\mathbf{y}$, the cosine similarity measures the cosine of the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$. Cosine distance is the complement of the cosine similarity, defined by Equation (2.4).

$$s_{Cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\sqrt{\mathbf{x} \cdot \mathbf{x}}\sqrt{\mathbf{y} \cdot \mathbf{y}}} \tag{2.4}$$

Euclidean Distance ($\boldsymbol{d_E}$)

The Euclidean Distance is the distance between two points in the Euclidean space. In general, in the n-dimensional space the Euclidean distance is given by Equation (2.5).

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.5}$$

2.8 Final Considerations

In this chapter, we presented the main study topics explored in this research, namely, opinion mining, sentiment analysis. In addition, we also introduced basic concepts of offensive language, machine learning and vectorization methods that are mentioned along of this proposal. Although there are some alternatives for transfer learning involving text, including the use of language models that allow transference introduced in Section 2.6, in this thesis we are only investigating transfer learning based on dataset similarity, which is measured according to the distance metrics, as the ones presented in Section 2.7, and multilingual embeddings from pretrained models as the Universal Sentence Encoder Multilingual (USEM). Our premise is that there are no labels for the target dataset and therefore it would not be possible to perform a fine tuning process. These choices are better explained in Chapter 4. In the next chapter, we discuss related work, exploring approaches related to election forecast using social media and computational techniques, and approaches that deal with classification tasks and lack of labeled data in the target domain.

Chapter 3

Related Work

This chapter is divided into two sections. First of all, a systematic literature review was conducted to better understand domain-related approaches that propose a means for forecasting elections based on social media, as described in Section 3.1. After noticing that the *lack of labeled data* in the electoral domain is a critical problem for election data analysis, we decided to conduct an ad hoc search to find researches that propose strategies for dealing with unlabeled data in the target domain and need to build machine learning classifiers, as described in Section 3.2. One example of that are the strategies for selecting datasets to be used as starting point for transfer learning tasks.

3.1 Approaches for forecasting elections based on social media

We have conducted a systematic literature review about approaches that forecast elections based on social media data to better understand the state-of-the-art of this domain problem. We focused on the following research questions:

- **Q1**: What are the main approaches for predicting election outcomes by using social media?
- **Q2**: What are the main data science limitations of the approaches that collect social media data in order to predict elections?
- **Q3**: What are the possible lines for future research on election prediction using social media from the AI point of view?

The full content of the systematic review is available in [120]. We conducted our search in the IEEE, ACM, Scopus and Science Direct digital libraries using the following search string:

```
(("election prediction"OR "election forecast") AND
("social media"OR "Twitter")).
```

This search string was executed in August 2020 and returned a total of 242 works. We filtered the papers published from 2014, resulting in 207 works to be analyzed. When analyzing the abstracts, we considered the following inclusion criteria: *(I1)* Papers that propose methods to predict *election outcomes* using *social media* and *(I2)* Papers that apply existing data and opinion methods for election outcomes prediction based on social media data. Our exclusion criteria were:

- *(E1)* papers predicting election outcomes not using social media posts. For instance, [80] presented a method to predict elections outcomes considering the results of previous elections and questionnaires; and [50] that predicted election outcomes of the 2016 Brazilian municipal elections relying on comments extracted from news websites instead of social media posts;
- *(E2)* works analyzing some aspects of electoral data extracted from social media but do not predict election outcomes. For instance, [130] analyzed data from Twitter to find out key topics and influencers for the left and right wings for the 2017 French presidential elections; [121] investigated the usage of sentiment analysis for datasets from several domains to predict people sentiment towards the 2018 Brazilian presidential elections; [63] adopted a bayesian network to predict the voting behavior of a given Facebook user in relation to the US 2016 presidential elections based on his Facebook profile; and [36] analyzed tweets posted by Italian deputies to discover the most mentioned topics by political alignment. All of them analyze aspects related to elections but they do not try to predict election outcomes;
- *(E3)* Papers not written in English.

Finally, after the process of filtering, removing the duplicated and unrelated papers and applying the inclusion and exclusion criteria on the 207 papers, we ended up with 53 works. We observed that some of these works use similar strategies for forecasting elections. We categorize the majority of the works into the following approaches for elections outcomes predictions, namely: *Counting Based Approach*, *Political Alignment Approach*, *Event Detection Approach* and *Popularity Based Approach.* The works that do not fit in these categories are grouped into *Other Works* category.

3.1.1 Data Collection

This section refers to information about how data were collected from social media, such as, data sources, quantity of data collected, keywords used for gathering data and collection period. Election data to be analyzed are collected from a given data source and usually are collected based on a given time period or based on keywords/search terms. We observed that the collection period was variable ranging from less than a month to more than six months. In relation to the search keywords, we identified that most of the papers use keyword of the following categories: (i) *candidate related*: terms or hashtags including candidate name or last name; (ii) *party names*: term or hashtags that refer to party names; and (iii) election keywords: terms or hashtags containing campaign slogans, for example. A summary about the opinion sources is presented as a Venn diagram in Figure 3.1. Concerning data sources, we notice that Twitter stands out among social networks for gathering political opinions in order to forecast election, *i.e.,* considering all the 53 papers (100%) analyzed in this research, 44 papers (83,01%) use only Twitter as source of social media election opinions, 1 (1,89%) uses Facebook as the only source of opinions, 1 (1,89%) combines data from Facebook, Twitter and websites (candidates webpages and Google), 1 (1,89%) combines data from Facebook, Twitter, Instagram, traditional polls and past elections, 1 (1,89%) combines data from Facebook and websites (e-news and magazines), and 1 (1,89%) uses data from Twitter and websites (blogs). Finally, 4 papers (7,55%) use exclusively other sources for mining opinions that are not adopted by other papers such as the Flickr, Reddit, the BSS, and a Taiwan forum. We have chosen not to illustrate the latter case in the diagram because these sources are not mentioned by more than one paper.

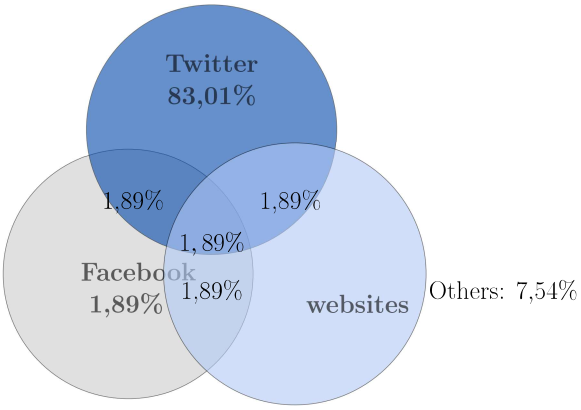


Figure 3.1: Opinion sources

Table 3.1 refers to the amount of data collected. The column *paper* is the reference to the paper in which the approach was detailed; the column *number of posts (x)* refers to the number of data instances collected to be analyzed (e.g: tweets, Facebook likes or comments). The ranges of data collection were organized as follows:

- $\mathbf{x} \leq 100\,000$: papers that collected up to 100 000 data instances;
- $100\,000 < \mathbf{x} \leq 500\,000$: papers that collected between 100 000 and 500 000 data instances;
- $500\,000 < \mathbf{x} < 1\,000\,000$: papers that collected more than 500 000 and less than 1 million data instances;
- $\mathbf{x} \geq 1\,000\,000$: papers that collected more than 1 million data instances.

Papers that do not explicitly inform how much instances were collected are grouped into the *not informed* field (see Table 3.1). Figure 3.2 illustrates this information using a bar chart where we can see that most works collect more than one million data. Table 3.2 exhibits a summary about the number of successful approaches according to the amount of data, where each row represents the number of papers of a given amount of data range that are associated with each one of the following possibilities: *success* – the paper predicted correctly the election winner in all their experiments; *partial* – the paper achieved success in predicting the election winner in at least one of their experiments but not in all experiments; *no* – the paper failed to predict the election winner; *N/A* – the paper does not present enough information about the success of their approach.

Table 3.1: Quantity of data instances collected

number of posts (x)	**paper**
$\mathbf{x} \leq 100\,000$	[108], [127], [87], [66], [70], [57], [128], [77], [16], [129], [60]
$100\,000 < \mathbf{x} \leq 500\,000$	[71], [152], [136], [104], [67][1], [85], [117], [11], [12]
$500\,000 < \mathbf{x} < 1\,000\,000$	[28], [139], [8]
$\mathbf{x} \geq 1\,000\,000$	[132], [23], [125], [73], [37], [6], [111] [62], [55], [43], [39], [26], [144] [140], [69], [137], [56], [147], [13], [21]
not informed	[2], [142], [154], [148], [143], [90], [9], [118], [22], [61]

[1] The quantity of tweets collected was not explicitly presented in this paper [67]. However it was inferred based on the number of tweets collected per day.

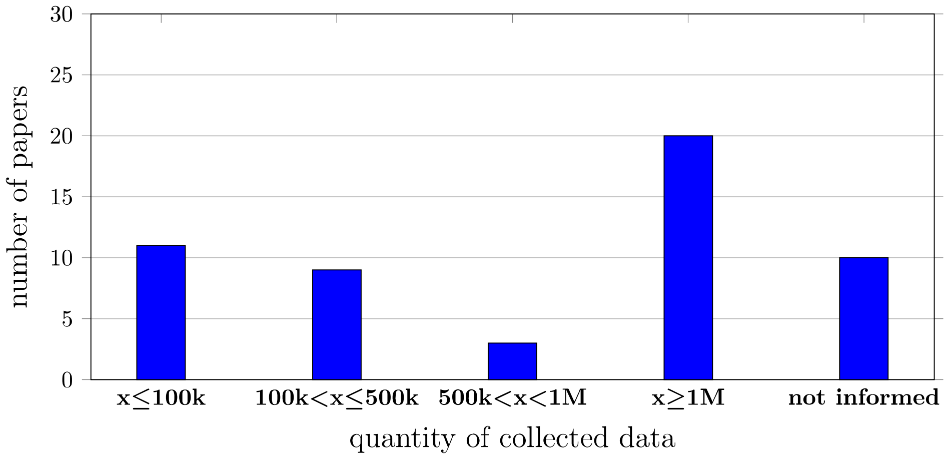


Figure 3.2: Number of papers by quantity of collected data

Table 3.2: Number of successful approaches according to the amount of collected data

	success	partial	no	N/A
x≤100k	6	1	2	2
100k<x≤500k	3	3	1	2
500k<x<1M	2	1	0	0
x≥1M	9	5	2	4

Although it may seem that collecting more data leads to better results, we cannot draw this conclusion, as the amount of works that collected more than 1 million posts is also much greater than the amount of works in the other data collection ranges.

Table 3.3 presents information about the period in which the data was collected. The column *paper* refers to the paper in which the approach is described; the column *period collection* refers to the period (number of months (x)) that each approach considered to collect data. The ranges of the period collection were organized as follows:

- **x ≤ 1 month**: papers that collected data in a period up to 1 month;
- **1 < x ≤ 3 months**: papers that collected data in a period between 1 and 3 months;
- **3 < x < 6 months**: papers whose period of data collection was between 3 and 6 months.
- **x ≥ 6 months**: papers that collected data in a period bigger than 6 months;

Papers that do not explicitly inform what was the period of data collection are grouped into the *not informed* (see Table 3.3). Figure 3.3 illustrates an overview about the collection period using a bar chart where we can see that most works adopt a period between 1 and 3 months.

Table 3.3: Period collection

period collection (x)	paper
$x \leq 1$ month	[108], [148], [127], [66], [132], [140], [111], [70], [43], [104], [26], [128], [129], [61], [11], [117], [12], [8], [118]
$1 < x \leq 3$ months	[28], [55], [6], [62], [73], [136], [39], [67], [87] [56], [139], [147], [13], [125], [144], [16]
$3 < x < 6$ months	[152], [23], [136], [137], [71], [22]
$x \geq 6$ months	[37], [85], [57], [21], [60]
not informed	[2], [154], [69], [143], [77], [9], [90]

We have observed that the collection period does not necessarily implies on a higher amount of data (as is the case of the papers [140], [132], [111], [43] in Tables 3.1 and 3.3, for example). Therefore, the amount of data also depends of the hashtags used for data collection.

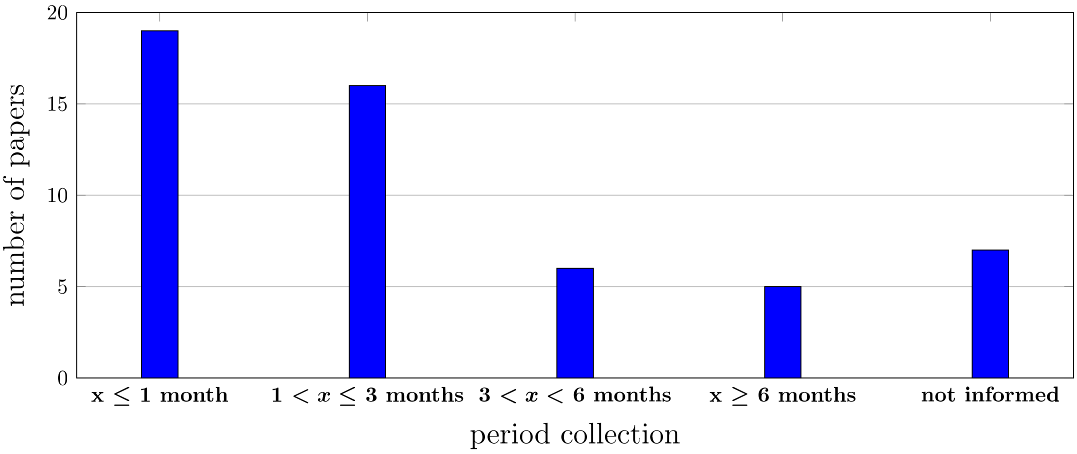


Figure 3.3: Number of papers by period collection

Table 3.4 exhibits a summary about the number of successful approaches according to the data collection period, where each row represents the number of papers of a given data collection period that are associated with each one of the following possibilities: *success* – the paper predicted correctly the election winner in all their experiments; *partial* – the paper achieved success in predicting the election winner in at least one of their experiments but not in all experiments; *no* – the paper failed to predict the election winner; *N/A* – the paper does not present enough information about the success of their approach. While it might seem that collecting data for a shorter time results in better predictions, we cannot draw this conclusion as most works have adopted a short data collection time.

Table 3.4: Number of successful approaches according to the data collection period

	success	**partial**	**no**	**N/A**
x $\leq$ 1 month	9	5	1	4
$1 < x \leq 3$ months	8	4	1	3
$3 < x < 6$ months	3	1	1	1
x $\geq$ 6 months	2	1	1	1

The keywords/terms used to collect data are summarized in Table 3.5. We observed that keywords related to candidates such as those that use parts of the candidate's first or last name and keywords related to election terms such as the ones that contain campaign slogans or combinations mentioning the name of the elections and the election year are the most popular types of keywords.

Table 3.5: Types of keywords used by the surveyed papers

keywords	**paper**
candidate related	[55], [66], [39], [23], [148], [73], [56], [87], [62], [111], [125], [154], [142], [37], [71], [147], [136], [104], [67], [143], [128], [77], [16], [57], [129], [11], [12], [8]
party names	[127], [23], [148], [73], [87], [37], [147], [136], [67], [128], [16], [129], [11], [12], [8], [118]
election keywords	[148], [73], [56], [137], [139], [125], [37], [69], [43], [70], [26], [143], [77], [90], [57], [11], [12], [22], [60]
not informed/ do not use keywords	[108], [132], [152], [144], [2], [28], [6], [13], [140], [117], [9], [118], [21], [61]

3.1.2 Data Preprocessing

As pointed out by [84], social media data are very noisy since they include different kinds of spelling, punctuation and grammatical errors. For this reason, before data analysis, it is important to conduct a preprocessing phase to clean data and remove noise. Table 3.6 exhibits a summary of the most used preprocessing techniques that were adopted after the data collection phase. In Table 3.6, the term *word extension* refers to words with duplicated letters such as *Loooove* instead of *Love*.

Table 3.6: Preprocessing steps

Preprocessing step	**papers**
lower case conversion	[108], [104], [128], [90], [11], [12]
hashtags removal	[108], [127], [104], [26], [90], [129]
stop words removal	[108], [127], [148], [62], [28], [137], [139], [104], [26], [67], [128], [90], [11], [12], [22]
URL removal	[108], [127], [148], [73], [62], [139], [104], [128], [90], [16], [129], [22]
emoticons removal	[55], [127], [28], [16]
duplicated removal	[55], [73], [12], [22]
retweets removal	[55], [127]
special characters removal	[127], [148], [67], [22], [77], [11]
punctuation removal	[127], [62], [28], [137], [104], [67], [128], [77], [90], [16], [11], [12]
username mentions removal	[108], [127], [148], [62], [139], [104], [90]
negation handling	[127], [66], [139]
ambiguous keywords removal	[136]
bots and non-personal accounts removal	[6], [62], [28], [39]
stemming	[28], [139], [11]
retweet keyword (RT)	[139], [62], [90], [16]
slang and word extension replacement	[139], [90]
translation	[128], [104]
not informed	[66], [39], [23], [132], [87], [56], [140], [111], [13], [125], [154], [142], [152], [37], [69], [144], [71], [2], [43], [147], [143], [57], [117], [9], [118], [21], [8], [60], [61]

We have observed that the removal of user mentions, URLs, punctuation, and stop words are the most popular preprocessing steps. A minimal number of works discard duplicated content or try to detect and discard content posted by bot accounts (spam). Most of the works did not inform the steps conducted for data preprocessing. A few works translate opinions during preprocessing step. Techniques to filter bots and non-personal accounts are little explored.

3.1.3 Data Labeling

The methods that were adopted by the papers for labeling data are presented in Table 3.7. Figure 3.4 summarizes the information about the labeling methods using a Venn diagram. From the 38 papers that *use methods for predicting sentence sentiment*, three of them ([108], [128], [8]) do not inform the method that was adopted to assign polarities to the sentences and for this reason were not considered in this analysis. From the remaining 35 papers (100%), the total of 17 papers (48,57%) rely only on lexicon dictionaries to determine the sentiment of a sentence. Methods that rely only on emoticon or hashtags that denote positivity/negativity are also trendy (22,86%) (8 papers). Only three papers (8,57%) are based on manually labeling a subset of the documents (semi-supervised approach) and 5,71% (*i.e.*: 2 papers) of the works combine all of these three methods (lexicon, emoticon/hashtag and manually annotated). One paper (2,86%) uses emoticons and lexicons to determine sentence sentiment. A total of 11,43% of the works (4 papers) use other methods that are adopted by only one approach, and, for this reason, we choose not to illustrate them in the diagram (which exhibits information about the major three methods). As an example of this case is the SAS software cited by [87] and the Aylien API adopted by [61], whose underlying methods to assign polarities are not explained.

Table 3.7: Data labeling methods

method	**papers**
emoticons/ hashtags based	[55], [39], [132], [148], [6], [56], [154], [70], [26], [143], [77], [12]
manually labeled	[127], [132], [6], [90], [118]
lexicon based	[136], [66], [23], [132], [73], [6], [62], [142], [144], [43], [147], [139], [125], [104], [67], [129], [11], [12], [22], [60]
others	[87], [69], [16], [61]
not informed	[108], [128], [8]

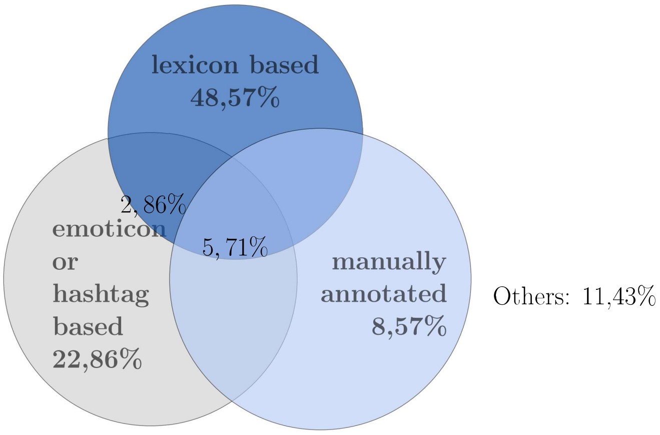


Figure 3.4: Labeling methods

The exploratory study about the sentiment analysis process of the 2018 Brazilian presidential elections [119] suggests that generic sentiment labeling methods may not be enough to capture the real sentiment of electoral tweets and points out that the use of automatic labeling strategies can be a threat to obtain reliable electoral analyzes based on social media. This work compares labels obtained with Microsoft Azure Sentiment Analysis API with labels obtained from manual labeling based on crowdsourcing with the majority voting strategy. Such study showed that the overall sentiment (positive, negative, neutral) of the sample of tweets obtained with the automatic labeling strategy was different from the overall sentiment calculated using the manual labeling strategy.

3.1.4 Demographic Information

Table 3.8 summarizes aspects related to the user profile and location. The column *user* is checked when the approach considers information related to the user profile in its analysis (e.g.: sex, age, education, income, etc). The column *location* is checked when the approach tries to detect the location of the post or of the user, as for example: the work in [139] – that uses POS-tag information to identify the location; the approach in [147] – that uses the Twitter geolocation tag; the study described in [56] – that searches for the location information in the user profile; or even the approach in [127] – that assumes that tweets written in Hindi language belong to Hindi users, and so on. The symbol "-" indicates that such approach does not consider user characteristics/ location information in its analysis. Figure 3.5 presents this information in a visual way. We have observed that 62,26% of the works (33 papers) do not consider location and user characteristics info, 7,55% of them (4 papers) consider only user characteristics, and 22,64% (12 papers) consider only location info[2]. A total of 7,55% (4 papers) of the works consider both location and user characteristics. The work in [117] and [6] infer automatically age and gender of the users based on the history of posts of the user. In addition, [6] also uses a name dictionary to infer user gender and a classifier to predict user social class. The work in [11] and [12] only applied geotagging during data collection when using keywords that are not exclusive to the given election.

Table 3.8: Demographic information

paper	**user**	**location**
[108], [55], [66], [23], [132], [87], [137], [125], [154], [142], [70], [43], [69], [144], [71], [2], [136]	-	-
[127], [148], [73], [56], [111], [13], [139], [37], [85], [11], [57], [12]	-	✓
[62], [140], [152], [117]	✓	
[39], [6], [28], [147]	✓	✓

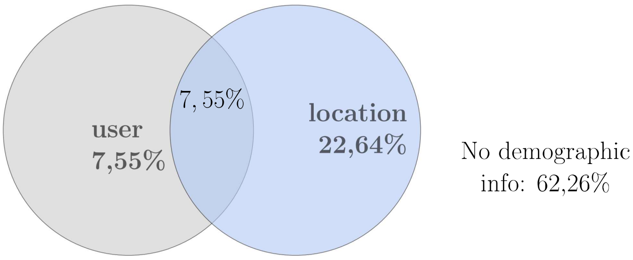


Figure 3.5: Demographic information

3.1.5 Machine Learning methods

Table 3.9 is related to the machine learning methods that were applied in each election forecast approach. The column *paper* refers to the paper in which the approach was described and the column *machine learning method* refers to the machine learning algorithm

[2] From this set of papers, [11] and [12] use location when keywords are not exclusive to elections.

used. The column *success rate* indicates the percentage of success of each algorithm considering all works that had success when using this method and the total of works that use it, disregarding works that do not explicitly inform that they predicted the winner of the elections correctly (N/A). In the case where the algorithm is only associated with works that did not achieve success when adopting such algorithm, the success rate is 0%. Also, the symbol "-"is used to indicate cases where the algorithm is only associated with works whose success is not informed (N/A). Papers that did not adopt machine learning methods are not mentioned in this table. The work in [117] adopted a software to automatically users demographic characteristics (age and gender) but it is not clear if such software uses machine learning methods.

Table 3.9: Machine learning methods

Paper	**Machine learning method**	**Success rate**
[132], [127], [28], [6] [69], [104], [77]	Support Vector Machine	57,4%
[127], [39], [104], [128]	Naive-Bayes	75%
[148]	Binary Multinomial Naive Bayes	100%
[73]	OLS Regression	100%
[66]	Hidden Markov Model	-
[154]	Adaboost, CVAR	100%
[6]	Random Forest	0%
[6], [21]	Multilayer Perceptron	50%
[6], [26]	Multinominal Naive-Bayes	0%
[128], [136], [21]	Linear Regression	100%
[136]	Sequential Minimal Optimization for regression	-
[55], [125], [56]	Convolutional Neural Networks (CNN)	50%
[16], [85]	Recurrent Neural Networks (RNN)	50%
[70]	DynamicLMC	-
[71], [136]	Gaussian Process Regression model	0%
[104]	Logistic Regression	100%
[104], [67], [128], [9]	Decision Tree	33,33%
[128]	K-Nearest Neighbors	100%
[142], [62]	not informed	-
[8]	Bayesian optimization model	0%
[11]	Biterm Topic Model (BTM)	100%
[137], [11], [28], [43]	Latent Dirichlet Allocation (LDA)	75%

Classification tasks: Several classification tasks were adopted in the surveyed approaches such as the detection of buzzer/ spammer accounts, demographic info classification (gender, social class, age), political alignment classification, and sentiment analysis. We observed that SVM, Naive-Bayes, and Decision Trees are the most common machine learning methods used to address this task. The works in [108] and [87] do not appear in Table 3.9 because it is not clear in these papers if the sentiment analysis is performed using machine learning.

Regarding Deep Learning, we observed that works using this kind of models are becoming more popular in recent years, *i.e.*, since 2017. The Deep Learning methods adopted in the surveyed papers are as follows: Convolutional Neural Networks (CNN) – adopted by papers published in 2017, namely, [55] and [125], and [56], which adopted an AlexNet model, specifically, and was published in 2018; Recurrent Neural Networks (RNN) – adopted by papers published in 2018, namely, [16] and [85], which adopted a RNN-LSTM model, specifically. The paper [13], which was published in 2018, used Deep Learning but did not inform the specific algorithm adopted. If we had created a specific category for Deep Learning, the success rate would be 40%.

Topic Modeling: We observed that the LDA was the most popular algorithm to address topic modeling, as it was adopted in [11, 28, 43, 137]. The BTM model was also adopted in [11] to address this task.

3.1.6 Approaches for Predicting Election Outcomes

Figure 3.6 presents a pie chart that illustrates the percentage of papers that belong to each approach. Although we have identified different approaches for handling the election prediction task such as analyzing candidate popularity, detecting events that are important for the course of campaigns, and analyzing user political alignment, we can clearly notice that the approach based on counting instances (considering sentiment or not) are still the most adopted strategy for forecasting election outcomes in the literature.

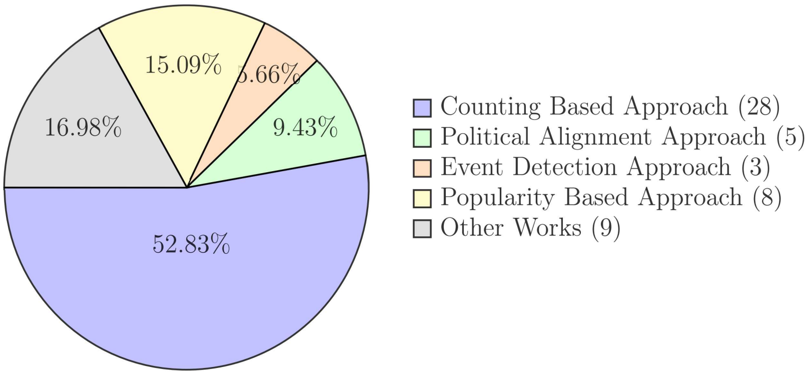


Figure 3.6: Papers by approach

3.1.7 Approaches Summary

Table 3.10 presents the general characteristics of the papers. It is organized as follows: the column *ap.* refers to the approach used by the paper to forecast election outcomes. We use numbers to distinguish the different approaches, as follows:

- 1 – Counting Based Approach;
- 2 – Political Alignment Approach;
- 3 – Event Detection Approach;
- 4 – Popularity Based Approach;
- 5 – Other Works.

The column *paper* indicates the paper in which the approach was presented; the column *vol.* informs if the approach uses at least one method based only on post counting (also called volume-based); the column *sent.* is checked when the approach presents at least one method based on sentiment analysis; the column *source* refers to the data source from where opinions were collected (e.g.: Twitter or Facebook); the column *alig.* informs if the paper predicts the political alignment of tweets/users to forecast election outcomes; the column *ev.* refers to the ones that use strategies related to event detection; the column *pop.* is checked when the paper proposes a means to calculate the candidate popularity; Finally, the column *success* indicates if the authors of the paper state that their approach predicted the election winner correctly. This field can be filled with four different labels:

- *partial* – the authors state that their methods achieve success in at least one of their experiments but not in all experiments;
- ✓ – the authors state that their results predicted the election winner in all their experiments;
- *N/A* – the authors did not explicitly inform if their approach predicted the election winner;
- *no* – the authors state that their approach has failed to predict the election winner correctly.

Cases of partial success are summarized as follows:

- *Papers that present more than one election prediction method*: this is the case of works in [55], [11], and [61], which belong to the Counting Based Approach and achieved success for the sentiment-based method and failed when using the volume-based method; the work presented in [2], which belongs to the Other Works section and achieved success when analyzed data from blogs and failed when analyzed data from Twitter;

- *Papers that only achieved success for some cities, states, districts or seats*: this is the case of the following Counting Based approaches: [6], [104], and [56]; the Event Detection Approach described in [137]; the Political Alignment Approach presented in [85]; and the following works presented in the Other Works section: [147] and [8].

- *Papers with more that one sentiment analysis method*: this is the case of the Counting Based Approach [127] that failed when adopted dictionaries to predict sentiment and achieved success when adopted machine learning methods to infer sentiment.

We have observed that although several approaches state that they achieved success, they do not consider real percentages as, most of them, assume that the candidate associated with the higher number of (positive) instances is the election winner. Also, while some approaches compare their results with the real outcomes, other ones compare their results only with traditional polls. Another point that we can see from Table 3.9 and Table 3.10, is that although some papers adopted more recent algorithms such as the ones based on some Deep Learning strategy, they failed to predict the winner or only achieved a partial success. This is the case of approaches [55], [56] and [13], for example. On the other hand, we observed that approaches that adopted traditional machine learning algorithms were able to achieve a correct prediction in many cases, as is the case of [132], [127], [28], [6], [127], [39], and [148], for example. The amount of collected data is not directly related to success. We can see this by looking at Table 3.10 and Table 3.1, where the approaches in [13] and [111] failed to predict the election winner even though they were in the group that collected more tweets (more than 1 000 000 tweets). In the opposite side, papers in the group that collected fewer tweets (less than 100 000 instances) were able in some cases to achieve success, namely [108] and [87]. Another point that we can notice looking at Table 3.10, is that most of the works that achieved success predicted the opinions sentiment in their analysis, independently of the approach adopted. For instance, event though [87], [28], [139], [144], [43] adopted different approaches (Counting Based Approach, Political Alignment Approach, Event Detection Approach, Popularity Based Approach, and Other Works, respectively), all of them combine their approaches with sentiment analysis. In addition, we also can observe that most of approaches that succeed (22 out of 25) use Twitter as source of opinions.

Table 3.10: General characteristics of the surveyed papers

ap.	paper	vol.	sent.	alig.	ev.	pop.	source	success
1	[87]		✓				Twitter	✓
1	[6]		✓				Twitter	partial
1	[39]		✓				Twitter	✓
1	[62]		✓				Twitter	✓
1	[73]		✓				Twitter	✓
1	[132]		✓				Twitter	✓
1	[23]		✓				Twitter	✓
1	[66]		✓				Twitter	N/A
1	[127]		✓				Twitter	partial
1	[148]		✓				Twitter	✓
1	[117]	✓					Twitter	✓
1	[140]	✓					Facebook	N/A
1	[55]	✓	✓				Twitter	partial
1	[111]	✓					Twitter	no
1	[108]		✓				Twitter	✓
1	[57]	✓					Twitter	no
1	[129]		✓				Twitter	✓
1	[104]	✓	✓				Twitter	partial
1	[56]	✓	✓				Twitter	partial
1	[16]	✓	✓				Twitter	✓
1	[90]	✓	✓				Twitter	N/A
1	[11]	✓	✓				Twitter	partial
1	[22]	✓	✓				Twitter	✓
1	[12]	✓	✓				Twitter	✓
1	[61]	✓	✓				Reddit	partial
1	[77]	✓	✓				Twitter	no
1	[128]	✓	✓				Twitter	✓
1	[118]	✓	✓				Twitter	✓
2	[9]			✓			Twitter	no
2	[28]			✓			Twitter	✓
2	[13]			✓			Twitter	no
2	[26]			✓			Twitter	N/A
2	[85]			✓			Twitter	partial
3	[139]		✓		✓		Twitter	✓
3	[137]				✓		BSS	partial
3	[125]	✓	✓		✓		Twitter	N/A
4	[154]		✓			✓	Flickr	✓
4	[69]		✓			✓	Twitter	✓
4	[37]					✓	Twitter	N/A
4	[144]	✓	✓			✓	Taiwan forum	✓
4	[152]	✓				✓	Twitter, Facebook, Google, webpages	✓
4	[142]		✓			✓	Twitter	✓
4	[143]	✓	✓			✓	Twitter	N/A
4	[67]	✓	✓			✓	Twitter	N/A
5	[70]	✓	✓				Twitter	N/A
5	[136]		✓				Twitter	N/A
5	[147]	✓	✓				Twitter	partial
5	[71]		✓				Twitter	no
5	[43]		✓				Twitter	✓
5	[2]						Twitter, blogs	partial
5	[60]		✓			✓	Facebook, e-news, magazines	✓
5	[8]		✓			✓	Twitter	partial
5	[21]						Twitter, Facebook, Instagram, past elections, traditional polls	✓

3.1.8 Discussions, Limitations and Challenges for Future Research

In this section, we resume the answer for the research questions **Q1, Q2, Q3** presented in Section 3.1. We firstly present our categorization about the main election forecasting approaches using social media found for answering **Q1**). After that, we identify gaps and limitations in the current literature for answering **Q2**, both from opinion mining point of view, discussing limitations in regard to the proposals presented in the literature, and from tradition election polls point of view, discussing how opinion mining in social media can help leverage their results. Finally, we point out directions for future research, mainly from the machine learning and Artificial Intelligence point of view, for answering **Q3**.

Main Election Forecasting Approaches: One of the goals of this SLR was to identify the main approaches for forecasting elections using social media. We observed that, in general, the surveyed approaches can be categorized into four main groups: (i) Counting Based Approach [6,11,12,16,22,23,39,55–57,61,62,66,73,77,87,90,104,108,111,117,118, 127–129,132,140,148] – this is the most simple approach, in which papers basically sum mentions to a specific party/candidate (volume-based) or sum the occurrence of positive opinions that mention a given party/candidate (sentiment-based) to predict election outcomes; (ii) Political Alignment Approach [9,13,26,28,85] – papers that try to predict the political alignment/leaning of the users to forecast election outcomes; (iii) Event Detection Approach [125,137,139] – papers that relate the victory of a candidate/party to the occurrence of political events and predict outcomes based on that; (iv) Popularity Based Approach [37, 67, 69, 142, 144, 152, 154] – papers that propose to use a formula to infer candidates popularity and assume that the most popular candidate will win the election.

In what follows we present a discussion about the surveyed papers according to their category. We mention some papers of each approach as example, highlighting some of their characteristics and emphasizing the ones that present particular aspects that differ them from other ones in the same category. In addition to the four approach categories identified in the literature, we create a category called *Other Works* [2,8,21,43,60,70,71,136,147] to group papers that do not fit in any of the identified categories.

Counting Based Approach: We verified that works that *only* consider volume were not successful in most of the cases. [55], [73], [11], [61] and [56] tested both methods – sentiment- and volume-based. While [55], [11], and [61] only achieved good results using the sentiment-based strategy, [73] achieved the expected election winner with both methods. [56] concluded that the volume-based result was equivalent to the sentiment-based result on the national level. However, in their experiments the sentiment-based method outperformed the volume-based when it comes to state level elections. [127] tested two methods using sentiment-based strategy, the first one uses machine learning algorithms and the other one is based only on dictionaries. In their case, they only predicted the correct winner in the experiments that use machine learning algorithms. Other approaches such as the one presented by [77] did not predict the correct election winner using the sentiment-based strategy. [140] adopted a counting approach based on the number of Facebook likes and concluded that election prediction based on Facebook is not accurate when compared to traditional polls.

Some papers adopted counting based approaches that are more elaborated. [39] go further by filtering non-personal accounts, slacktivists and spam users. In this context, [62] built a classifier to detect buzzer accounts aiming at removing data noise. [6] also removed some spam and news accounts and considered demographic information. [39] used both demographic and geolocation information. [132] and [73] tried to predicted not only vote share but also seat share. [132] and [128] tried to find out the number of seats. [148] and [128] presented a subtle difference in relation to the other sentiment-based approaches since this approach [148] also takes into account the number of tweets with negative sentiment. While [148] assume that a negative opinion can be interpreted as a vote to the opposition, [128] presented a formula to compute the actual sentiment score that depends on the negative score. [90] adopted a counting based approach to find out the most supported agenda and from that predict the election winner. Finally, [11] find out tweet topics using word co-occurrences and infer the sentiment of the tweet based on the sentiments of the topics.

Political Alignment Approach: From the works that classify tweets according to the political alignment (e.g: as “republican"or “democrat" [85]), the success rate was variable. While [26] argued that they achieved promising results, [28] predicted the correct winner, [13] and [9] failed to predict the right result, and [85] presented the correct prediction only for some districts.

Some of the papers were concerned with other aspects in their analysis. For example, [28] proposed to compute a reputation score based on the number of friends and followers that a Twitter account has to discard bot accounts. [28] and [13] take into account geolocation information.

Event Detection Approach: [137] assume that events can have positive or negative impact on people opinions in relation to a given candidate. They try to detect the occurrence of events that are related to the elections based on terms that appear in users comments on an online forum. By using rules, they determine the winner party based on events occurrence. With this approach they predicted the election winner correctly for some cities. On the other hand, [125] and [139] use event detection approaches that were successful to predict the election winner. [139] present a more elaborated approach, by clustering tweets that belong to the same event/topic based on their terms. They use part-of-speech tagging and named entity recognition to determine events location. Additionally, the work by [139] also predicted tweets sentiment and proposed some rules to detect sarcasm constructions.

Popularity Based Approach: [154] achieved a successful result using a strategy that try to predict the candidates popularity by combining textual and image features using data from Flickr. [142] and [152] were also successful to predict the correct winner. Different from the other approaches that were grouped into this category, [152] use demographic information and considered many sources of information to infer candidates popularity (Twitter, Facebook, Google, and candidates' campaign websites and offline data from pollsters). [37] and [69] presented approaches where the candidate popularity was computed based on an analysis of graphs of user interactions in Twitter, achieving good results. [144] also argued that they predicted the election winner correctly, by using an approach that takes into account rating records of candidate related articles in a popular Taiwan forum. [67] proposed a popularity formula that is based on sentiment score of positive and neutral tweets. achieving the correct outcome. [143] predicted the correct winner in one of their experiments, which computes candidate popularity based on the score of keywords related to him.

We noticed that most of the works that achieved successful predictions adopted a *sentiment analysis* step and used *Twitter* as source of opinions, independently of the approach adopted. For this reason, we believe that this may indicate that Twitter is a promising source for collecting electoral opinions and that sentiment analysis should be considered for those who want to achieve better electoral predictions using social media. An example of this are the papers: [87], [28], [139], [144] and [43], which adopted different approaches (Counting Based Approach, Political Alignment Approach, Event Detection Approach, Popularity Based Approach, and Other Works, respectively). Also, most of the successful works have attempted to predict presidential election results, which leads us to believe that social media election surveys are more suitable for analyzing elections at national level.

Limitations and Challenges for Data Science: There are many aspects that can lead to wrong predictions in traditional polls. [133] and [155] pointed out that last minute changes, *i.e.*, a shift in vote share towards one of the parties between the final polls and election day, may be one of the reasons for wrong predictions in traditional polls. [20] highlights that fake news and social media bots had a high influence on voters opinions in the 2016 US Presidential elections, factor that could be responsible for vote changes in a short period. According to Michael Bruter (2017) [27], a political scientist at the London School of Economics, another reason for wrong predictions is the fact that some people only make up their minds on the eve of the election. [155] argues that wrong predictions may also occur when pollsters fail to achieve a representative sample, due to the lack of accurate phone databases or when pollsters assume that people who did not vote in past elections will not vote in the next elections, for example. In this way, opinion mining and social media data analysis can be a helpful tool for leveraging the prediction power of traditional tools for (i) detecting fake news and social media bots; (ii) detecting opinion changing of the people regarding to the candidates over time; and (iii) identification of different types of voters considering an analysis of their behavior and profile, increasing the representativeness of the population in the polls. In the last case, tools for supporting the identification of voters profile may be adopted.

On the other hand, although social media has emerged as a promising way of collecting opinions in real-time, there are yet many limitations when social media data is used to predict election outcomes from computer science point of view:

- **Methods cannot be generalized**: Usually, the methods found in the literature for election forecast using social media only consider specific elections, implying that the results cannot be general enough to contemplate other elections;

- **Non-availability of datasets**: One of the gaps found in this domain is that the electoral datasets are not freely available for the community that work on this topic. It is not possible to evaluate the success of existing forecasting methods without analyzing their datasets. We cannot guarantee, for example, if they were successful because of the election (*i.e.*, the case in which the election is not a close dispute) or if they were successful due to the effectiveness of the adopted methodology. We believe that this can be the reason of some similar forecasting strategies achieving the correct prediction in some papers (that refer to a given election) and the wrong in other ones. In order to address this issue, a temporal analysis could be conducted evaluating data from time to time in different time periods that preceded the election to check if the predicted election winner changes over time. Also, we cannot compare existing approaches by predictive accuracy since they refer to different datasets/elections;

- **Filtering potential users**: Data from social media can be posted by non-person users, such as organizations. Few approaches discard this kind of post. Additionally, the majority of approaches consider in their calculus different posts by the same user. This can impact the final results since each person only can vote one time. Another problem is that most of these approaches also did not analyze if the social media user account belongs to a person that is permitted to vote in the given election or even if it belongs to a real person, *i.e.*, they did not discard fake accounts and accounts that belong to non-voters. In this context, data collected from social media can be posted by bots (spam). Therefore, the nature of data collected from the Web can have many biases, reducing trust and the credibility of the results. Furthermore, the general profile of people that use social media is different from the voters, mainly in developing and non-developing countries, *i.e.*, the majority of Twitter users are young men that live in urban areas as pointed out in [39]. Taking into account these factors, sometimes a big amount of data can not reflect a statistically representative sample of the general population;

- **Sentiment analysis challenges**:

  Data Labeling: We noticed that many approaches that forecast the election outcomes based on sentiment analysis rely on straightforward methods for labeling the sentiment of the sentences (such as emoticons or dictionaries) (see Section 3.1.3 – Figure 3.4), ignoring that predictions may be misleading due to *the difference between domains*. This is because the polarity of a word depends on the context that it is inserted. For instance, the word *scary* can express a negative sentiment when extracted from posts related to general contexts and positive sentiment when extracted from opinions about horror movies. In addition, terms that denote sentence sentiment can vary according to the domain, *i.e.,* although the word *cheap* indicates a positive sentiment in the product reviews domain, this word does not denote a positive sentiment in tweets talking about a given political candidate;

  Sarcasm and Irony: Another challenge to infer sentiment polarity of texts is that ironic and sarcastic posts are prevalent on social media. In this way, a text thanking a person, for example, can be expressing the contrary opinion. Although was observed by recent approaches social media texts related to elections are full of ironic content [38], only one work analyzed in this literature survey takes into account this issue. However, even in such a case only a simple mechanism is presented to deal with sarcasm/irony;

- **Absence of a methodological pattern**: There is not a default methodology to predict elections based on data from social media, *i.e.*, approaches use different steps to collect data and estimate the prediction. We have observed that each research surveyed considers different periods to begin/end the data collection. Additionally, each one of them collect data containing different kinds of terms (for instance, candidates' names, parties' names, campaign slogans, and so on) and considering the different quantity of posts (see Section 3.1.1);

- **Accuracy of the polls based on social media**: In general, predictions based on polls that use social media have lower accuracy than predictions based on traditional polls, *i.e.*, most of times traditional polls present results closer to the actual results. However, data analysis and opinion mining tools on social media in election scenarios can be very important for improving traditional polls results;

- **Absence of patterns for evaluating the predictions**: There is not a consensus in the literature about how to evaluate the predictions. While some papers compare their prediction with the real predictions (*post hoc* analysis), other ones compare their result with the result of predictions based on traditional polls. Also, the majority of approaches that argue that they were successful only take into account the absolute election winner (and not vote share);

- ***Post hoc* analysis**: Several approaches present a *post hoc* analysis, analyzing social media data, and calculating the prediction after the occurrence of the real election. According to [51], this cannot be considered a *prediction* at all. So, developing tools for tracking social media users behavior regarding to their candidates is very important in this scenario;

- **Elector behavior**: as pointed out by [142], the behavior of the elector can affect the accuracy of the predictions methods based on social media. This is because while most supporters of a given candidate A may not attack its adversaries on social media, the supporters of a candidate B may usually attack the other ones, posting a huge amount of data;

- **Annotator bias**: Supervised machine learning techniques for sentiment analysis/opinion mining rely on labeled datasets that can be annotated by a small set of persons that also do not reflect the characteristics of the electorate. This fact may affect results on classification tasks.

Open Issues and Future Research from the AI Point of View: In what follows, we highlight some lines for future research in opinion mining for elections outcomes predictions:

- **Opinion mining using multimodal data**: Only one of the papers surveyed considered images shared on social media in order to predict election results [154]. This field could be better explored by future related lines of research since many social media posts contains images and not only text;

- **Data Streaming Mining**: We have identified a lack of approaches that deal with election forecast using data stream mining methods. These methods are interesting as they adapt the machine learning model over time [15]. Moreover, there are also methods that allow to identify concept drift, which could be interesting in electoral scenario. On the other hand, one main challenge for this is acquiring labeled data. In this way, unsupervised or semi-supervised data stream mining methods could be explored;

- **Active learning**: We have noticed that none of the papers in our survey adopt active learning methods [134]. These methods include the human in the machine learning loop. Basically, they select the most representative instances to be labeled by humans to deal with the lack of labeled data in diverse domains. These methods can be investigated in future researches to cope with the problem of short period for labeling domain specific data, helping to improve election outcomes prediction;

- **Domain adaptation and transfer learning**: We did not find in the literature papers that predict elections using transfer learning and domain adaptation strategies [96], which are techniques that can be applied when labeled data is scarce (or not available) in the target domain, as in electoral scenarios. For example, pre-trained word embedding techniques could be better explored in order to improve the accuracy of the results. Additionally, recent lines of research based on language modeling such as ULMFit [59], ELMo [102] and BERT [35], could also be investigated. These methods take advantage of the ability of language modeling networks to representation and semantics and fine-tuning them to perform classification tasks. Also, as a good practice, the community that work on topics related to electoral domain problems could concentrate some efforts to make available existing labeled domain datasets and existing machine learning models to facilitate forecasting tasks of future elections as well to enable additional analyzes of existing ones. Another important aspect of this line of research is verifying the possibility of transfer learning considering datasets of different languages (datasets from previous elections in different languages could be used) or different domains, as well as proposing new mechanisms to search which datasets are more suitable to this task. This is an important line of research as the number of datasets available on internet for opinion mining grows and the number of experiments needed to choose the appropriated classifier can also exponentially grow.

- **Gamification**: Given the importance of having labeled data in the electoral domain to build reliable classifiers, we believe that gamification labeling strategies [92] can be explored to motivate manual labeling of social media electoral opinions.

3.2 Approaches for Dealing with Unlabeled Data

In this section we present approaches that propose means to deal with classification tasks when labeled data in the target domain is not available. An ad-hoc search was conducted to find the approaches described in this section.

Many approaches have been proposed in the literature to deal with unlabeled data in the target domain. One example are the approaches for learning an embedding space to reduce the difference between source and target domains such as the Structural Correspondence Learning (SCL) [17, 81]. A common representation is obtained using *pivot* features (words) that occur frequently in both source and target domains to create a correspondence between features from these two datasets. Similarly, the work presented in [94] presents the Spectral Feature Alignment (SFA) that aligns domain-specific words from different domains into clusters, using domain independent words as a bridge.

Sentiment graphs (SG) are used by Wu and Huang (2016) [150] to extract polarity relations among words to compare different domains. Two types of relations are explored, namely: (i) sentiment coherent relation; and (ii) sentiment opposite relation. Those sentiment relations are identified according to manually selected rules. For example, words connected by conjunction prepositions are linked by sentiment coherent relation. On the other hand, words connected by adversative conjunctions are linked by sentiment opposite relations. A sentiment graph is created to each domain, where nodes represent words and edges represent the sentiment relations between words. A high domain similarity occurs when a pair of domains have many sentiment word pairs in common and the polarity relation scores of these word pairs are similar.

Reverse Classification Accuracy (RCA) methods are another possibility to choose proper source datasets, by estimating the performance drop of a model when evaluated on a new unlabeled target domain [40,42,158]. Basically, a given source dataset is used to build a classifier C_1 that will predict labels for the unlabeled target dataset. In turn, the new labeled dataset is used to train a classifier C_2. After that, the performance of classifier C_1 is compared to the performance of classifier C_2 using a subset of the source dataset as test data. The RCA* is a variation that uses one more classifier at the beginning of the process, to generate new labels for the source dataset and build from these new labels the classifier C_1. Confidence Based Measures (CBM) consists in using confidence scores – the certainty of the model over its predictions – as a similarity estimator. Therefore, this method requires not only the predicted labels but also the confidence scores [40].

Target Vocabulary Covered (TVC) was explored by Dai et el. (2019) [33] to choose proper source datasets based on similarity. It computes the number of words in the intersection between source and target datasets divided by the number of words of the target dataset. A variant of this measure that only considers nouns, verbs, adjectives was also explored. A Language Modeling (LM) based approach was also investigated by these authors. The basic idea is that, every time a language model trained on the source dataset finds a sentence in the target dataset that is very unlikely, then the model will assign a low probability and a high perplexity value. The closest source dataset is selected by taking into account all sentences from the target dataset. Finally, Dai et el. (2019) [33] explored an approach called Word Vector Variance (WVV) whose first step is to train a word vector on the source data using the skipgram model. Next, this trained word vector wv_1 is used to initialize weights of a new model trained on the target data, generating the trained word vector wv_2. The general idea is that the smaller the word vector variance is, the more similar the two datasets are.

Table 3.11 presents a summary of the approaches aforementioned presented. Column *App.* indicates the name of the approach, column *Efforts* refers to the required efforts to implement the approach – where n = number of source datasets and m = number of target datasets, column *Sem.* indicates if the approach is able to capture semantic information (e.g.: using embeddings), column *Diff. lang.* indicates if the approach supports the selection of datasets from different languages, column *Based on* refers to the main elements

Table 3.11: Approaches to Deal with Unlabeled Target Data

App.	**Efforts**	**Sem.**	**Diff. lang.**	**Based on**
SFA SCL	learn a common feature representation meaningful for source and target domains	✓		pivot features
SG	create n sentiment graphs	✓		sentiment graphs
TVC	compute how many words are in the vocabularies intersection			dataset vocabulary
WVV	train $n + (n \times m)$ word vectors[1]	✓		word vectors
LM	train n language models	✓		language models
RCA	train $n + (n \times m)$ classifiers[1]	✓	✓	classification models
RCA*	train $2 \times n + (n \times m)$ classifiers[1]	✓	✓	classification models
CB	train n classifiers	✓	✓	classification models
Our proposal	create similarity ranking	✓	✓	similarity ranking

[1] if there is no intersection between the set of source and target datasets.

used to improve/select source datasets.

There are many aspects that differ our approach from related work. For example, RCA requires that two classifiers (or more as in the variation called RCA*) are built for each source dataset before computing similarity between source and target datasets. To measure dataset similarity between datasets using the LM approach, language models must be trained for each one of the source datasets. Also, the WVV requires the training of many word vectors. Techniques such as [17], [81], [94] may not work properly when source and training data do not share much information. The proposal of Wu and Huang (2016) [150] depends on manually selected rules to identify sentiment relations. We believe that rules are subjective and cannot be enough to represent similarity aspects in specific domains.

Different from these approaches, the dataset selection method presented in this thesis does not require any model or classifier to be trained. The dataset ranking is created based only on the analysis of semantic similarity between the source and target datasets. Then, our heuristic allows us to compare several datasets quickly and select the ones more similar to our domain of interest, being more appropriate for scenarios subjected to time restrictions and large amounts of unlabeled data, as in electoral scenarios. Furthermore, our proposed heuristic is independent of the selection of the algorithm that will be later adopted to train sentiment classifiers. The adoption of multilingual embeddings – which is another point that distinguishes our approach from the others – allows us to take advantage of data from different languages, a factor that may be very interesting in domains such as the electoral one.

3.3 Final Considerations

In this chapter, we presented our findings obtained in a systematic literature review. Our review considered papers from 2014 to 2020. From this, we observed that there are many gaps for predicting elections through opinion mining in social media. One important aspect we observed was related to the complexity of labeling data in this scenario. In addition to that, we also presented in this chapter some strategies for dealing with unlabeled data in the target domain in a scenario where classification tasks need to be performed. Based on that, we decided to explore *transfer learning* for analyzing opinions in this scenario, focusing on the improvement of the *sentiment analysis* task. In this context, we investigated an approach that take advantage of existing datasets to be used in transfer learning for sentiment analysis tasks in the electoral domain. As pointed out along of this chapter, there are still many other future work that can be conducted to fill gaps and open issues of this field. In the next chapter, we describe the thesis proposal.

Chapter 4

Thesis Proposal

This chapter presents the proposal of this thesis. Section 4.1 relates our study hypothesis to the main study topics, which are sentiment analysis and transfer learning. Section 4.2 depicts our proposal general structure and presents the methodology steps.

4.1 Proposal

The electoral domain is a real scenario that motivated our research due to the importance of electoral analysis to society and the several computational issues involved that make data analysis in this domain a complex task. This issues include: short time for labeling data and high level of data complexity. The systematic literature review (SLR) that we conducted to better understand this domain problem pointed out many open issues and limitations in approaches for forecasting electoral outcomes, as was discussed in Section 3.1. Therefore, we chose to focus on the *sentiment analysis* task in this domain using social media and computational techniques, as it is a closely related task to understanding people's opinions that was explored for most of election forecasting approaches found in the systematic literature review. Also, with the SLR, we observed that most of the sentiment analysis approaches of this domain rely only on strategies for automatic labeling electoral data using generic lexical dictionaries [23], [139], [136] or emoticons [39], [55], for example. This is because it is not possible to manually analyze thousands of electoral opinions (or even millions) in a timely manner. However, the usage of generic labeling methods that do not consider information specific to the target domain to assign polarities, may affect sentiment analysis predictions in specific domains, such as the electoral one.

Sentiment Analysis: In fact, dealing with none or limited labeled data in the domain of interest is a challenge that has been commonly tackled in *sentiment analysis* tasks [1, 124]. While there may be plenty of training labeled data in sentiment analysis domains, there is no guarantee that they follow the same distribution of the specific target domain of interest. The difference in the distribution between training and target datasets called *domain shift*, may considerably impact the success rates of classification tasks on target data [40]. One factor that contributes to explaining the decreasing of classification success rates when domain shift occurs is that expressions or words used to denote sentiment and characterize a sentence as positive, negative, or neutral may vary from domain to domain [151]. For example, while for product reviews words such as *cheap* and *useful* are terms that denote positive sentiment, these words are not helpful for detecting positive sentences for other domains such as movie reviews or political domains. Analogously, while the word *unpredictable* denotes a positive sentiment for book or movie reviews, this word may indicate a negative orientation for automobile reviews (e.g.: "The steering of car is unpredictable") [68]. We elicited several reasons for sentiment prediction problems due to domain shift, as follows [3, 40, 82, 156]:

- **Polysemy and Polarity Divergence:** Words can have different meanings and may denote different sentiments in different contexts;
- **Feature Divergence:** Terms that denote positive/negative sentiments may not be the same for source and target domains;
- **Sparsity:** Different datasets may have very different vocabularies and words may appear frequently in the target domain but may not appear (or appear rarely) in the source domain;
- **Writing Style:** Writing style may vary from domain to domain.

According to Liu (2020) [84], generic sentiment classification commercial systems usually do not perform well when applied to predict sentiment related to political data due to the complexity of this domain. The study presented in [119] reinforces this idea. In such a study, a set of election-related tweets were categorized according to their sentiment, using two approaches: (i) machine learning sentiment classifiers using generic data – Microsoft Azure Sentiment Analysis module; (ii) a manual labeling method, where tweets are analyzed by several human annotations, and the final label is obtained using majority voting. After that, the labels achieved with each one of these two approaches were compared, and a high level of divergence between labels was identified, i.e., labels were different for more than half of the tweets considered.

Transfer Learning: Transfer learning [95] is based on the idea that exploiting prior knowledge from a given task or domain may be used to help build models for other domains or tasks. Aiming to improve opinions analysis regarding the electoral scenario, we decided to explore transfer learning strategies that consider dataset similarity. The idea is to minimize prediction problems that may be the result of different data distributions between the training dataset – used to train the sentiment analysis classifier – and the electoral dataset.

Our preliminary experiments published in the Brazilian Conference on Intelligent Systems (BRACIS) 2019 [121] suggested that dataset similarity can help in choosing promising training datasets. In such study, the similarity is computed taking into account dataset vocabularies. Jaccard distance and euclidean distance based on Glove embedding metrics are explored for selecting training datasets for the sentiment analysis task of a dataset containing tweets related to the 2018 Brazilian Presidential Elections.

This thesis hypothesis is as follows, as was presented in Chapter 1:

Hypothesis (H): *If there is a high degree of similarity between a source labeled sentiment analysis dataset and a target electoral dataset, then machine learning classifiers trained with this source dataset will achieve proper sentiment predictions for the target electoral dataset.*

This hypothesis follows the intuition that similar sentiment analysis datasets tend to share terms that denote sentiment. To emphasize such a hypothesis, we performed a survey of the 50 most frequent terms of examples that are labeled as negative, using two airlines datasets (*Airlines ES* and *Airlines EN* described in Chapter 5). The first dataset is written in spanish and has opinions about spanish airlines and the second one is an english dataset that contains opinions about american airlines. We verified that the two datasets share several words that have exactly the same meaning among the top 50 most frequent terms: flight, vuelo; service, servicio; hours, horas; customer, cliente; plane, avión; hour, hora; flights, vuelos; bag, maleta; waiting, esperando; thanks, gracias; delay, retraso; luggage, equipaje; airport, aeropuerto; bags, maletas; fly, volar.

The most frequent terms related to negative sentiment in the first dataset are summarized in Figure 4.1 that illustrates a word cloud, and the 50 most frequent word list is as follows: flight, united, usairways, americanair, southwestair, jetblue, cancelled, service, hold, hours, help, customer, time, still, plane, delayed, hour, call, flightled, us, one, flights, bag, phone, gate, need, waiting, back, late, please, thanks, airline, trying, never, minutes, worst, like, delay, wait, today, luggage, guys, going, even, told, day, airport really, bags, fly.

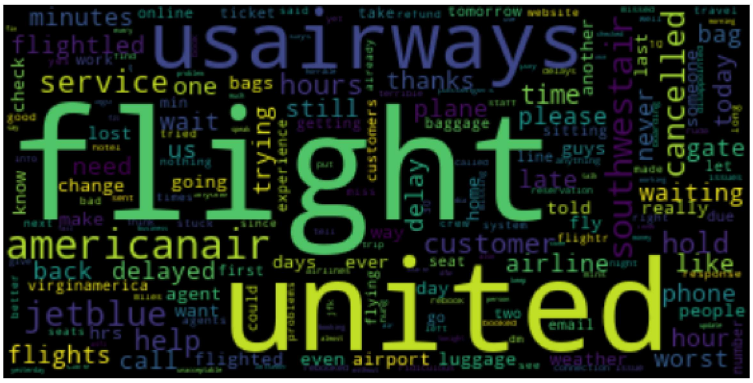


Figure 4.1: Word Cloud of Negative Examples - Airlines EN

The most frequent terms related to negative sentiment in the second dataset are summarized in Figure 4.2 that illustrates a word cloud, and the 50 most frequent word list is as follows: iberia, ryanair, vuelo, españa, maleta, vuelos, hola, huelga, avión, madrid, spanair, accidente, video, letal, vueling, décadas, hacer, gracias, pilotos, pasajeros, horas, días, solo, maletas, ryanair_es, equipaje, servicio, retraso, cliente, día, aeropuerto, respuesta, dinero, vez, iberiaexpress, volar, vía, billete, hora, ahora, esperando, tierra, siempre, destino, pasajero, atención, suerte, mal, menos, años.

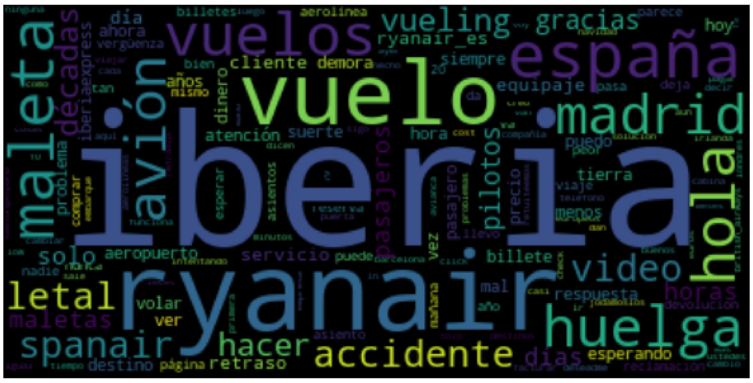


Figure 4.2: Word Cloud of Negative Examples - Airlines ES

On the other hand, when we collected the top 50 most frequent terms related to the negative examples of a music festival dataset (*Music Festival EN* dataset presented in Section 5), the same is not true. The list of the top 50 most frequent words is as follows: coachella2015, coachella, lineup, line, year, drake, go, like, going, weak, fuck, see, tickets, really, better, responsible, impressed, good, one, looks, acdc, know, disappointed, years, playing, whos, festival, headlining, wtf, steely, dan, last, excited, sucks, bands, disappointing, still, sad, day, worst, ever, gonna, even, lol, sold, could, though, ticket, far, seen. The word cloud illustrated in Figure 4.3 presents a summary about the most frequent terms related to negative examples in the music festival dataset.

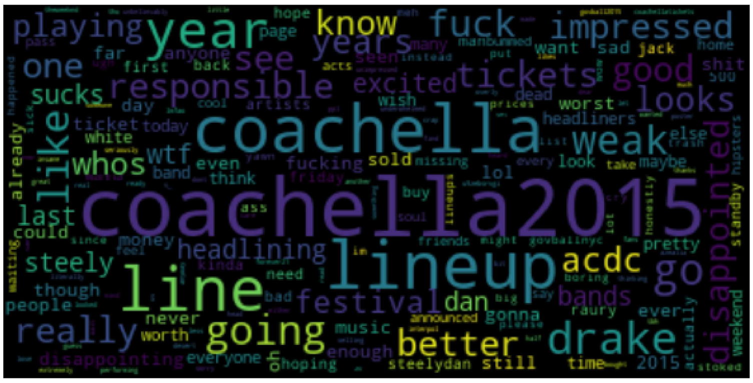


Figure 4.3: Word Cloud of Negative Examples - Music Festival EN

In what follows, we will use the term dataset to refer to a set of sentences that belongs to a certain domain (e.g. movie reviews, election opinions, product reviews, etc) and may have been collected to analyze a certain task (e.g. sentiment analysis, presence of offensive speech, next word prediction, etc). We use the term *source dataset* to refer to the dataset from which we will try to take advantage of knowledge to perform analyzes on another dataset that we call *target dataset.* To address the hypothesis in our case study, we need to perform an analysis of existing datasets labeled for *sentiment analysis* task and investigate *transfer learning* methods that apply to our scenario, where there is a lack of labels for the target domain. So far, prediction tasks related to elections are not reusable and comparable since the manual annotation process demands time and human effort, implying only small annotated datasets that are specific to an election. Therefore, the method that we will adopt is transferring learning of sentiment analysis tasks from other datasets to the target one. The notation D_s and D_t is used to describe the source and target domains, respectively. The term T represents the task that the datasets are related to, and T_s and T_t refer to the source and target tasks, respectively.

At first sight, three methods could be adopted to deal with transfer learning between a source and a target dataset.

(a) Transfer within the same domain and the same task, but with different elections (elections of other years or other types). This scenario can be formally represented as: $P_{D}s(x) \neq P_{D}t(x), Ds = Dt$ and $Ts = Tt$, where P denotes the probability distribution;

(b) Language transfer implies transferring learning from other tasks. This case refers to methods in which sentiment classifiers are built on top of language models, followed by a fine-tuning step. This can be formally represented as: $P_Ds(x) \neq P_Dt(x), Ds = Dt$ or $Ds \neq Dt$ and $Ts \neq Tt$;

(c) Adopting existing labeled datasets for the same task as the starting point for transfer learning, aiming at reusing knowledge. This scenario can be formally represented as: $P_Ds(x) \neq P_Dt(x), Ds = Dt$ or $Ds \neq Dt$ and $Ts = Tt$;

As a result of the systematic literature review (SLR), we have identified the lack of annotated datasets for sentiment analysis in the electoral domain in Portuguese. Also, our premise is that there are no labels for the target dataset due to time restriction and data complexity; therefore, it would not be possible to perform a fine-tuning process for the sentiment analysis task. For this reason, we have chosen to discard alternatives (a) and (b) and to proceed with the alternative (c), combining them with embeddings from pre-trained models. Figure 4.4 illustrates an example where sentiment analysis labeled datasets of several domains (movie reviews, product reviews, TV show reviews) are available, but there are no labeled examples for the dataset that belong to the domain of interest (electoral dataset). The idea is that the labeled datasets could be used (independently or combined) as training data to predict the sentiment in the unlabeled dataset of the target domain.

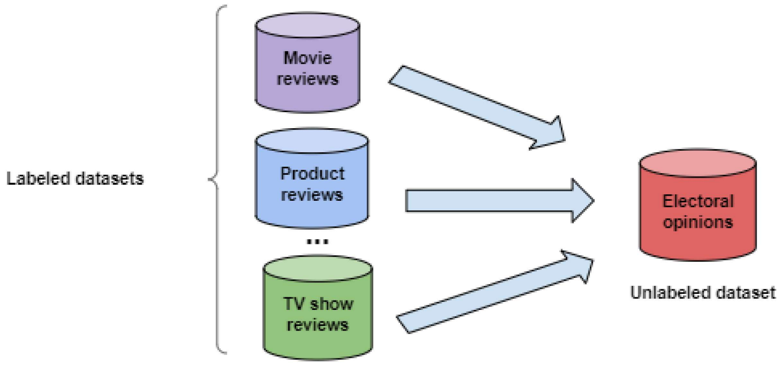


Figure 4.4: Transfer learning between different sentiment analysis datasets

To test our hypothesis, our experimental analysis will consider labeled datasets from different languages and different domains labeled for sentiment analysis. Datasets from different languages but in the same domain (elections scenario) will be tested to enlarge the possibilities of dataset selection. This is because we understand that due to the nature of the analyzed scenario – elections are events that occur regularly in democratic environments around the world –, datasets from the electoral domain may hold relevant shared information. To that, examples are represented as embeddings gathered from pre-trained models, as will be explained in the next section.

In this context, this thesis is proposing a method for selecting training datasets based on measuring the semantic similarity between potential source datasets and the target electoral dataset, considering data from different languages and domains. According to the hypothesis, sentiment analysis datasets that contain similar content are more likely to lead to sentiment predictions closer to the ones obtained with classifiers trained with in-domain data if they existed. The basic premise is that the target dataset has no labels, and the similarity function must work in this scenario. Therefore, our proposal does not depend on building models since training datasets are selected based only on a ranking created according to dataset similarity analysis.

The evaluation method applied to validate the thesis hypothesis relies on the analysis of the F1-score values achieved when using the selected datasets for training classifiers with a set of machine learning algorithms. For evaluating the results related to the electoral dataset of the 2018 Brazilian presidential elections, we use labels obtained with both automatic and manual labeling processes, as is better described in Section 5.1. The Friedman test [48] is applied to check if there is a statistical difference between the F1-score of the datasets being compared. Also, a unified F1-score rank is created to point out the source datasets that achieved the best results. Finally, the Nemenyi [34] is adopted as a post hoc analysis to identify statistically equivalent datasets, in cases where this test is applicable. All these steps will be detailed in the next section.

4.2 Proposed Method

The proposed method relies on the idea that existing labeled datasets (and knowledge acquired from them) could be useful for analyzing electorate opinions. Taking advantage of them would be a way to help prediction tasks related to sentiment analysis in social media texts that refer to the electoral domain. Our hypothesis is that the usage of existing labeled datasets of other domains would avoid (or reduce) the need to manually label electoral datasets during the election course. To test this hypothesis, we conduct the following steps: (i) selecting potential training datasets; (ii) after that, a pre-processing phase is applied over all datasets to clean up their content (this phase involves discarding irrelevant information and normalizing data); (iii) next, similarity metrics are employed to build a unified ranking that will be used to select proper datasets as training data.

The main idea of our proposed method is to select promising datasets to be used to solve sentiment analysis tasks when in-domain data is not available (or is scarce). The diagram in Figure 4.5 illustrates the main steps for executing our dataset selection method, whose main activities (1 to 4) are described in what follows. The thin-bordered circle represents the beginning of the process. Process activities are denoted by round-cornered rectangles. The thick-bordered circle denotes the control that ends the activity. Dotted rectangles group-related activities.

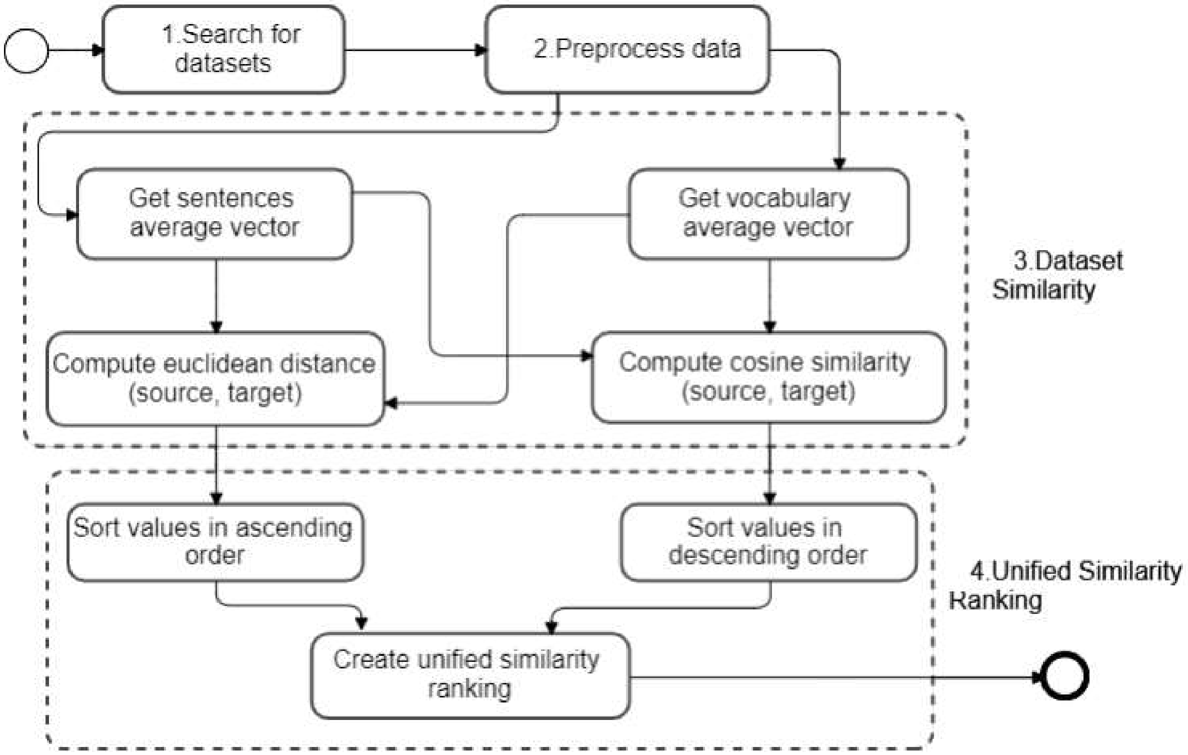

1.Search for datasets
2.Preprocess data
Get sentences average vector
Get vocabulary average vector
3.Dataset Similarity
Compute euclidean distance (source, target)
Compute cosine similarity (source, target)
Sort values in ascending order
Sort values in descending order
4.Unified Similarity Ranking
Create unified similarity ranking

Figure 4.5: Dataset Selection Method Overview

4.2.1 Dataset Search

The first step of our method consists of acquiring labeled training data, assuming that existing knowledge may be useful for sentiment classification tasks when labeled data is unavailable or is scarce in a given target domain. For this reason, we search for sentiment analysis labeled datasets, given our focus on those problems.

4.2.2 Data Preprocessing

Preprocessing is applied to source and target datasets to remove noise data. Data cleaning is based on the following steps: removing special characters, punctuation, accents, and numbers; discarding stop words (except the ones that denote contrast or negation); and converting all words to lowercase. Also, sentiment classes of the datasets are analyzed, maintaining only instances whose classes are common to all candidate datasets. Next, the datasets are converted to numerical representations (e.g., BoW or embeddings).

4.2.3 Dataset Similarity

We hypothesize that selecting similar datasets before training machine learning classifiers is related to a greater chance of achieving satisfactory prediction results when labeled data in the target dataset is unavailable or scarce. Two traditional similarity/distance metrics[1] are used in this approach:

Cosine Similarity: Given two vectors $\mathbf{x}$ and $\mathbf{y}$, the cosine similarity measures the cosine of the angle between them, defined by Equation 4.1.

$$s(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}} \tag{4.1}$$

Euclidean Distance: Given two vectors $\mathbf{x}$ and $\mathbf{y}$, the euclidean distance measures the absolute value of the numerical difference of their coordinates, defined by Equation 4.2.

$$d\left(x,y\right) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \tag{4.2}$$

Given a set of datasets D, and D_i, D_j a pair in this set. Based on the set of words that belong to D_i is $W_i = w_1 i, ...w_i n$ and the set of words $W_j = w_1 j, ...w_j n$ that belong to D_j, we propose four ways to analyze the semantic similarity between datasets, which will be explained as follows.

Euclidean Distance based on Sentences: For each example that belongs to the dataset, we get the corresponding numerical representation according to the vectorization method adopted. After that, we compute a single embedding vector by averaging all of the sentence embeddings of the dataset. In this way, every dataset will have an embedding that will represent its general context. The similarity between datasets is measured according to the Euclidean distance, i.e., the smaller the value of the Euclidean distance between

[1] We did not adopt the jaccard distance metric because it is not suitable for comparing embeddings.

the average embeddings of a source-target dataset pair, the greater the similarity between this dataset pair.

Cosine Similarity based on Sentences: This case is similar to the aforementioned case since an average embedding is computed for each dataset, but cosine similarity is used to measure dataset similarity. Therefore, the greater the value of cosine similarity between the average embeddings of a source-target dataset pair, the greater the similarity between this dataset pair.

Euclidean Distance based on Vocabulary: In this case, we compare the vocabulary of two datasets to measure their similarity. We consider that the vocabulary of each dataset is composed of the words that appear in it, and a single embedding vector is computed by averaging embeddings of these words. Each word is only considered once. Euclidean distance is used to compute the similarity between the dataset vocabularies.

Cosine Similarity based on Vocabulary: Similarly to the last case mentioned, the vocabulary of the dataset pairs are compared using the cosine similarity metric.

These similarity methods will be used to compare pairs of datasets according to their content and to select training datasets.

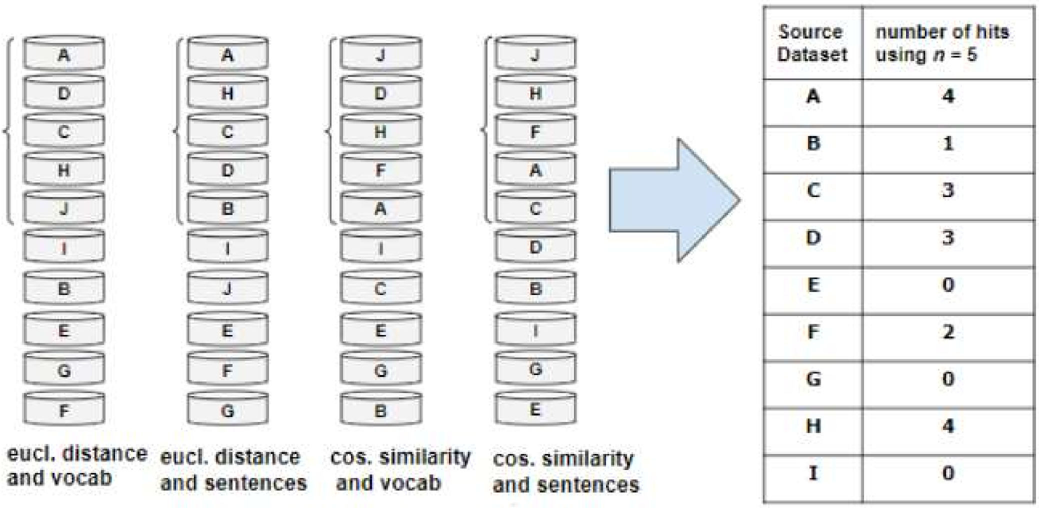
4.2.4 Unified Similarity Ranking

Source datasets are sorted according to the similarity best values. All the similarity values for each source-target dataset pair are computed using four possibilities: *Euclidean Distance based on Sentences*, *Cosine Similarity based on Sentences*, *Euclidean Distance based on Vocabulary*, and *Cosine Similarity based on Vocabulary*. Therefore, four sorted lists are created for the given target dataset, where the most similar source datasets appear on the first positions and the most dissimilar appear on the last positions in each list. A unified ranking is created based on the $Hit(n)$ method, which computes how many times each source dataset appears in the top n positions in the four sorted lists, where n is a hyperparameter. The $Hit(n)$ formula is presented in Equation 4.3.

$$Hits(n) = \frac{1}{R} \sum_{r \in R} f[r \leq n] \tag{4.3}$$

where R denotes the set of ranks for all predicted most likely conclusions, f is the indicator function – that returns 1 if the condition is true and 0 otherwise, and n refers to the number of top positions.

Therefore, datasets that are likely to achieve good results are the ones associated with the highest values in the unified ranking, whose cell values may be: 0 – source dataset does not appear within the first n positions in any of the lists ordered according to the values of similarity metrics; 1 – source dataset appears once in the first n positions; 2 – source dataset appears twice in the first n positions; 3 – source dataset appears three times in the first n positions; 4 – source dataset appears four times in the first n positions. In all the cases we are considering, they appear in the first positions of the lists ordered according to the values of similarity metrics. Figure 4.6 presents an illustration where several source datasets identified by $A, B, C, D, E, F, G, H, I, J$ are sorted according to the four proximity metrics and the unified similarity ranking in relation to a hypothetical target dataset is built. In this example, the hyperparameter n is equal to five, and the unified similarity ranking points out datasets A and H to be adopted as they are associated with the highest values in the similarity ranking.



Source Dataset	number of hits using $n = 5$
A	4
B	1
C	3
D	3
E	0
F	2
G	0
H	4
I	0

Figure 4.6: Illustration of the Hit (n) method, using $n = 5$

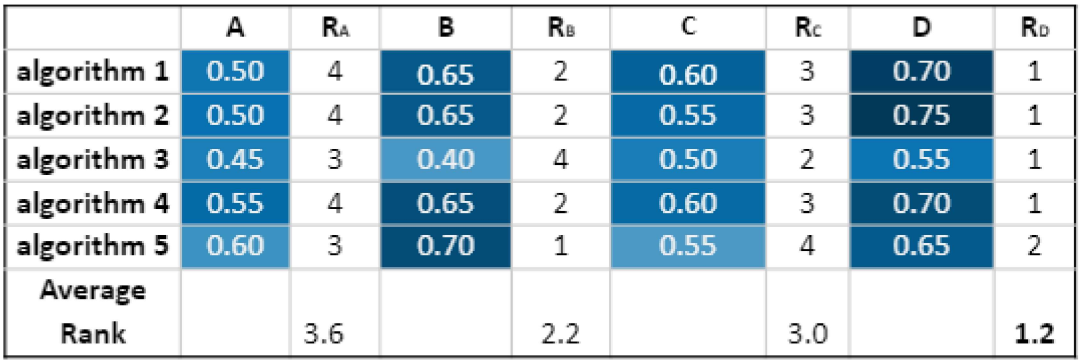
4.3 Proposed Method Validation

The proposed method validation consists of building a unified F1-score ranking considering the list of F1-score achieved with the predicted outputs returned by a list of arbitrary machine learning classifiers. The selected datasets by our method are compared to the dataset list with the best results.

The creation of the unified F1-score ranking used by our dataset selection method is detailed as follows.

Unified F1-score ranking The unified F1-score ranking is built following the idea of the Friedman [48] test rank. The Friedman test is a non-parametric statistical test that is used to detect differences in treatments across multiple test attempts. Our approach uses the source datasets as treatments and the classifiers built with the different algorithms using the source-target pairs as test attempts. It is used to check whether or not there is a statistically significant difference between the means of three or more groups. First of all, each row (algorithm) is ranked together. After that, the values of ranks are computed by averaging the values of the columns (datasets). Figure 4.7 illustrates the process for creating the unified F1-score ranking. In this example, each one of the source datasets A, B, C, D are used to train five classifiers with algorithms 1, 2, 3, 4, and 5, listed as rows in the table illustrated in Figure 4.7. The obtained F1-score values when these classifiers are applied to a given target data can be viewed for each pair of dataset-algorithm. Columns R_A, R_B, R_C, and R_D correspond to the position of each dataset in relation to a given algorithm – i.e., the one that achieved the highest F1-score value is associated with the first position (value 1), and the one that achieved the lowest F1-score value is associated with the last position (in this example, value 4 indicates the last position as there are only 4 datasets). Finally, the Average Rank row exhibits the values of the average rank of each one of the columns R_A, R_B, R_C, and R_D. Therefore, the example indicates that dataset D achieved the best F1 score as its average rank is 1.2. The second dataset best positioned in the rank is dataset B as its average rank is 2.2, followed by dataset C, which has an average rank of 3.0, and dataset A as its average rank is 3.6. The validation ranking for the example in Figure 4.7 is as follows:

- Dataset D average rank: 1.2
- Dataset B average rank: 2.2
- Dataset C average rank: 3.0
- Dataset A average rank: 3.6



	A	R_A	B	R_B	C	R_C	D	R_D
algorithm 1	0.50	4	0.65	2	0.60	3	0.70	1
algorithm 2	0.50	4	0.65	2	0.55	3	0.75	1
algorithm 3	0.45	3	0.40	4	0.50	2	0.55	1
algorithm 4	0.55	4	0.65	2	0.60	3	0.70	1
algorithm 5	0.60	3	0.70	1	0.55	4	0.65	2
Average Rank		3.6		2.2		3.0		**1.2**

Figure 4.7: Example of the process for creating the Unified F1-score rank

The Friedman test is adopted to verify if there are differences among treatments across multiple test attempts. In our proposal, each dataset is viewed as a treatment, and each machine learning algorithm is viewed as a test attempt. It is an extension of the Wilcoxon signed-rank test and the nonparametric analog of one-way repeated measures. The Friedman test checks the null hypothesis that k related variables come from the same population. The k variables are ranked for each case from 1 to k. The test statistic is based on these ranks.

The Friedman test calculus is based on a set of ranks as the ones illustrated in Figure 4.7). After the rank is built, the chi-square is computed by using Equation 4.4, where R is the mean rank, k is the number of rows (algorithms being tested as in Figure 4.7) and N the number of columns (datasets being compared as in Figure 4.7). The FF is calculated based on the chi-square value and is given by Equation 4.5. When FF value is greater than the critic value [48], the Friedman null hypothesis H_0 is rejected, what indicates that the elements being compared are not equal.

$$\chi_F{}^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{4.4}$$

$$FF = \frac{(N-1)\chi_F{}^2}{N(K-1)-\chi_F{}^2} \tag{4.5}$$

The Nemenyi [34] post-hoc test is adopted to detect groups of statistically equivalent datasets in cases where the Friedman test points out that there is a statistical difference between the set of datasets being compared. It computes a critical distance and assumes that differences between datasets are significant if the average rank difference between them is greater than the critical distance. The critical distance (CD) is given by Equation 4.6, where k is the number of rows (algorithms being tested as in Figure 4.7) and N the number of columns (datasets being compared as in Figure 4.7), and q_α is a predefined value that is given by the combination of k and the confidence value [34].

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{4.6}$$

Finally, we validate our proposal by comparing the datasets selected by our method and the ones that are in the first position of the unified F1-score ranking (or that are equivalent according to Nemenyi).

4.4 Final Considerations

In this chapter, we related our research hypothesis to the main study topics of this work. We also presented the method that will be applied to investigate our research hypothesis. The next chapter describes how the manual labeled electoral dataset related to the 2018 Presidential elections was built, shows the results of the experiments that were applied to evaluate our hypothesis, and presents a discussion about them.

Chapter 5

Experiments

This chapter presents the process for creating a manually labeled electoral dataset related to the 2018 Brazilian Presidential Elections and our experimental analysis. In addition to describe the steps for data gathering and data labeling, we also present a divergence analysis taking into account annotators disagreement and the overall distribution of labels in tweets. The experimental analysis is based on a set of monolingual and multilingual experiments. We explore two type of comparisons: intra-lingual – where only datasets that belong to the same language are compared; and inter-lingual – where datasets that belong to different languages are compared. Our experiments also include: intra-domain analysis – datasets that belong to the same domain are compared; inter-domain analysis – datasets that belong to different domains are compared. Our aim is to verify if it is possible to take advantage of similar datasets that belong to similar domains even if they belong to different languages. This is an interesting point for the electoral domain as elections are regular events that occur along the world.

5.1 2018 Brazilian Presidential Election Dataset

As previously mentioned in this thesis, one contribution of this work was to provide a manually labeled dataset of opinions written in Portuguese from social media related to the election domain[1]. We adopted Twitter social media to gather data related to the electoral scenario of the 2018 Brazilian Presidential Elections. Tweets were collected using keywords related to the election candidates and political parties. Opinions mentioning the name of at least one of the the three most mentioned candidates in the second round of the

[1] https://docs.google.com/spreadsheets/d/1JpVQ6EdFN7fvPYRtTvf0mbWhEQCqiDg7eupy_sGRBDk/edit?usp=sharing

elections: *bolsonaro*, *lula*, and *haddad* were collected, resulting in a total of 57 808 tweets. Initially, this tweets were labeled using the Microsoft Azure Sentiment Classification API[2]. This API receives as input the textual content to be analyzed and a parameter informing the language. It returns the sentiment label (positive, negative, neutral or mixed) and the sentiment score for the classes positive, negative and neutral, which is a value that varies from 0 to 1. After that, a subset of this dataset containing 406 tweets was annotated by several volunteers via an online form[3]. Each tweet was annotated by at least three annotators in the three dimensions: Sentiment Analysis (SA), Offensive Speech Presence (OS), and Candidate Support (CS).

5.1.1 Data Labeling Process

The data labeling process includes a manual labeling approach, where human judges manually analyze the electoral tweets in the three dimensions, as is better described as follows.

- SA – users are asked to inform the tweet's general sentiment (positive, negative, or neutral). In cases where the given tweet content is associated with mixed sentiments, the volunteers are instructed to inform only the predominant sentiment.
- OS Presence – users are asked to tag the electoral tweets as offensive or non-offensive. Our definition of offensive speech is that they are tweets that contain insults that aim to offend an individual or a group;
- CA Support – users are asked to inform whether the tweet contains content for or against each one of the candidates. The not-applicable label is also available to indicate that the tweet is not related to a particular candidate. Tweets were displayed randomly to the volunteer annotators in an online form, and there was no minimum or maximum limit of tweets that each annotator could label. We left tweets being labeled in the online form until they were reviewed by at least three annotators.

Although in this thesis we focus our experiments on the sentiment analysis dimension, we decided to ask the users to also annotate in the offensive speech and candidate support dimensions to enable future investigations. In addition to the sentiment analysis

[2] https://docs.microsoft.com/pt-br/rest/api/cognitiveservices-textanalytics/3.0/sentiment/sentiment
[3] `http://www.ic.uff.br/~jessicasoares/elections`

dimension, the dimensions of the presence of offensive speech and candidate support are briefly explored to investigate the degree of divergence in the labeling process. In this direction, we performed an analysis of inter-annotator divergence and entropy analysis of electoral tweets. Such analyses were useful to confirm the difficulty of labeling data in this domain and identify the main characteristics of electoral tweets that make data annotation difficult.

5.1.2 Divergence Analysis

Regarding the divergence in annotation, our purpose is measuring (1) the overall distribution of labels in tweets, leading us to measure the divergence of annotators, not considering each specific tweet or annotator; (2) the divergence of annotation per tweet considering different annotators.

1. Measuring divergence among annotators – Inter-rater Agreement: We adopted the Krippendorff's alpha (α) [76] to measure the general agreement level among the independent annotators for each one of the manual labeling tasks, namely: SA classification, OS classification, and CA classification. Krippendorff's alpha (α) agreement coefficient looks at the overall distribution of annotations/labels, not considering which annotators produced these annotations [18]. Differently from other metrics such as Cohen's Kappa [31] (which computes the agreement level between a pair of annotators) and Fleiss' Kappa [45] (which is a generalization of Cohen Kappa and allows more than two annotators), the metric Krippendorff's alpha α can be applied to evaluate labeling agreement among multiple annotators even when there are missing values. Allowing missing values in the annotations is particularly important to our experiments, as voluntary annotators who manually labeled electoral tweets were not required to label the same subset of tweets. Instead, they were asked to annotate a random subset of tweets without requiring a minimum or a maximum number of annotations. We adopted this procedure to maximize and diversify the number of labeled instances. The responses of all annotators for a single example is called a unit. The metric α is given by Equation 5.1, where D_o is the observed disagreement among values assigned to units of analysis and D_e is the disagreement one would expect when the coding of units is attributable to chance rather than to the properties of these units [76]. Both D_o and D_e are computed based on the frequencies of values in coincidence matrices. In a scenario where annotators perfectly agree, $D_o = 0$ and $\alpha = 1$ but when there is a complete disagreement $\alpha = 0$.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{5.1}$$

2. Measuring divergence in tweets – Labeling Entropy Analysis: In order to try to identify how much the annotators disagree with each tweet and possible reasons why this happens, we made an analysis based on the number of labels each tweet received for each class in each of the classification tasks. We adopted the concept of Entropy from Information Theory [126], which states that Entropy from a random variable is the average level of "information", "surprise" or "uncertainty" in the variable's possible outcomes. Given a random variable X with possible outcomes x_1, x_2, ..., x_n, which occur with probability $P(x_i)$, Entropy $H(X)$ is calculated by Equation 5.2. In our case, each tweet is X and the possible outcomes x_1, x_2, ..., x_n are {"positive", "negative", "neutral"} for the SA task; {"offensive", "non-offensive"} for the OS detection task; and {"for", "against", "not applicable"} for the CA task. In a scenario where all annotators agree with all labels of a task for a given instance, entropy $H(X) = 0$. In this way, the higher the entropy, the higher the annotation divergence.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{5.2}$$

The exploratory analysis of the divergence in the labeling process of electoral opinions is presented in detail in [119]. In short, we measured the divergence among annotators – also called inter-rater agreement – using the Krippendorff's alpha (α) coefficient [76] and the divergence in each tweet, by calculating the entropy. An alpha value equal to 1 indicates complete agreement, and an alpha value equal to 0 indicates complete disagreement among annotators. The analysis presented in such study [119] proves the great difficulty in labeling electoral data extracted from social media regarding SA and OS, as we obtained $\alpha = 0.39$ and $\alpha = 0.54$, respectively. In addition, the task of analyzing whether a tweet is "for", "against" or "not applicable" concerning the candidate Bolsonaro was the one that obtained the highest degree of agreement among the annotators ($\alpha =$ 0.85). This information is summarized in Table 5.1.

	Sentiment Analysis (SA)	Offensive Speech (OS)	Candidate Analysis (CA) Support		
			Lula	Haddad	Bolsonaro
α	0.39	0.54	0.70	0.71	0.85

Table 5.1: Krippendorff's alpha (α) agreement coefficient

We also performed an entropy [126] analysis in such study [119] to measure divergence in each tweet, specifically. This analysis is illustrated in Figures 5.1 and 5.2. From that, we can identify the tweets that are related to a *high level of confusion in the labeling process*. As we obtained high α for SA and OS dimensions, we selected the tweets associated with the top 15 highest entropy values to identify the main reasons that may lead to high levels of confusion in SA and OS detection tasks and, consequently, label divergence.

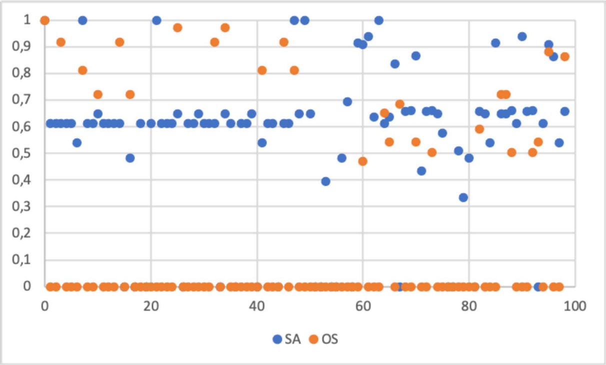


Figure 5.1: Entropy Analysis – SA and OS dimensions

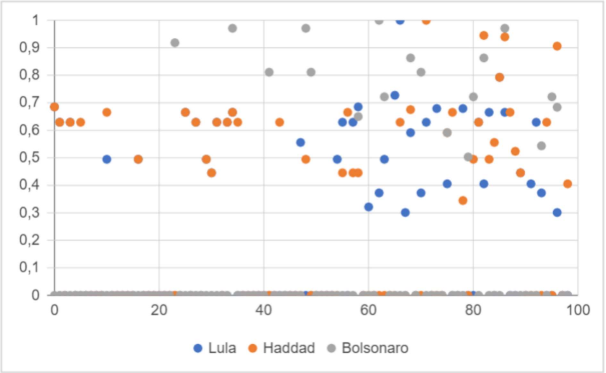


Figure 5.2: Entropy Analysis – CA dimension

We observed some common characteristics of these tweets associated with high entropy that we believe may explain the high level of complexity of labeling them, as follows:

- **non-textual content**: tweets that, in addition to textual content, also use links to external content like news, images, or gifs to express their opinions, which may lead to not being possible to infer the correct sentiment by looking only at the textual content of the tweet;

- **irony or humor**: tweets containing jokes or ironic opinions about elections;

- **external knowledge**: tweets that mention facts that occurred before or during electoral campaigns, which may require external knowledge about the political context to understand the real intention of the opinion;

- **negative content and support hashtag**: tweets that denote a predominant negative sentiment but are full of hashtags in favor of a given political candidate;

- **neutral content and support hashtag**: tweets that denote a neutral sentiment but are full of hashtags in favor of a given political candidate;
- **mixed sentiment**: tweets containing both positive and negative opinions related to different entities, such as tweets where the user supports one candidate and rejects other entities (whether they are other candidates or even a particular population group).

5.2 Experimental Analysis

We present in this section the evaluation steps adopted to validate our method and our obtained results, considering the electoral scenario as our domain of interest. Although the proposed dataset selection method is unsupervised and does not require target dataset labels, we had to consider labeled target datasets in this experimental analysis to validate our results.

Datasets: We began by searching for existing labeled datasets for the sentiment analysis task in public repositories to be used in our experiments. This search for publicly available datasets was conducted in repositories such as Kaggle[4], GitHub[5], and Google Dataset Search[6]. Three languages were considered: Portuguese, English, and Spanish. We selected Portuguese as it is our mother language, English because of the large available resources and the similarity of the 2016 USA electoral campaign and 2018 Brazilian campaign, and Spanish because it is a language with the same roots as Portuguese while vastly spoken in Latin America. We also searched for datasets that belong to the electoral domain to analyze this scenario as the target. The 2018 Brazilian Presidential Election dataset (2018 BR Election PT) was the only dataset used in our experiments whose data was gathered specifically for this thesis (as discussed in Section 5.1). A total of 16 datasets[7] was considered in our experiments, as follows:

1. *2018 BR Election PT* – posts on Twitter related to the 2018 Brazilian presidential elections;
2. *Restaurants PT* – opinions about Brazilian restaurants extracted from Foursquare;
3. *2016 US Election EN* – posts on Twitter about the 2016 United States presidential elections;

[4] kaggle.com

[5] https://github.com/

[6] https://datasetsearch.research.google.com/

[7] `https://docs.google.com/spreadsheets/d/1Olp_gFMwiXLY7QIGroU641jYKF5J5jiT/edit?usp=sharing&ouid=110616407102804776423&rtpof=true&sd=true`

4. *GOP Debate EN*: Twitter posts in English about a political presidential debate that occurred in 2016;
5. *2012 US Election EN*: Twitter opinions about the 2012 United States presidential elections;
6. *TV PT*: Twitter opinions about Brazilian TV shows;
7. *Music Festival EN*: opinions posted on Twitter about a US music festival;
8. *Urban Problems PT*: Twitter opinions about urban problems in Minas Gerais, a Brazilian state.
9. *Airlines EN*: Twitter opinions written in English about airlines;
10. *Movies 1 EN*: movie reviews in English;
11. *Movies PT*: movie reviews in Portuguese;
12. *Movies 2 EN*: movie reviews in English;
13. *Apple EN*: Twitter opinions written in English about Apple products;
14. *Airlines ES*: Twitter opinions about Spanish airlines;
15. *2018 CO Election ES*: Twitter opinions about the 2018 Colombian presidential elections;
16. *Sports ES*: Twitter opinions about sports in Spanish.

After selecting the datasets, we performed the preprocessing steps where special characters and stop words were removed. Also, we discarded instances associated with the class *neutral* since some of these datasets only have *positive* and *negative* instances. Details about these datasets can be viewed in Table 5.2.

Table 5.2: Datasets Info

ID	**Name**	**Positive**	**Neutral**	**Negative**	**Total**
1	2018 BR Election PT [8]	11205	20210	26392	57807
1	2018 BR Election PT[9]	136	108	162	406
2	Restaurants PT	902	0	886	1788
3	2016 US Election EN	4793	2677	3006	10476
4	GOP Debate EN	671	0	3082	3753
5	2012 US Election EN	285183	539840	174977	1000000
6	TV PT	4793	2677	3006	10476
7	Music Festival EN	2282	928	553	3763
8	Urban Problems PT	103	2453	345	2901
9	Airlines EN	2363	3098	9178	14639
10	Movies 1 EN	32584	0	31068	63652
11	Movies PT	24522	0	24522	49044
12	Movies 2 EN	24999	0	25000	49999
13	Apple EN	422	2161	1219	3802
14	Airlines ES	1489	2609	3769	7867
15	2018 CO Election ES	20979	0	12255	33234
16	Sports ES	11004	0	9489	20493

Considering that datasets that belong to similar domains can be found in different

[8] This refers to the automatic labeled version of this dataset.

[9] This refers to the manual labeled version of this dataset.

languages we explore a multilingual analysis in our experiments. We use the Universal Sentence Encoder Multilingual[10] [29,153], a pre-trained cross-lingual model that was trained on 16 languages and can embed text from these different languages in a single vector space. The multilingual embeddings that this model provides allow us to compare and use datasets belonging to different languages.

Also, we execute experiments where the proposed dataset selection method is applied to only consider datasets that belong to the same language. In these experiments, monolingual embeddings are adopted for each of the three languages (Portuguese, Spanish, and English), and the dataset comparison step is restricted to datasets belonging to the same language. The choice of the models was taken based on the available pretrained models for each language in the Hugging Face [138] website, excluding the ones that are tagged as deprecated. Finally, models BETO [24] (spanish), BERTimbau [131] (portuguese), and T5 [91] (english) were select.

Dataset similarity rankings: Semantic similarity was computed between each source-target dataset pair according to the similarity metrics presented in Section 4.1. While all the datasets will serve as source data, only those related to the electoral domain are selected as targets since this is our domain of interest. The code of our dataset selection method[11] and the manually labeled dataset[12] provided by this research are available online.

5.2.1 Monolingual Experiments

We conducted monolingual experiments using data from three languages that were analyzed independently: Spanish, English, and Portuguese. We tested all pairs of source-target datasets with a set of five traditional machine learning algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Multi-Layer Perceptron (MLP), and XGBoost.

Spanish: We performed an analysis only considering datasets written in Spanish and Spanish embeddings gathered from a Hugging Face [138] model[13] [24] called BETO. Results are summarized in Tables 5.3 – 5.7, cells with the strongest color represent the best results. We can notice that according to all distance/ similarity metrics, the *Sports ES* is the closest dataset. A previous study [109] shows that the polarization level of

[10] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

[11] https://github.com/sjessicasoaress/ds_recommender

[12] https://docs.google.com/spreadsheets/d/1JpVQ6EdFN7fvPYRtTvfOmbWhEQCqiDg7eupy_sGRBDk

[13] https://huggingface.co/hiiamsid/sentence_similarity_spanish_es

soccer and political datasets are quite similar, we believe that this factor may explain this experimental result. This dataset was the one that achieved the best F1 scores for all the classifiers built with different algorithms.

Table 5.3: Cosine Similarity based on Sentences - Spanish

	2018 CO Election ES
Airlines ES	0.6551
Sports ES	0.7651

Table 5.4: Euclidean Distance based on Sentences - Spanish

	2018 CO Election ES
Airlines ES	6.3963
Sports ES	5.5541

Table 5.5: Cosine Similarity based on Vocabulary - Spanish

	2018 CO Election ES
Airlines ES	0.9582
Sports ES	0.9743

Table 5.6: Euclidean Distance based on Vocabulary - Spanish

	2018 CO Election ES
Airlines ES	3.2816
Sports ES	2.5845

Table 5.7: F1-score Summary - Spanish

2018 CO Election ES	**Airlines ES**	**Sports ES**
SVM	0.6082	0.6567
Logistic Regression	0.6193	0.6698
Decision Tree	0.5606	0.5953
MLP	0.6212	0.6723
XGBoost	0.5840	0.6719

Portuguese: We performed an analysis of the datasets written in Portuguese using Portuguese embeddings gathered from a Hugging Face [138] model[14] [131] called BERTimbau. Results are summarized in Tables 5.8 – 5.14. This analysis considered the labels that were manually assigned to the 2018 BR Election dataset. According to the analyzes based on Vocabulary, the closest dataset is the Movies PT. Tables 5.12 and 5.13 present the selected datasets using the Hit(n) method using n = 1, and n = 2, respectively. On the other hand, by analyzing the average of datasets sentences, the closest dataset is the TV PT. Table 5.14 shows the F1-score summary.

[14] https://huggingface.co/neuralmind/bertbaseportuguesecased

Table 5.8: Cosine Similarity based on Sentences - Portuguese

	2018 BR Election PT
Restaurants PT	0.8510
TV PT	0.9372
Urban Problems PT	0.9273
Movies PT	0.8543

Table 5.9: Euclidean Distance based on Sentences - Portuguese

	2018 BR Election PT
Restaurants PT	3.3650
TV PT	2.1537
Urban Problems PT	2.3176
Movies PT	3.5498

Table 5.10: Cosine Similarity based on Vocabulary - Portuguese

	2018 BR Election PT
Restaurants PT	0.9727
TV PT	0.9894
Urban Problems PT	0.9878
Movies PT	0.9912

Table 5.11: Euclidean Distance based on Vocabulary - Portuguese

	2018 BR Election PT
Restaurants PT	1.4137
TV PT	0.8861
Urban Problems PT	0.9483
Movies PT	0.8215

Table 5.12: Unified Similarity Ranking, $n = 1$ - Portuguese

	2018 BR Election PT
Restaurants PT	0
TV PT	2
Urban Problems PT	0
Movies PT	2

Table 5.13: Unified Similarity Ranking, $n = 2$ - Portuguese

	2018 BR Election PT
Restaurants PT	0
TV PT	4
Urban Problems PT	2
Movies PT	2

Table 5.14: F1-score Summary - Portuguese

	Restaurants PT	**TV PT**	**Urban Problems PT**	**Movies PT**
SVM	0.5238	0.6372	0.5994	0.7010
LR	0.5384	0.6372	0.5645	0.7220
DT	0.4994	0.4252	0.5033	0.5277
MLP	0.5624	0.6623	0.5698	0.6631
XGBoost	0.4412	0.6677	0.4177	0.6304

From Table 5.14, the validation ranking obtained is as follows:

- Movies PT average rank: 1.2,
- TV PT average rank: 2.2,
- Urban Problems PT average rank: 3.0,
- Restaurants PT average rank: 3.6.

As we can notice, the most similar datasets (Movies PT and TV PT) were the ones that appeared in the first and second positions, respectively, in the validation ranking. Those datasets are also the selected datasets by the Hit(n) method. When adopting $n = 2$, the TV PT dataset is selected. When $n = 1$ is adopted, both datasets TV PT and Movies PT are selected as illustrated in Tables 5.12 and 5.13.

According to the Friedman test, there is a statistical difference as the *pvalue* was 0.0071, which is a value $<$ than 0.05. The Nemenyi test was applied, and the critical difference found was 1.8709, using $\alpha = 0.1$. We can observe from our analysis that datasets Movies PT, TV PT, and Urban Problems had equivalent results. We believe that the dataset related to restaurant reviews may be less similar to the others due to the magnitude of the sentiment and engagement associated with this type of context compared to the others. However, an analysis of the magnitude of sentiment in each context is outside the scope of this thesis but can be addressed later in future studies.

English: We performed an analysis of the datasets written in English using English embeddings gathered from a Hugging Face [138] model[15] [91]. Results are summarized in

[15] https://huggingface.co/sentencetransformers/sentencet5large

Tables 5.15 – 5.22. By looking at the similarity and distance metrics based on sentences, the closest dataset is the GOP Debate EN for both the 2016 US Election EN and 2012 US Election EN datasets. On the other hand, by looking at the similarity and distance metrics based on the vocabulary, the closest dataset is the 2012 US Election EN for the 2016 US Election EN and vice versa. Therefore, vocabulary based analysis seems to be better for T5 embeddings because it points to the datasets that are at the top of the ranking. However, it should be emphasized that other factors related to the context of datasets can be considered in choosing the most appropriate dataset. For example, even though it is the closest to the 2012 US Election dataset, it may not be advantageous to adopt the 2016 US Election dataset as a pre-trained dataset since it refers to a later election.

Table 5.15: Cosine Similarity based on Sentences - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	0.9533
GOP Debate EN	0.9579	0.9673
2012 US Election EN	0.9533	-
Music Festival EN	0.8599	0.8654
Airlines EN	0.9132	0.9237
Movies 1 EN	0.9302	0.9378
Movies 2 EN	0.8682	0.8767
Apple EN	0.9262	0.9343

Table 5.16: Euclidean Distance based on Sentences - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	0.2556
GOP Debate EN	0.2414	0.2128
2012 US Election EN	0.2556	-
Music Festival EN	0.4584	0.4493
Airlines EN	0.3478	0.3259
Movies 1 EN	0.3103	0.2927
Movies 2 EN	0.4489	0.4344
Apple EN	0.3193	0.3013

Table 5.17: Cosine Similarity based on Vocabulary - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	0.9980
GOP Debate EN	0.9970	0.9974
2012 US Election EN	0.9980	-
Music Festival EN	0.9936	0.9951
Airlines EN	0.9945	0.9961
Movies 1 EN	0.9941	0.9957
Movies 2 EN	0.9952	0.9971
Apple EN	0.9937	0.9953

Table 5.18: Euclidean Distance based on Vocabulary - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	0.0558
GOP Debate EN	0.0679	0.0629
2012 US Election EN	0.0558	-
Music Festival EN	0.0999	0.0875
Airlines EN	0.0926	0.0782
Movies 1 EN	0.0968	0.0830
Movies 2 EN	0.0861	0.0664
Apple EN	0.0988	0.0854

Table 5.19: Unified Similarity Ranking, $n = 1$ - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	2
GOP Debate EN	2	2
2012 US Election EN	2	-
Music Festival EN	0	0
Airlines EN	0	0
Movies 1 EN	0	0
Movies 2 EN	0	0
Apple EN	0	0

Table 5.20: Unified Similarity Ranking, $n = 2$ - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	4
GOP Debate EN	4	4
2012 US Election EN	4	-
Music Festival EN	0	0
Airlines EN	0	0
Movies 1 EN	0	0
Movies 2 EN	0	0
Apple EN	0	0

Table 5.21: Unified Similarity Ranking, $n = 3$ - English

	2016 US Election EN	**2012 US Election EN**
2016 US Election EN	-	4
GOP Debate EN	4	4
2012 US Election EN	4	-
Music Festival EN	0	0
Airlines EN	0	0
Movies 1 EN	2	2
Movies 2 EN	2	2
Apple EN	0	0

Table 5.22: F1-score Summary - English - US 2016

	GOP Debate EN	**2012 US Election EN**	**Music Festival EN**	**Airlines EN**	**Movies 1 EN**	**Movies 2 EN**	**Apple EN**
SVM	0.5988	0.6681	0.5877	0.5908	0.5606	0.5834	0.5811
LR	0.6063	0.6499	0.6106	0.6019	0.5901	0.6287	0.5735
DT	0.5567	0.5632	0.5225	0.5359	0.526	0.5261	0.5423
MLP	0.6065	0.6403	0.556	0.5688	0.5048	0.5953	0.5739
XGBoost	0.5785	0.6016	0.6364	0.6129	0.5742	0.5995	0.5691

Table 5.23: F1-score Summary - English - US 2012

	2016 US Election EN	**GOP Debate EN**	**Music Festival EN**	**Airlines EN**	**Movies 1 EN**	**Movies 2 EN**	**Apple EN**
SVM	0.6231	0.5899	0.5984	0.6281	0.5339	0.5751	0.5614
LR	0.6415	0.5923	0.5942	0.6275	0.5679	0.5927	0.5561
DT	0.4908	0.5868	0.4704	0.5349	0.5532	0.5047	0.5161
MLP	0.6361	0.577	0.5755	0.5968	0.4862	0.5733	0.5656
XGBoost	0.6405	0.5542	0.5932	0.647	0.502	0.5666	0.5453

The validation ranking of F1-scores for the 2016 US Election dataset EN is as follows:

- 2012 US Election EN average rank: 1.4
- GOP Debate EN average rank: 3.0
- Movies 2 EN and Airlines EN average rank: 3.8
- Music Festival EN average rank: 4.2
- Apple EN average rank: 5.4
- Movies 1 EN average rank: 6.4

The validation ranking of F1-scores for the 2012 US Election dataset EN is as follows:

- Airlines EN average rank: 1.8
- 2016 US Election EN average rank: 2.4

- GOP Debate EN average rank: 3.6
- Music Festival EN average rank: 4.0
- Movies 2 EN average rank: 4.6
- Movies 1 EN and Apple EN average rank: 5.8

We can observe that in all the cases, our method selected the 2016 US Election EN and GOP Debate EN for the 2012 US Election, and the 2012 US Election EN and the GOP Debate EN for the 2016 US Election EN. The Nemenyi critical distance (CD) for both cases is 3.6790, using $\alpha = 0.1$. From the results, our dataset selection method was able to select datasets that are in the three top positions in the validation ranking for n = 1, n = 2, and n = 3. Also, both the 2016 US Election EN and GOP Debate EN may be considered equivalent to the Airlines EN dataset in the 2012 US Election validation ranking.

5.2.2 Multilingual Experiments

Two main experiments were executed to analyze our approach using multilingual embeddings. The main difference between them is that the manual labels for the 2018 BR Election PT datasets were unavailable in the first experiments, so we automatically annotated the 2018 BR Election PT dataset using the Microsoft Azure sentiment analysis labeling method. The results of this experiment were published in [122]. In the second experiment, we adopted the subset of the 2018 BR Election dataset that was labeled with the help of a set of volunteers. A comparative analysis about the manual labeling and the automatic labeling approach is presented in [119].

Experiment 1: Tables 5.24, 5.25, 5.26, and 5.27, present the results for the *Euclidean Distance based on Sentences*, *Cosine Distance based on Sentences*, *Euclidean Distance based on Vocabulary* and *Cosine Distance based on Vocabulary* metrics, respectively. The columns of these tables represent target datasets, and rows represent source datasets. The best similarity values are highlighted with stronger colors, while the worst are highlighted with lighter colors. The unified similarity ranking was built according to the *Hit*(n) method adopting n = *5*. Therefore, each table cell value corresponds to the number of hits, i.e., how many times a given source-target dataset is at the top-five first positions according to the similarity results presented in Tables 5.24, 5.25, 5.26 and 5.27. Table 5.28 illustrates the similarity ranking as a heat map. Columns represent target datasets, and rows represent source datasets. Then, for each column, we can see selected source datasets for a given target dataset as being the ones with the strongest color.

Table 5.24: Euclidean Distance based on Sentences - Experiment 1

	2018 BR	2016 US	2012 US	2018 CO
	Election	Election	Election	Election
	PT	EN	EN	ES
2018 BR Election PT	-	0.5227	0.5254	0.4133
Restaurants PT	0.6374	0.6332	0.6417	0.4945
2016 US Election EN	0.5227	-	0.4517	0.4208
GOP Debate EN	0.5309	0.4568	0.4254	0.4910
2012 US Election EN	0.5254	0.4517	-	0.4384
TV PT	0.5648	0.5892	0.5644	0.4083
Music Festival EN	0.8554	0.8170	0.7911	0.7268
Urban Problems PT	0.5756	0.5811	0.5647	0.4842
Airlines EN	0.6392	0.5633	0.5532	0.4597
Movies 1 EN	0.6223	0.5467	0.5447	0.4408
Movies PT	0.6386	0.6062	0.5867	0.4787
Movies 2 EN	0.6695	0.6098	0.5893	0.5133
Apple EN	0.6039	0.5276	0.5093	0.4357
Airlines ES	0.5897	0.5810	0.5819	0.4252
2018 CO Election ES	0.4133	0.4208	0.4384	-
Sports ES	0.5897	0.5789	0.5643	0.3925

Table 5.25: Cosine Similarity based on Sentences - Experiment 1

	2018 BR	2016 US	2012 US	2018 CO
	Election	Election	Election	Election
	PT	EN	EN	ES
2018 BR Election PT	-	0.5677	0.5651	0.7107
Restaurants PT	0.2780	0.1911	0.1747	0.3527
2016 US Election EN	0.5677	-	0.6415	0.6267
GOP Debate EN	0.5899	0.6692	0.7148	0.5598
2012 US Election EN	0.5651	0.6415	-	0.5957
TV PT	0.4351	0.2937	0.3576	0.5536
Music Festival EN	0.1527	0.1715	0.2291	0.2607
Urban Problems PT	0.4670	0.3926	0.4299	0.4792
Airlines EN	0.2253	0.3147	0.3456	0.3764
Movies 1 EN	0.2397	0.3283	0.3394	0.3868
Movies PT	0.2569	0.2377	0.2923	0.3672
Movies 2 EN	0.1984	0.2485	0.3039	0.2990
Apple EN	0.2829	0.3718	0.4239	0.3917
Airlines ES	0.3533	0.2747	0.2780	0.4723
2018 CO Election ES	0.7107	0.6267	0.5957	-
Sports ES	0.3901	0.3306	0.3689	0.6001

Table 5.26: Euclidean Distance based on Vocabulary - Experiment 1

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.0831	0.0941	0.0555
Restaurants PT	0.1168	0.1259	0.1092	0.1124
2016 US Election EN	0.0831	-	0.0484	0.0912
GOP Debate EN	0.1078	0.0587	0.0504	0.1024
2012 US Election EN	0.0941	0.0484	-	0.0932
TV PT	0.0858	0.1047	0.0856	0.0848
Music Festival EN	0.1541	0.1208	0.0857	0.1453
Urban Problems PT	0.0906	0.0963	0.0978	0.0906
Airlines EN	0.1216	0.0802	0.0674	0.1152
Movies 1 EN	0.1115	0.0730	0.0830	0.1091
Movies PT	0.0669	0.0686	0.0735	0.0706
Movies 2 EN	0.1008	0.0654	0.0724	0.1066
Apple EN	0.1527	0.1136	0.0935	0.1428
Airlines ES	0.0970	0.1085	0.0932	0.0814
2018 CO Election ES	0.0555	0.0912	0.0932	-
Sports ES	0.0680	0.0881	0.0797	0.0595

Table 5.27: Cosine Similarity based on Vocabulary - Experiment 1

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.9905	0.9878	0.9961
Restaurants PT	0.9812	0.9783	0.9837	0.9829
2016 US Election EN	0.9905	-	0.9968	0.9887
GOP Debate EN	0.9838	0.9952	0.9965	0.9857
2012 US Election EN	0.9878	0.9968	-	0.9882
TV PT	0.9901	0.9851	0.9901	0.9903
Music Festival EN	0.9677	0.9803	0.9901	0.9716
Urban Problems PT	0.9885	0.9872	0.9868	0.9889
Airlines EN	0.9791	0.9911	0.9938	0.9819
Movies 1 EN	0.9829	0.9927	0.9906	0.9839
Movies PT	0.9938	0.9935	0.9926	0.9933
Movies 2 EN	0.9859	0.9941	0.9928	0.9846
Apple EN	0.9673	0.9822	0.9880	0.9722
Airlines ES	0.9868	0.9837	0.9880	0.9910
2018 CO Election ES	0.9961	0.9887	0.9882	-
Sports ES	0.9940	0.9895	0.9914	0.9952

Table 5.28: Unified Similarity Ranking Summary, $n = 5$ - Experiment 1

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	2	2	4
Restaurants PT	0	0	0	0
2016 US Election EN	4	-	4	2
GOP Debate EN	2	4	4	1
2012 US Election EN	2	4	-	1
TV PT	3	0	0	3
Music Festival EN	0	0	0	0
Urban Problems PT	1	1	1	0
Airlines EN	0	0	2	0
Movies 1 EN	0	2	0	0
Movies PT	2	2	2	2
Movies 2 EN	0	2	2	0
Apple EN	0	1	1	0
Airlines ES	0	0	0	3
2018 CO Election ES	4	2	2	-
Sports ES	2	0	0	4

Table 5.29: F1-score Summary - Experiment 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 - SVM	-	0.72	0.72	0.58	0.72	0.60	0.63	0.46	0.75	0.47	0.56	0.63	0.70	0.72	0.73	0.78
1 - LR	-	0.72	0.73	0.57	0.74	0.61	0.59	0.46	0.75	0.44	0.57	0.61	0.73	0.76	0.72	0.76
1 - DT	-	0.52	0.61	0.54	0.57	0.62	0.5	0.48	0.58	0.50	0.56	0.54	0.57	0.65	0.62	0.63
1 - MLP	-	0.69	0.66	0.51	0.71	0.62	0.48	0.47	0.66	0.48	0.61	0.63	0.65	0.70	0.67	0.77
1 - XGBoost	-	0.73	0.75	0.67	0.72	0.67	0.57	0.5	0.72	0.54	0.66	0.63	0.66	0.72	0.74	0.72
3 - SVM	0.66	0.53	-	0.63	0.68	0.59	0.6	0.52	0.60	0.54	0.63	0.63	0.65	0.61	0.58	0.63
3 - LR	0.66	0.53	-	0.63	0.69	0.55	0.56	0.53	0.62	0.55	0.64	0.63	0.65	0.60	0.60	0.61
3 - DT	0.55	0.5	-	0.55	0.56	0.53	0.51	0.49	0.54	0.51	0.51	0.52	0.54	0.54	0.53	0.56
3 - MLP	0.63	0.49	-	0.61	0.68	0.51	0.57	0.54	0.61	0.51	0.55	0.52	0.63	0.57	0.58	0.63
3 - XGBoost	0.62	0.49	-	0.62	0.69	0.53	0.58	0.54	0.60	0.54	0.62	0.6	0.59	0.61	0.62	0.60
5 - SVM	0.64	0.58	0.71	0.55	-	0.65	0.58	0.57	0.67	0.58	0.58	0.57	0.65	0.61	0.55	0.68
5 - LR	0.63	0.60	0.70	0.51	-	0.6	0.57	0.56	0.67	0.53	0.59	0.57	0.68	0.63	0.55	0.68
5 - DT	0.59	0.49	0.62	0.52	-	0.52	0.54	0.53	0.52	0.47	0.52	0.53	0.6	0.55	0.57	0.54
5 - MLP	0.6	0.54	0.66	0.55	-	0.61	0.53	0.59	0.63	0.54	0.55	0.59	0.63	0.58	0.56	0.63
5 - XGBoost	0.61	0.55	0.69	0.54	-	0.56	0.57	0.61	0.62	0.56	0.59	0.63	0.60	0.58	0.58	0.65
15 - SVM	0.67	0.62	0.61	0.56	0.61	0.65	0.64	0.57	0.67	0.53	0.59	0.60	0.65	0.63	-	0.66
15 - LR	0.67	0.63	0.63	0.57	0.62	0.65	0.63	0.55	0.68	0.53	0.60	0.61	0.66	0.65	-	0.67
15 - DT	0.59	0.54	0.59	0.52	0.55	0.58	0.51	0.53	0.60	0.49	0.53	0.53	0.56	0.57	-	0.58
15 - MLP	0.65	0.59	0.62	0.51	0.58	0.63	0.60	0.57	0.63	0.50	0.55	0.57	0.63	0.62	-	0.66
15 - XGBoost	0.67	0.62	0.66	0.62	0.64	0.65	0.63	0.56	0.68	0.53	0.60	0.59	0.65	0.65	-	0.66

We tested all pairs of source-target datasets with a set of five traditional machine learning algorithms, using the same pre-trained embeddings employed during the similarity analysis: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Multi-Layer Perceptron (MLP), and XGBoost. Four of these sixteen datasets belong to this experimental analysis's target domain (2018 BR Election PT, 2016 US Election EN, 2012 US Election EN, 2018 CO Election ES). For each of the five machine learning algorithms, we trained different classifiers using each one of the source datasets individually as training data. The classifiers were applied to the four electoral datasets, individually. Table 5.29 presents the F1-score results of each classifier. Datasets are identified by a number instead of their name. The numeric identifier is the same as the one used before: 1 – corresponds to the 2018 BR Election PT, 3 – corresponds to the 2016 US Election EN, 5 – corresponds to the 2012 US Election EN, and 15 – corresponds to the 2018 CO Election ES. Table 5.29 columns refer to the identifier of the source dataset used to train classifiers. Rows refer to the identifier of the target dataset followed by the abbreviation of the machine learning algorithm. The best F1-score values are highlighted with stronger colors, while the worst are highlighted with lighter colors. For each combination of target dataset-algorithm, the obtained F1-score classification results – when using the different source datasets as training data – are stored in a list that is sorted according to F1-score best values.

The validation ranking for the 2018 BR Election PT - Experiment 1 is as follows:

- Sports ES average rank: 2.4
- Airlines ES average rank: 3.6
- 2018 CO Election ES average rank: 4.5
- 2016 US Election EN e Airlines EN average rank: 5.0
- 2012 US Election EN average rank: 5.2
- Restaurants PT average rank: 6.9
- Apple EN average rank: 8.5
- TV PT average rank: 9.3
- Movies 2 EN average rank: 10.9
- GOP Debate EN average rank: 11.9
- Movies PT average rank: 12.0
- Music Festival EN average rank: 13.1

- Movies 1 EN average rank: 15.0
- Urban Problems PT average rank: 15.8

We can observe that the datasets selected by our method, namely, 2016 US Election EN and 2018 CO Election ES, have the third and fourth best positions in the validation ranking.

The validation ranking for the 2016 US Election EN - Experiment 1 is as follows:

- 2012 US Election EN average rank: 1.1
- 2018 BR Election PT average rank: 3.0
- GOP Debate EN average rank: 5.1
- Apple EN average rank: 5.4
- Sports ES average rank: 5.5
- Movies PT average rank: 7.6
- Airlines EN average rank: 7.9
- Airlines ES average rank: 8.5
- 2018 CO Election ES average rank: 8.9
- Movies 2 EN average rank: 9.8
- Music Festival EN average rank: 11.4
- TV PT average rank: 12.9
- Movies 1 EN average rank: 13.7
- Urban Problems PT average rank: 14.6
- Restaurants PT average rank: 15.5

We can observe that the datasets selected by our method, namely, 2012 US Election EN and GOP Debate EN, have the first and third best positions in the validation ranking.

The validation ranking for the 2012 US Election EN - Experiment 1 is as follows:

- 2016 US Election EN average rank: 1.0
- Sports ES average rank: 3.2
- Apple EN average rank: 3.8

- 2018 BR Election PT and Airlines EN average rank: 5.2
- Airlines ES average rank: 7.2
- TV PT average rank: 8.3
- Movies 2 EN average rank: 8.4
- Urban Problems PT average rank: 9.2
- Movies PT average rank: 10.1
- 2018 CO Election ES average rank: 10.3
- Music Festival EN average rank: 10.7
- Restaurants PT average rank: 12.1
- Movies 1 EN average rank: 13.3
- GOP Debate EN average rank: 14.0

We can observe that the datasets selected by our method, namely, 2016 US Election EN and GOP Debate EN appear in the first and the last positions in the validation ranking, respectively.

The validation ranking for the 2018 CO Election PT - Experiment 1 is as follows:

- Airlines EN average rank: 1.7
- 2018 BR Election PT average rank: 2.1
- Sports ES average rank: 2.9
- TV PT average rank: 5.1
- Apple EN average rank: 5.4
- 2016 US Election EN average rank: 6.5
- 2018 CO Election ES average rank: 6.7
- Airlines ES average rank: 6.7
- Music Festival EN average rank: 9.8
- Restaurants PT average rank: 9.9
- 2012 US Election EN average rank: 10.1
- Movies 2 EN average rank: 12.5
- Movies PT average rank: 13.0

- Urban Problems PT average rank: 13.7
- GOP Debate EN average rank: 13.9
- Movies 1 EN average rank: 16.0

We can observe that the datasets selected by our method, namely, 2018 BR Election PT and Sports ES appear in the second and the third positions in the validation ranking, respectively.

For all the cases, the Nemenyi critical distance is CD: 8.9356. According to the Nemenyi test, differences between datasets are significant if the average rank difference between them is greater than the critical distance. We can notice that our method selected datasets that are in the first positions in the validation ranking or are significantly equivalent, except for the GOP Debate EN dataset when it was selected to the 2012 US Election EN. We believe this may occur when the datasets have similar content but there is a high polarity divergence between them. This would be a limitation of our method, although, in practice, it is not possible to measure polarity divergence in cases where there is no data annotated in the target domain as target labels are not available. We noticed that this dataset was also selected by other dataset selection approaches (namely, RCA [40, 42, 158] and LM [33]), as is better explained in Section 5.3. Also, our method was the one that provided the best selections when compared to other dataset selection approaches (see Section 5.3).

The results of Experiment 1 were published in [122][16].

Experiment 2: Our approach using multilingual embeddings was also tested in the manually labeled version of the 2018 BR Election dataset. After performing the preprocessing steps, datasets were balanced, and the distance and similarity metrics were computed, as can be viewed in Tables 5.31 – 5.32. After that, we built the unified similarity ranking by considering different values for the n, as illustrated in Tables 5.34 – 5.38. We tested values for n from 1 to 5.

[16] There is a slightly difference in the approach presented in [122], as the validation method adopted in such work was based only on the Hit (n) strategy.

Table 5.30: Euclidean Distance based on Sentences - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.5227	0.5310	0.4331
Restaurants PT	0.6439	0.6272	0.6384	0.4931
2016 US Election EN	0.5227	-	0.4248	0.4108
GOP Debate EN	0.4948	0.3869	0.3255	0.3786
2012 US Election EN	0.5310	0.4248	-	0.4326
TV PT	0.5743	0.5732	0.5630	0.4097
Music Festival EN	0.8604	0.8073	0.7945	0.7249
Urban Problems PT	0.5592	0.5563	0.5486	0.4766
Airlines EN	0.6457	0.5563	0.5507	0.4612
Movies 1 EN	0.6290	0.5391	0.5453	0.4515
Movies PT	0.7567	0.7509	0.7198	0.6606
Movies 2 EN	0.7840	0.7491	0.7165	0.6811
Apple EN	0.6104	0.5221	0.5040	0.4380
Airlines ES	0.6002	0.5741	0.5757	0.4260
2018 CO Election ES	0.4331	0.4108	0.4326	-
Sports ES	0.5965	0.5704	0.5611	0.3985

Table 5.31: Cosine Similarity based on Sentences - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.5660	0.5543	0.6795
Restaurants PT	0.2636	0.1886	0.1660	0.3509
2016 US Election EN	0.5660	-	0.6749	0.6373
GOP Debate EN	0.5796	0.6980	0.7915	0.6223
2012 US Election EN	0.5543	0.6749	-	0.5991
TV PT	0.4194	0.3213	0.3515	0.5513
Music Festival EN	0.1454	0.1843	0.2152	0.2653
Urban Problems PT	0.5227	0.4676	0.4857	0.5417
Airlines EN	0.2046	0.3104	0.3317	0.3599
Movies 1 EN	0.2426	0.3485	0.3395	0.3781
Movies PT	0.2199	0.1488	0.2234	0.2259
Movies 2 EN	0.1707	0.1623	0.2393	0.1838
Apple EN	0.2707	0.3719	0.4249	0.3850
Airlines ES	0.3316	0.2769	0.2796	0.4682
2018 CO Election ES	0.6795	0.6373	0.5991	-
Sports ES	0.3754	0.3340	0.3614	0.5819

Table 5.32: Euclidean Distance based on Vocabulary - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.1624	0.1512	0.1218
Restaurants PT	0.2144	0.2437	0.2264	0.2214
2016 US Election EN	0.1624	-	0.0918	0.1519
GOP Debate EN	0.1607	0.1053	0.0989	0.1633
2012 US Election EN	0.1512	0.0918	-	0.1532
TV PT	0.1786	0.2175	0.2017	0.1715
Music Festival EN	0.2306	0.1645	0.1684	0.2202
Urban Problems PT	0.1246	0.1811	0.1654	0.1686
Airlines EN	0.2083	0.1514	0.1461	0.2060
Movies 1 EN	0.2293	0.1618	0.1514	0.2162
Movies PT	0.1430	0.1676	0.1297	0.1563
Movies 2 EN	0.1953	0.1505	0.1145	0.2019
Apple EN	0.2012	0.1623	0.1486	0.2054
Airlines ES	0.1876	0.2206	0.2007	0.1779
2018 CO Election ES	0.1218	0.1519	0.1532	-
Sports ES	0.1690	0.1912	0.1798	0.1530

Table 5.33: Cosine Similarity based on Vocabulary - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0.9673	0.9698	0.9807
Restaurants PT	0.9391	0.9253	0.9335	0.9365
2016 US Election EN	0.9673	-	0.9897	0.9712
GOP Debate EN	0.9673	0.9863	0.9877	0.9662
2012 US Election EN	0.9698	0.9897	-	0.9695
TV PT	0.9593	0.9413	0.9483	0.9627
Music Festival EN	0.9370	0.9683	0.9671	0.9424
Urban Problems PT	0.9789	0.95877	0.9640	0.9627
Airlines EN	0.9447	0.9717	0.9732	0.9465
Movies 1 EN	0.9356	0.9685	0.9726	0.9431
Movies PT	0.9715	0.9653	0.9780	0.9678
Movies 2 EN	0.9491	0.9717	0.9829	0.9470
Apple EN	0.9465	0.9671	0.97142	0.9454
Airlines ES	0.9510	0.9382	0.9463	0.9581
2018 CO Election ES	0.9807	0.9712	0.9695	-
Sports ES	0.9644	0.9550	0.9594	0.9707

Table 5.34: Unified Similarity Ranking, $n = 1$ - Experiment 2

	2018 BR	2016 US	2012 US	2018 CO
	Election	Election	Election	Election
	PT	EN	EN	ES
2018 BR Election PT	-	0	0	3
Restaurants PT	0	0	0	0
2016 US Election EN	0	-	2	0
GOP Debate EN	0	2	2	1
2012 US Election EN	0	2	-	0
TV PT	0	0	0	0
Music Festival EN	0	0	0	0
Urban Problems PT	0	0	0	0
Airlines EN	0	0	0	0
Movies 1 EN	0	0	0	0
Movies PT	0	0	0	0
Movies 2 EN	0	0	0	0
Apple EN	0	0	0	0
Airlines ES	0	0	0	0
2018 CO Election ES	4	0	0	-
Sports ES	0	0	0	0

From Table 5.34, we can observe that when $n = 1$, our dataset selection method selects: the 2018 CO Election ES for the 2018 BR Election PT; the 2012 US Election EN and the GOP Debate EN for the 2016 US Election dataset; the 2016 US Election EN and the GOP Debate EN datasets for the 2012 US Election EN; and the 2018 BR Election PT for the 2018 CO Election ES.

Table 5.35: Unified Similarity Ranking, $n = 2$ - Experiment 2

	2018 BR	2016 US	2012 US	2018 CO
	Election	Election	Election	Election
	PT	EN	EN	ES
2018 BR Election PT	-	0	0	3
Restaurants PT	0	0	0	0
2016 US Election EN	0	-	4	3
GOP Debate EN	2	4	4	1
2012 US Election EN	0	3	-	0
TV PT	0	0	0	0
Music Festival EN	0	0	0	0
Urban Problems PT	2	0	0	0
Airlines EN	0	0	0	0
Movies 1 EN	0	0	0	0
Movies PT	0	0	0	0
Movies 2 EN	0	0	0	0
Apple EN	0	0	0	0
Airlines ES	0	0	0	0
2018 CO Election ES	4	1	0	-
Sports ES	0	0	0	1

From Table 5.35, we can observe that when $n = 2$, our dataset selection method selects: the 2018 CO Election ES for the 2018 BR Election PT; the GOP Debate EN for the 2016 US Election dataset; the 2016 US Election EN and the GOP Debate EN datasets for the 2012 US Election EN; and the 2018 BR Election PT and the 2016 US Election EN datasets for the 2018 CO Election ES.

Table 5.36: Unified Similarity Ranking, $n = 3$ - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	0	0	3
Restaurants PT	0	0	0	0
2016 US Election EN	2	-	4	3
GOP Debate EN	2	4	4	2
2012 US Election EN	0	4	-	0
TV PT	0	0	0	1
Music Festival EN	0	0	0	0
Urban Problems PT	2	0	0	0
Airlines EN	0	0	0	0
Movies 1 EN	0	0	0	0
Movies PT	2	0	0	0
Movies 2 EN	0	2	2	0
Apple EN	0	0	0	0
Airlines ES	0	0	0	0
2018 CO Election ES	4	2	2	-
Sports ES	0	0	0	3

From Table 5.36, we can observe that when $n = 3$, our dataset selection method selects: the 2018 CO Election ES for the 2018 BR Election PT; the GOP Debate EN for the 2016 US Election dataset; the 2016 US Election EN and the GOP Debate EN datasets for the 2012 US Election EN; and the 2018 BR Election PT, the 2016 US Election EN and the Sports ES datasets for the 2018 CO Election ES.

Table 5.37: Unified Similarity Ranking, $n = 4$ - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	1	1	3
Restaurants PT	0	0	0	0
2016 US Election EN	2	-	4	4
GOP Debate EN	2	4	4	2
2012 US Election EN	4	4	-	3
TV PT	0	0	0	1
Music Festival EN	0	0	0	0
Urban Problems PT	2	0	0	0
Airlines EN	0	2	0	0
Movies 1 EN	0	0	0	0
Movies PT	2	0	2	0
Movies 2 EN	0	2	2	0
Apple EN	0	1	1	0
Airlines ES	0	0	0	0
2018 CO Election ES	4	2	2	-
Sports ES	0	0	0	3

From Table 5.37, we can observe that when $n = 4$, our dataset selection method points out: the 2018 CO Election ES and the 2012 US Election EN for the 2018 BR Election PT; the GOP Debate EN and the 2012 US Election EN for the 2016 US Election dataset; the 2016 US Election EN and the GOP Debate EN datasets for the 2012 US Election EN; and the 2016 US Election EN dataset for the 2018 CO Election ES.

Table 5.38: Unified Similarity Ranking, $n = 5$ - Experiment 2

	2018 BR Election PT	2016 US Election EN	2012 US Election EN	2018 CO Election ES
2018 BR Election PT	-	2	2	3
Restaurants PT	0	0	0	0
2016 US Election EN	2	-	4	4
GOP Debate EN	4	4	4	2
2012 US Election EN	4	4	-	3
TV PT	0	0	0	1
Music Festival EN	0	0	0	0
Urban Problems PT	4	1	1	0
Airlines EN	0	2	2	0
Movies 1 EN	0	0	0	0
Movies PT	2	0	2	2
Movies 2 EN	0	2	2	0
Apple EN	0	1	1	0
Airlines ES	0	0	0	1
2018 CO Election ES	4	4	2	-
Sports ES	0	0	0	4

From Table 5.38, we can observe that when $n = 5$, our dataset selection method points out as promising datasets: the 2018 CO Election ES, the 2012 US Election EN, the GOP Debate EN and the Urban Problems PT for the 2018 BR Election PT; the GOP Debate EN, the 2012 US Election EN and the 2018 CO Election ES for the 2016 US Election dataset; the 2016 US Election EN and the GOP Debate EN datasets for the 2012 US Election EN; and the 2016 US Election EN and Sports ES datasets for the 2018 CO Election ES.

Table 5.39: F1-score Summary - Experiment 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1-SVM	-	0.61	0.62	0.70	0.56	0.47	0.54	0.39	0.56	0.48	0.70	0.61	0.68	0.72	0.58	0.61
1-LR	-	0.61	0.61	0.64	0.49	0.52	0.60	0.41	0.61	0.49	0.64	0.62	0.63	0.66	0.58	0.64
1-DT	-	0.52	0.65	0.70	0.55	0.61	0.53	0.43	0.62	0.45	0.64	0.63	0.54	0.67	0.66	0.52
1-MLP	-	0.59	0.50	0.64	0.51	0.60	0.57	0.43	0.58	0.45	0.59	0.58	0.62	0.68	0.62	0.62
1-XGBoost	-	0.69	0.66	0.68	0.59	0.58	0.62	0.38	0.66	0.52	0.59	0.73	0.58	0.70	0.66	0.65
3-SVM	0.63	0.53	-	0.69	0.64	0.52	0.58	0.37	0.58	0.51	0.61	0.65	0.63	0.57	0.55	0.61
3-LR	0.63	0.47	-	0.69	0.64	0.55	0.57	0.42	0.59	0.50	0.64	0.63	0.65	0.55	0.55	0.61
3-DT	0.49	0.49	-	0.64	0.50	0.60	0.49	0.42	0.59	0.44	0.58	0.59	0.46	0.61	0.50	0.58
3-MLP	0.62	0.43	-	0.62	0.64	0.54	0.53	0.44	0.61	0.46	0.57	0.49	0.62	0.57	0.62	0.62
3-XGBoost	0.60	0.60	-	0.69	0.56	0.59	0.57	0.37	0.59	0.45	0.58	0.67	0.60	0.58	0.62	0.58
5-SVM	0.57	0.53	0.61	0.53	-	0.53	0.48	0.33	0.65	0.38	0.56	0.58	0.57	0.56	0.47	0.59
5-LR	0.593	0.523	0.6177	0.55	-	0.57	0.52	0.40	0.63	0.38	0.58	0.59	0.60	0.50	0.48	0.60
5-DT	0.54	0.47	0.56	0.55	-	0.52	0.48	0.40	0.48	0.54	0.54	0.51	0.49	0.52	0.57	0.47
5-MLP	0.57	0.48	0.63	0.56	-	0.53	0.57	0.41	0.52	0.41	0.54	0.60	0.58	0.48	0.46	0.61
5-XGBoost	0.54	0.50	0.57	0.51	-	0.50	0.53	0.35	0.54	0.42	0.57	0.56	0.54	0.50	0.51	0.58
15-SVM	0.63	0.65	0.61	0.66	0.50	0.61	0.65	0.41	0.68	0.49	0.64	0.65	0.66	0.64	-	0.68
15-LR	0.64	0.64	0.61	0.67	0.52	0.62	0.66	0.52	0.71	0.50	0.64	0.66	0.66	0.65	-	0.66
15-DT	0.54	0.55	0.58	0.65	0.49	0.61	0.56	0.44	0.63	0.51	0.56	0.60	0.54	0.64	-	0.60
15-MLP	0.63	0.54	0.59	0.65	0.51	0.60	0.63	0.52	0.63	0.50	0.61	0.58	0.62	0.62	-	0.68
15-XGBoost	0.65	0.62	0.63	0.64	0.53	0.62	0.61	0.40	0.68	0.50	0.60	0.65	0.66	0.65	-	0.66

The validation ranking for the 2018 BR Election PT - Experiment 2 is as follows:

- Airlines ES average rank: 1.4
- GOP Debate EN average rank: 2.6
- Movies 2 EN e Movies PT average rank: 6.0
- 2018 CO Election ES average rank: 6.4
- Apple EN and Restaurants PT average rank: 7.0
- Sports ES average rank: 7.2
- 2016 US Election EN average rank: 7.6
- Airlines EN average rank: 7.8
- Music Festival EN average rank: 10.6
- TV PT average rank: 10.6
- 2012 US Election EN average rank: 11.0
- Movies 1 EN average rank: 13.8
- Urban Problems PT average rank: 15.0

The validation ranking for the 2016 US Election EN - Experiment 2 is as follows:

- GOP Debate EN average rank: 1.7
- Movies 2 EN average rank: 5.0
- 2012 US Election EN average rank: 5.8
- Apple EN average rank: 6.0
- Sports ES average rank: 6.4
- 2018 BR Election PT average rank: 6.5
- Movies PT average rank: 7.0
- Airlines EN average rank: 7.2
- 2018 CO Election ES average rank: 7.4
- Airlines ES average rank: 8.2
- TV PT average rank: 9.0
- Music Festival EN average rank: 10.0
- Restaurants PT average rank: 11.4
- Movies 1 EN average rank: 13.6
- Urban Problems PT average rank: 14.8

The validation ranking for the 2012 US Election EN - Experiment 2 is as follows:

- 2016 US Election EN average rank: 1.8
- Sports ES 4.4
- 2018 BR Election PT 5.0
- Movies 2 EN 5.2
- Airlines, Apple, Movies PT 6.0
- GOP Debate EN 7.6
- Music Festival EN 9.6
- TV PT 9.6
- 2018 CO Election ES 9.8
- Airlines ES 10.0
- Restaurants PT 11.6
- Movies 1 EN 12.8

- Urban Problems PT 14.6

The validation ranking for the 2018 CO Election ES - Experiment 2 is as follows:

- Airlines EN average rank: 2.4
- GOP Debate e Sports ES average rank: 3.0
- Apple EN average rank: 5.4
- Movies 2 EN average rank: 5.8
- Airlines ES average rank: 6.0
- Music Festival EN average rank: 7.0
- 2018 BR Election PT average rank: 7.8
- Movies PT and TV PT average rank: 9.2
- 2016 US Election EN and Restaurants PT average rank: 9.6
- 2012 US Election EN average rank: 13.6
- Movies 1 EN and Urban Problems PT average rank: 14.2

For all the cases, the Nemenyi critical distance is CD: 8.9356, using $\alpha = 0.1$. According to the Nemenyi test, differences between datasets are significant if the average rank difference between them is greater than the critical distance. We observed that all the datasets selected by our method are in the first positions in the validation ranking or are significantly equivalent when $n = 1$, $n = 2$, and $n = 3$. When n is changed to 4 or 5, our method start to point out datasets that are out of the best positions in the validation ranking. This occurred when the 2012 US Election EN was selected to the 2018 BR Election PT ($n = 4$ and $n = 5$), and when the Urban Problems PT was selected to the 2018 BR Election PT ($n = 5$).

We observed that the suggestions obtained with the multilingual embeddings in this experiment are mostly inline with the suggestions obtained with the monolingual embeddings, as the datasets pointed out by the monolingual embeddings appeared in the third top positions (considering datasets of the given language) in all cases. The portuguese monolingual embeddings provided a best dataset selection than the multilingual approach. For the spanish and english experiments, both monolingual and multilingual experiments pointed out the same datasets, which are the ones in that language that achieved the best F1-scores.

5.3 Comparison with other Approaches

Finally, we compared our method to other methods for choosing proper training datasets presented in Chapter 3. In this analysis, we considered the datasets adopted in the multilingual experiment 1. From the approaches mentioned in Chapter 3, we could not compare our method to the SG strategy as it requires sentiment graphs that are not available for each one of the source datasets. Also, we do not consider the SFA/SCL strategies in our comparison as they are not based on source dataset selection. Instead, they focus on creating a mapping between the source and the target domain, not considering the existence of multiple source datasets. We believe this strategy may be useful in cases where only a single source dataset is available for the target task. Our comparison results are presented in Table 5.46, and details are available in Tables 5.40–5.45, where the best values are highlighted in bold. All the values presented in Table 5.46 are the names of the datasets selected by each dataset selection method. The first row values of Tables 5.40–5.45 refer to the source dataset identifiers, and the first column refers to the identifier of the target datasets (1 – 2018 BR Election PT, 3 – 2016 US Election EN, 5 – 2012 US Election EN, 15 – 2018 CO Election ES).

Table 5.40: TVC Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	-	0.113	0.064	0.044	0.081	0.201	0.022	0.082	0.034	0.059	**0.372**	0.080	0.026	0.069	0.139	0.154
3	0.084	0.025	-	0.239	0.288	0.034	0.080	0.012	0.164	0.360	0.152	**0.454**	0.101	0.033	0.039	0.091
5	0.144	0.055	0.390	0.290	-	0.074	0.104	0.030	0.195	0.395	0.222	**0.496**	0.124	0.062	0.080	0.152
15	0.163	0.051	0.035	0.018	0.053	0.081	0.009	0.032	0.014	0.031	0.166	0.048	0.009	0.161	-	**0.358**

Table 5.41: WVV Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	-	2.714	2.489	0.233	1.224	**0.164**	0.914	1.731	2.083	0.571	0.392	0.841	0.467	0.521	0.996	1.255
3	6.241	4.832	-	0.371	1.910	**0.270**	2.353	0.785	3.137	0.424	0.750	0.408	1.336	0.801	3.825	5.333
5	4.539	3.444	2.598	1.079	-	**0.694**	4.038	0.815	4.681	0.976	1.126	0.911	2.0572	0.851	3.227	3.066
15	1.813	2.395	3.662	0.315	2.350	0.247	1.858	0.330	3.493	1.026	**0.243**	1.214	1.021	0.582	-	0.285

Table 5.42: RCA Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	-	0.027	0.008	**0.001**	0.017	0.007	0.003	**0.001**	0.009	0.004	0.012	0.002	0.036	0.015	0.012	0.004
3	0.008	0.003	-	0.013	0.005	0.007	0.005	**0.002**	0.008	0.007	0.007	0.003	0.007	0.009	0.006	0.004
5	0.002	0.020	0.010	**0.001**	-	0.004	0.005	0.002	0.0198	0.013	0.015	0.007	0.025	0.007	0.017	0.004
15	0.004	0.005	0.009	0.019	0.012	0.001	0.005	**0**	0.001	0.001	0.006	0.001	0.021	0.005	0.006	0.006

Table 5.43: RCA* Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	**-**	0.185	**0.034**	0.096	0.138	0.116	0.181	0.475	0.141	0.066	0.121	0.158	0.065	0.059	0.065	0.154
3	0.198	0.258	**-**	**0.061**	0.129	0.176	0.166	0.428	0.243	0.078	0.137	0.108	0.172	0.170	0.117	0.186
5	0.252	0.206	**0.032**	0.151	**-**	0.121	0.173	0.408	0.196	0.095	0.128	0.122	0.149	0.151	0.147	0.175
15	0.211	0.188	0.129	0.115	0.243	0.132	0.117	0.405	0.216	**0.101**	0.135	0.117	0.188	0.171	**-**	0.196

Table 5.44: LM Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	**-**	0.00070	0.00002	0.00001	0.00035	**0.00109**	0.00002	0.00076	0.00001	0.00001	0.00108	0.00002	0.00001	0.00023	0.00019	0.00016
3	0.00086	0.00025	**-**	**0.00204**	0.00183	0.00011	0.00104	0.00005	0.00134	0.00148	0.00047	0.00166	0.00124	0.00012	0.00000	0.00097
5	0.00048	0.00009	0.00074	**0.00111**	**-**	0.00005	0.00077	0.00003	0.00047	0.00060	0.00018	0.00071	0.00061	0.00037	0.00015	0.00042
15	0.00101	0.00016	0.00003	0.00001	0.00049	0.00018	0.00000	0.00017	0.00002	0.00000	0.00030	0.00002	0.00001	0.00175	**-**	**0.00181**

Table 5.45: CB Method Summary

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	**-**	0.66116	0.71902	0.73463	0.73880	0.75056	0.73164	0.72622	**0.80713**	0.67769	0.64981	0.64792	0.64484	0.66254	0.72490	0.80185
3	0.79839	0.71434	**-**	0.71098	0.75538	0.75552	0.69730	0.71595	**0.81584**	0.63767	0.64361	0.65507	0.67229	0.68820	0.71915	0.79750
5	**0.81243**	0.66832	0.73399	0.76639	**-**	0.75064	0.69730	0.75463	0.76034	0.62645	0.69036	0.67711	0.66387	0.67337	0.73752	0.80914
15	0.82057	0.69888	0.75076	0.76631	0.77628	0.75326	0.69730	0.71963	0.79109	0.65155	0.67860	0.68013	0.68382	0.70311	**-**	**0.82526**

Table 5.46: Selected Datasets by Approach

	TVC	WVV	RCA	RCA*	LM	CB	Our method
1	Movies PT	TV PT	GOP Debate EN,Urban Problems PT	2016 US Election EN	TV PT	Airlines EN	2016 US Election EN, 2018 CO Election ES
3	Movies 2 EN	TV PT	Urban Problems PT	GOP Debate EN	GOP Debate EN	Airlines EN	GOP Debate EN,2012 US Election EN
5	Movies 2 EN	TV PT	GOP Debate EN	2016 US Election EN	GOP Debate EN	2018 BR Election PT	2016 US Election EN, GOP Debate EN
15	Sports ES	Movies PT	Urban Problems PT	Movies 1 EN	Sports ES	Sports ES	2018 BR Election PT, Sports ES

The remainder columns refer to the approaches adopted (TVC, WVV, RCA, RCA*, LM, CB) for selecting training datasets.

For the 2018 BR Election PT, the datasets selected and positions in the validation ranking are as follows: the Movies PT (11th position) was selected by the TVC method, the TV PT (8th position) was selected by the WVV and LM methods, the GOP Debate EN (10th position) was selected by the RCA method, the Urban Problems PT (14th position) was selected by the RCA method, the Airlines EN dataset (4th position) was selected by the CB method, the 2016 US Election EN (4th position) was selected by the RCA* and our method, and the 2018 CO Election ES (3rd position) was selected by our method. Therefore, our method was the one that achieved the best dataset selections for the 2018 BR Election PT. The Friedman test detected a statistical difference between the datasets, and the Nemenyi critical distance is CD: 8.9356. We can notice that the datasets that are in positions higher than the 9th position in the 2018 BR Election PT validation ranking of Experiment 1 are not equivalent to the first position in the ranking. Finally, we concluded that the selections of methods WVV, LM, CB, RCA* were also acceptable.

For the 2016 US Election EN, the selections and positions in the validation ranking are as follows: the Movies 2 EN (10th position) was selected by the TVC method, the TV PT (12th position) was selected by the WVV method, the Urban Problems PT (14th position) was selected by the RCA method, the GOP Debate EN (3rd position) was selected by the RCA*, LM and our method, the Airlines EN dataset (7th position) was selected by the CB method, the 2012 US Election EN (1st position) was selected by our method. Therefore, our method achieved the best selections for the 2016 US Election EN. The Friedman test detected that there is a statically difference between the datasets and the Nemenyi critical distance is CD: 8.9356. We can observe that the datasets that are in positions higher than the 9th position in the 2016 US Election EN validation ranking of Experiment 1 are not equivalent to the first position in the ranking. Finally, we concluded that the selections of methods RCA*, LM and CB were also acceptable.

For the 2012 US Election EN, the selections and positions in the validation ranking are as follows: the Movies 2 EN (7th position) was selected by the TVC method, the TV PT (6th position) was selected by the WVV method, the GOP Debate EN (14th position) was selected by the RCA, LM and our method, the 2016 US Election EN (1st position) was selected by the RCA*, and our method, the 2018 BR Election PT (4th position) was selected by the CB method. The Friedman test detected a statistical difference between the datasets and the Nemenyi critical distance of CD: 8.9356. We can observe that the datasets that are in positions higher than the 8th position in the 2012 US Election EN validation ranking of Experiment 1 are not equivalent to the first position in the ranking. Therefore, the RCA* method was the one that achieved the best selections for the 2012 US Election EN. Finally, we concluded that the selections provided by methods TVC, WVV, CB, and one of our selections were also acceptable.

For the 2018 CO Election ES dataset, the selections and positions in the validation ranking are as follows: Sports ES (3rd position) was selected by the TVC, LM, CB, and our method, Movies PT (12th position) was selected by the WVV method, Urban Problems PT (13th position) was selected by the RCA method, the Movies 1 EN (15th position) was selected by the RCA* method, the 2018 BR Election PT (2nd position) was selected by our method. The Friedman test detected a statistical difference between the datasets and the Nemenyi critical distance of CD: 8.9356. We can notice that the datasets that are in positions higher than the 10th position in the 2018 CO Election ES validation ranking of Experiment 1 are not equivalent to the first position in the ranking. Therefore, our method was the one that achieved the best selections for the 2018 CO Election ES. Finally, we concluded that the selections of methods TVC, LM, and CB were also acceptable.

As we can notice, the strategy presented in our proposed method achieved better dataset selections than the others in general. This is probably because our domain of interest involved datasets from different languages, and many of these strategies were not designed to deal with multilingual dataset comparison (e.g.: WVV and TVC). Another factor that may explain this is that some approaches depend on very large source datasets (e.g., LM) to better identify semantic relationships.

Chapter 6

Conclusions

This chapter presents our conclusions, limitations and threats to validity, lines for future work and points out the publications, presentations and awards related to this thesis.

In this thesis, we raised the following research questions:

RQ1: How machine learning algorithms and natural language processing may aid electoral analysis using social media?

RQ2: What are the existing computational approaches to analyze elections using social media?

In order to answer our first two RQs, we conducted a Systematic Literature Review, from which we constructed a survey, where we indicated many future research lines. We could observe that one interesting research line was related to out RQ3:

RQ3: How to take advantage of existing labeled datasets (of other domains) to better analyze electoral opinions using computational techniques?

To address research questions **Q1** and **Q2**, we conducted a systematic literature review (SLR) to survey approaches that forecast elections using social media data and computational techniques. As a result of the SLR we identified the main steps taken to perform electoral analysis and concluded that social media data analysis about elections can be used as thermometers but there are still many open issues in this field of study, as detailed in Section 3.1. We identified four categories of approaches to forecast elections using social media, namely: *Counting Based Approach*, *Political Alignment Approach*, *Event Detection Approach* and *Popularity Based Approach*. After analyzing those approaches, we concluded that *sentiment analysis* is a key task for electoral analysis and it is performed by several surveyed papers that achieved success considering these four categories. Therefore, we chose to improve sentiment analysis predictions aiming at obtaining better election analysis. Another points that were observed are the (i) lack of annotated data in the electoral domain, (ii) the complexity of electoral data, and (iii) the time restriction that makes difficult the task of manually labeling data during the short period of campaigns. In this context, we investigated how to take advantage of existing labeled datasets, as stated in research question **Q3**. Our hypothesis is that dataset similarity can help one to achieve better predictions:

Hypothesis (H): *If there is a high degree of similarity between a source labeled sentiment analysis dataset and a target electoral dataset, then machine learning classifiers trained with this source dataset will achieve proper sentiment predictions for the target electoral dataset.*

We proposed a dataset selection method for selecting datasets in scenarios where conditions (i), (ii) and (iii) take place and presented a case study that focused on the sentiment analysis task in the electoral domain. The central idea of our method is helping one to choose, from a set of sentiment analysis labeled datasets, possible candidates to induce classifiers to unlabeled data. The dataset selection method relies on analyzing dataset semantic similarity between labeled and unlabeled examples. The usage of multilingual embeddings as the vectorization technique allows us to compare and use datasets from different languages (English, Portuguese, Spanish) as classifiers training data. This factor is desirable in the electoral scenario since elections are recurring events around the world. We tested the usage of two similarity measures, namely, cosine similarity and euclidean distance combined with building vectors from averaging examples and averaging words vectors. In addition, we also tested the proposed approaches using monolingual embeddings in scenarios where only datasets that belong to the same language are compared.

The main advantage of our method regarding other approaches in the literature to select proper training datasets is that only similarity metrics between the datasets are calculated, being much less expensive than building models or classifiers for all the possible training datasets. Also, we are able to consider source datasets from different languages as we explore multilingual embeddings. In our opinion, this specific point would be interesting for low resources languages as we could find similar labeled datasets in other languages that could be used as starting point for training classifiers. It is important to mention that multilingual transference would only be possible if there is a crosslingual model trained in both source and target languages.

To evaluate the quality of the dataset selection in the electoral case study, we analyze both dataset similarity measures and F1-score ranking results. For both monolingual and multilingual experiments, our findings indicate that the analysis of dataset semantic similarity can be beneficial when one needs to choose a dataset to be used as starting point for training classifiers, as was pointed out in the thesis hypothesis. The presented investigation also shows that results related to high values of semantic similarity between datasets could in some cases surpass results obtained with datasets in the same language, leading us to believe that similar domains can contribute to better results regardless of language. This specific point can be observed by looking at the prediction results obtained using electoral datasets as training data. Some of the source datasets that appeared out of the top five first positions in our ranking were also able to achieve good predictions results. This factor may have occurred due to the existence of other common characteristics between the source-target dataset pairs that were not captured by using only the similarity metrics considered in this research. Also, factors such as unbiased data can improve the generalization ability of a model, what could explain models performing well despite low dataset similarity [75].

It is important to emphasize that sometimes heuristic inference methods may not be able to present the optimal datasets but being able to point out datasets that can achieve values that are close to the optimal ones. However, adopting dataset selection heuristics to select similar datasets beforehand has the advantage of not having to run all possible experiments to get a satisfactory result, saving time and reducing computational costs. Furthermore, if there is no label for the target domain, it is not possible to be sure which is the most proper source dataset and strategies like the one proposed in this thesis can be adopted. Our experimental results suggest that dataset similarity may be considered, even when datasets belong to different languages, to minimize negative effects that may occur due to domain shift in sentiment classification tasks.

Finally, we also provided as a contribution an election dataset[1] containing tweets related to the 2018 Brazilian Presidential Election that was manually labeled with the help of several volunteers using an online form. During the data annotation process we ask volunteers to assign labels to the electoral tweets in three dimensions, namely: sentiment analysis (SA), candidate support (CS), and offensive speech (OS). Each tweet of the provided dataset was labeled by at least three volunteers in these three dimensions. Although in this thesis we focused our investigation in the sentiment analysis dimension, we believe that the other two dimensions can be useful to allow other interesting future

[1] https://docs.google.com/spreadsheets/d/1JpVQ6EdFN7fvPYRtTvfOmbWhEQCqiDg7eupy_sGRBDk/edit?usp=sharing

work.

Another contribution of this research is an analysis of the labeling divergence process in two dimensions: inter-annotation agreement and the divergence of annotation per tweet. The inter-annotation agreement measures the level of divergence between annotators in each one of the tasks (SA, CS and OS) using the Krippendorff alpha method. The divergence of annotation between tweets measures how much the annotators disagree in each tweet by computing the entropy. This analysis highlights the great difficulty in labeling electoral data extracted from social media in regard to SA and OS dimensions and points out some common characteristics of electoral tweets that make the annotation process a complex task (non textual content, irony or humor, external knowledge, negative or neutral content and support hashtag, and mixed sentiment).

Therefore, we believe that the dataset selection method presented in this thesis can be useful for other domains that also have complex data as the ones with very specific terms, dynamic vocabulary, and in cases where there is not enough time to reliably label data from the target domain.

6.1 Limitations and Threats to Validity

As a limitation of this research we can cite the fact that our proposed method depends on the use of pretrained multilingual models. Therefore, if there are no multilingual models available for the languages of the datasets being compared, the multilingual approach cannot be explored. A threat to validity that we observed is the fact that our method does not analyze the labels of the candidate datasets, the similarity analysis only takes into account the examples (dataset sentences). Therefore, even if the selected dataset has a content very close to the target dataset, it is not possible to guarantee that the annotation process of the selected dataset occurred properly, avoiding bias and guaranteeing the quality of the annotation. Still thinking about this point, cases involving polarity divergence, that is, similar instances that have different labels are also not treated by our method. However, as one of the premises of this thesis is the fact that the target dataset does not have available labels, it would not be possible to carry out an analysis involving the labels of the source and target datasets to deal with this issue. Finally, not using a filter to discard spam data can be considered a limitation of this approach, since our systematic literature review revealed that electoral data from Twitter is related to a high level of spam. Analyzes in this sense were outside the scope of this thesis.

6.2 Future Work

As next steps, we intend to investigate the offensive speech detection problem in the electoral domain. One direction would be to explore the dataset selection method for selecting similar offensive speech labeled datasets to be used as starting point for electoral analysis. One can investigate, for example, if a high level of offensive content related to a candidate is related to a high level of candidate rejection. An offensive speech classification model would also allow analyzing whether people who support candidates whose posts are full of offensive content also tend to propagate offensive posts. Also, an analysis of the combination of subsets of the selected datasets could be considered as an issue for further investigation to improve the classification tasks. Other direction for future work is the investigation of semi-supervised or active learning approaches that take advantage of crowdsourcing labels to analyze the overall sentiment of the thousands of collected tweets related to the 2018 Brazilian Presidential Elections. Another possible research direction is to carry out a cross-cultural analysis of the electoral scenario to understand how people express themselves about elections in different cultures and the impacts of cultural factors.

6.3 Publications, Presentations and Awards

The list of papers related to this thesis already published is as follows:

- Santos, J. S.; Paes, A.; Bernardini, F. Similarity-based Dataset Recommendation across Languages and Domains to Sentiment Analysis in the Electoral Domain. International IFIP Electronic Government Conference (EGOV), Linkoping, Sweden, 2022.

- Santos, J. S.; Bernardini, F.; Paes, A. A survey on the use of data and opinion mining in social media to political electoral outcomes prediction. Social Network Analysis and Mining (SNAM), 11, 1 (2021), 1 – 39.

- Santos, J. S.; Bernardini, F.; Paes, A. Measuring the degree of divergence when labeling tweets in the electoral scenario. In Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2021), 2021.

- Santos, J. S.; Paes, A.; Bernardini, F. Investigating Transfer Learning Approaches for Mining Opinions in the Electoral Domain. In: Journal LXAI Workshop co-located with the Neural Information Processing Systems (NeurIPS) conference, Vancouver, Canada, 2019.

- Santos, J. S.; Paes, A.; Bernardini, F. Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In 2019 8th Brazilian Conference on Intelligent Systems (BRACIS) (2019), IEEE, pp. 455 – 460, 2019.

A subset of results of this research was presented by the author at:

- the Brazilian Conference on Intelligence Systems (BRACIS), Salvador, Brazil in October 2019;
- the Latin America Meeting in Artificial Intelligence (Khipu), Montevideo, Uruguay in November 2019;
- the LXAI Workshop co-located with the Neural Information Processing Systems (NeurIPS) conference, Vancouver, Canada in December 2019;
- the Microsoft Research PhD Summit, virtual, in December 2020;
- the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), virtual, July 2021.

The research presented in this thesis is related to the Microsoft Research Latin America PhD Award[2], received by the thesis author in 2020.

[2] http://aka.ms/AA8s9pc

References

[1] Abdullah, N. A.; Feizollah, A.; Sulaiman, A.; Anuar, N. B. Challenges and recommended solutions in multi-source and multi-domain sentiment analysis. *IEEE Access 7* (2019), 144957–144971.

[2] Ajito, M.; Kawahata, Y.; Ishii, A. Analysis of national election using mathematical model of hit phenomenon. In *Big Data (Big Data), 2017 IEEE International Conference on* (2017), IEEE, pp. 4722–4724.

[3] Al-Moslmi, T.; Omar, N.; Abdullah, S.; Albared, M. Approaches to cross-domain sentiment analysis: A systematic literature review. *Ieee access 5* (2017), 16173–16192.

[4] Alakrot, A. *Detection of anti-social behaviour in online communication in Arabic.* Tese de Doutorado, 2019.

[5] Allen, C.; Hospedales, T. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning* (2019), PMLR, pp. 223–231.

[6] Almeida, J. M.; Pappa, G. L., et al. Twitter population sample bias and its impact on predictive outcomes: A case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (2015), ACM, pp. 1254–1261.

[7] Asur, S.; Huberman, B. A. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (2010), IEEE Computer Society, pp. 492–499.

[8] Awais, M.; Hassan, S.-U.; Ahmed, A. Leveraging big data for politics: predicting general election of pakistan using a novel rigged model. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–9.

[9] Bachhuber, J.; Koppeel, C.; Morina, J.; Rejström, K.; Steinschulte, D. Us election prediction: A linguistic analysis of us twitter users. In *Designing Networks for Innovation and Improvisation.* Springer, 2016, pp. 55–63.

[10] Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (2017), International World Wide Web Conferences Steering Committee, pp. 759–760.

[11] Bansal, B.; Srivastava, S. On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science 135* (2018), 346–353.

[12] Bansal, B.; Srivastava, S. Lexicon-based twitter sentiment analysis for vote share prediction using emoji and n-gram features. *International Journal of Web Based Communities 15*, 1 (2019), 85–99.

[13] Bastos, M.; Mercea, D. Parametrizing brexit: mapping twitter political space to parliamentary constituencies. *Information, Communication & Society 21*, 7 (2018), 921–939.

[14] Ben-David, A.; Matamoros-Fernández, A. Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain. *International Journal of Communication 10* (2016), 1167–1193.

[15] Bifet, A.; Gavaldà, R.; Holmes, G.; Pfahringer, B. *Machine Learning for Data Streams with Practical Examples in MOA*. MIT Press, 2018. `https://moa.cms.waikato.ac.nz/book/`.

[16] Bilal, M.; Asif, S.; Yousuf, S.; Afzal, U. 2018 pakistan general election: Understanding the predictive power of social media. In *2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (2018), IEEE, pp. 1–6.

[17] Blitzer, J.; McDonald, R.; Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (2006), pp. 120–128.

[18] Bobicev, V.; Sokolova, M. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *RANLP* (2017), vol. 97.

[19] Bovet, A.; Morone, F.; Makse, H. A. Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Scientific reports 8*, 1 (2018), 8673.

[20] Breur, T. Us elections: How could predictions be so wrong?, 2016.

[21] Brito, K. d. S.; Adeodato, P. J. L. Predicting brazilian and us elections with machine learning and social media data. In *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), IEEE, pp. 1–8.

[22] Budiharto, W.; Meiliana, M. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data 5*, 1 (2018), 1–10.

[23] Burnap, P.; Gibson, R.; Sloan, L.; Southern, R.; Williams, M. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies 41* (2016), 230–233.

[24] Ca´ ete, J.; Chaperon, G.; Fuentes, R.; Ho, J.-H.; Kang, H.; Pérez, J. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020* (2020).

[25] Cabanas Martí, M. Study of algorithms of supervision of systems based on artificial intelligence. Master's thesis, Universitat Politècnica de Catalunya, 2019.

[26] Campanale, M.; Caldarola, E. G. Revealing political sentiment with twitter: the case study of the 2016 italian constitutional referendum. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018), IEEE, pp. 861–868.

[27] Castelvecchi, D. Why the polls got the UK election wrong? *Nature News* (2017).

[28] Castro, R.; Vaca, C. National leaders' twitter speech to infer political leaning and election results in 2015 venezuelan parliamentary elections. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on* (2017), IEEE, pp. 866–871.

[29] Chidambaram, M.; Yang, Y.; Cer, D.; Yuan, S.; Sung, Y.-H.; Strope, B.; Kurzweil, R. Learning cross-lingual sentence representations via a multi-task dual-encoder model. pp. 250–259.

[30] Chowdhury, G. G. Natural language processing. *Annual review of information science and technology 37*, 1 (2003), 51–89.

[31] Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement 20*, 1 (1960), 37–46.

[32] Crosset, V.; Tanner, S.; Campana, A. Researching far right groups on twitter: Methodological challenges 2.0. *new media & society 21*, 4 (2019), 939–961.

[33] Dai, X.; Karimi, S.; Hachey, B.; Paris, C. Using similarity measures to select pretraining data for ner. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), pp. 1460–1470.

[34] Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research 7* (2006), 1–30.

[35] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[36] Di Giovanni, M.; Brambilla, M.; Ceri, S.; Daniel, F.; Ramponi, G. Content-based classification of political inclinations of twitter users. In *2018 IEEE International Conference on Big Data (Big Data)* (2018), IEEE, pp. 4321–4327.

[37] Dokoohaki, N.; Zikou, F.; Gillblad, D.; Matskin, M. Predicting swedish elections with twitter: A case for stochastic link structure analysis. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on* (2015), IEEE, pp. 1269–1276.

[38] Duarte, L.; Macedo, L.; Oliveira, H. G. Exploring emojis for emotion recognition in portuguese text. In *EPIA Conference on Artificial Intelligence* (2019), Springer, pp. 719–730.

[39] Dwi Prasetyo, N.; Hauff, C. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (2015), ACM, pp. 149–158.

[40] Elsahar, H.; Gallé, M. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 2163–2173.

[41] Ezeibe, C. C. Hate speech and electoral violence in nigeria. In *Two-day National Conference on the* (2015).

[42] Fan, W.; Davidson, I. Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), pp. 147–156.

[43] Fano, S.; Slanzi, D. Using twitter data to monitor political campaigns and predict election results. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (2017), Springer, pp. 191–197.

[44] Farías, D. I. H.; Patti, V.; Rosso, P. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT) 16*, 3 (2016), 1–24.

[45] Fleiss, J. L.; Levin, B.; Paik, M. C., et al. The measurement of interrater agreement. *Statistical methods for rates and proportions 2*, 212-236 (1981), 22–23.

[46] Forsythe, R.; Myerson, R. B.; Rietz, T. A.; Weber, R. J. An experiment on coordination in multi-candidate elections: The importance of polls and election histories. *Social Choice and Welfare 10*, 3 (1993), 223–247.

[47] Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR) 51*, 4 (2018), 1–30.

[48] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association 32*, 200 (1937), 675–701.

[49] Gao, L.; Kuppersmith, A.; Huang, R. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Taipei, Taiwan, Nov. 2017), Asian Federation of Natural Language Processing, pp. 774–782.

[50] Garcia, A. C. B.; Silva, W.; Correia, L. The prednews forecasting model. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (2018), pp. 1–6.

[51] Gayo-Avello, D. No, you cannot predict elections with twitter. *IEEE Internet Computing 16*, 6 (2012), 91–94.

[52] Gollapudi, S. *Practical machine learning.* Packt Publishing Ltd, 2016.

[53] Guerra, P. C.; Veloso, A.; Meira, W. J.; Almeida, V. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (2011), ACM, pp. 150–158.

[54] Harris, Z. S. Distributional structure. *Word 10*, 2-3 (1954), 146–162.

[55] Heredia, B.; Prusa, J.; Khoshgoftaar, T. Exploring the effectiveness of twitter at polling the united states 2016 presidential election. In *Collaboration and Internet Computing (CIC), 2017 IEEE 3rd International Conference on* (2017), IEEE, pp. 283–290.

[56] Heredia, B.; Prusa, J. D.; Khoshgoftaar, T. M. Social media for polling and predicting united states election outcome. *Social Network Analysis and Mining 8*, 1 (2018), 48.

[57] Hinch, J. # makeamericaspollsgreatagain: Evaluating twitter as a tool to predict election outcomes.

[58] Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[59] Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (01 2018), pp. 328–339.

[60] Huang, J.-Y. Web mining for the mayoral election prediction in taiwan. *Aslib Journal of Information Management 69*, 6 (2017), 688–701.

[61] Hwang, B. Reddit sentiment analysis to improve election predictions. In *In International Conference Big Data Analytics, Data Mining and Computational Intelligence* (07 2019), pp. 204–208.

[62] Ibrahim, M.; Abdillah, O.; Wicaksono, A. F.; Adriani, M. Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on* (2015), IEEE, pp. 1348–1353.

[63] Idan, L.; Feigenbaum, J. Show me your friends, and i will tell you whom you vote for: Predicting voting behavior in social networks. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2019), IEEE, pp. 816–824.

[64] James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

[65] Jorge, V. L.; Faria, A. M. T. d.; Silva, M. G. d. Posicionamento dos partidos políticos brasileiros na escala esquerda-direita: dilemas metodológicos e revisão da literatura. *Revista Brasileira de Ciência Política*, 33 (2020), e227686.

[66] Jose, R.; Chooralil, V. S. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach. In *2016 international conference on data mining and advanced computing (SAPIENCE)* (2016), IEEE, pp. 64–67.

[67] Joseph, F. J. J. Twitter based outcome predictions of 2019 indian general elections using decision tree. In *2019 4th International Conference on Information Technology (InCIT)* (2019), IEEE, pp. 50–53.

[68] Joshi, M.; Prajapati, P.; Shaikh, A.; Vala, V. A survey on sentiment analysis. *International Journal of Computer Applications 163*, 6 (2017), 34–38.

[69] Kagan, V.; Stevens, A.; Subrahmanian, V. Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election. *IEEE Intelligent Systems*, 1 (2015), 2–5.

[70] Kalampokis, E.; Karamanou, A.; Tambouris, E.; Tarabanis, K. A. On predicting election results using twitter and linked open data: The case of the uk 2010 election. *J. UCS 23*, 3 (2017), 280–303.

[71] Kassraie, P.; Modirshanechi, A.; Aghajan, H. K. Election vote share prediction using a sentiment-based fusion of twitter data with google trends and online polls. In *DATA* (2017), pp. 363–370.

[72] Kharde, V.; Sonawane, P., et al. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971* (2016).

[73] Khatua, A.; Khatua, A.; Ghosh, K.; Chaki, N. Can# twitter_trends predict election results? evidence from 2014 indian general election. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (2015), IEEE, pp. 1676–1685.

[74] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[75] Kouw, W. M.; Loog, M. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806* (2018).

[76] Krippendorff, K. Computing krippendorff's alpha-reliability.

[77] Kristiyanti, D. A.; Umam, A. H., et al. Prediction of indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis. In *2019 5th International Conference on New Media Studies (CONMEDIA)* (2019), IEEE, pp. 36–42.

[78] Kumar, A.; Sebastian, T. M., et al. Sentiment analysis: A perspective on its past, present and future. *International Journal of Intelligent Systems and Applications 4*, 10 (2012), 1–14.

[79] Lee, J.; Ryu, H.; Mon, L.; Park, J. S. Citizens' use of twitter in political information sharing in south korea.

[80] Li, B.; Guo, D.; Chang, M.; Li, M.; Bian, A. The prediction on the election of representatives. In *Security, Pattern Analysis, and Cybernetics (SPAC), 2017 International Conference on* (2017), IEEE, pp. 329–334.

[81] Li, N.; Zhai, S.; Zhang, Z.; Liu, B. Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2017), vol. 31.

[82] Li, Y.; Guo, H.; Zhang, Q.; Gu, M.; Yang, J. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowledge-Based Systems 160* (2018), 1–15.

[83] Liu, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*, 1 (2012), 1–167.

[84] Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2 ed. Studies in Natural Language Processing. Cambridge University Press, 2020.

[85] Lopardo, A.; Brambilla, M. Analyzing and predicting the us midterm elections on twitter with recurrent neural networks. In *2018 IEEE International Conference on Big Data (Big Data)* (2018), IEEE, pp. 5389–5391.

[86] Mahendiran, A.; Wang, W.; Lira, J. A. S.; Huang, B.; Getoor, L.; Mares, D.; Ramakrishnan, N. Discovering evolving political vocabulary in social media. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)* (2014), IEEE, pp. 1–7.

[87] Maldonado, M.; Sierra, V. Can social media predict voter intention in elections? _x000d_ the case of the 2012 dominican republic presidential election.

[88] Martínez-Cámara, E.; Martín-Valdivia, M. T.; Urena-López, L. A.; Montejo-Ráez, A. R. Sentiment analysis in twitter. *Natural Language Engineering 20*, 1 (2014), 1–28.

[89] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[90] Naiknaware, B. R.; Kawathekar, S. S. Prediction of 2019 indian election using sentiment analysis. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on* (2018), IEEE, pp. 660–665.

[91] Ni, J.; Hernandez Abrego, G.; Constant, N.; Ma, J.; Hall, K.; Cer, D.; Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022* (Dublin, Ireland, may 2022), Association for Computational Linguistics, pp. 1864–1874.

[92] Öhman, E.; Kajava, K.; Tiedemann, J.; Honkela, T. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2018), pp. 24–30.

[93] Okeowo, A. Hate on the rise after trump's election. *The New Yorker 17* (2016).

[94] Pan, S. J.; Ni, X.; Sun, J.-T.; Yang, Q.; Chen, Z. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web* (2010), pp. 751–760.

[95] Pan, S. J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*, 10 (2009), 1345–1359.

[96] Pan, S. J.; Yang, Q., et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*, 10 (2010), 1345–1359.

[97] Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2*, 1–2 (2008), 1–135.

[98] Pennebaker, J. W.; Francis, M. E.; Booth, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001 (2001), 2001.

[99] Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[100] Perkins, D. N.; Salomon, G., et al. Transfer of learning. *International encyclopedia of education 2* (1992), 6452–6457.

[101] Persily, N. The 2016 us election: Can democracy survive the internet? *Journal of democracy 28*, 2 (2017), 63–76.

[102] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of North American Association for Computational Linguistics (NAACL)* (2018).

[103] Phillips, L.; Dowling, C.; Shaffer, K.; Hodas, N.; Volkova, S. Using social media to predict the future: a systematic literature review. *arXiv preprint arXiv:1706.06134* (2017).

[104] Praciano, B. J. G.; da Costa, J. P. C. L.; Maranhão, J. P. A.; de Mendonça, F. L. L.; de Sousa Júnior, R. T.; Prettz, J. B. Spatio-temporal trend analysis of the brazilian elections based on twitter data. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (2018), IEEE, pp. 1355–1360.

[105] Przeworski, A.; Alvarez, R. M.; Alvarez, M. E.; Cheibub, J. A.; Limongi, F., et al. *Democracy and development: political institutions and well-being in the world, 1950-1990*, vol. 3. Cambridge University Press, 2000.

[106] Raina, R.; Ng, A. Y.; Koller, D. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 713–720.

[107] Ramos, J., et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (2003), vol. 242, Piscataway, NJ, pp. 133–142.

[108] Ramzan, M.; Mehta, S.; Annapoorna, E. Are tweets the real estimators of election results? In *2017 Tenth International Conference on Contemporary Computing (IC3)* (2017), IEEE, pp. 1–4.

[109] Riemer, P. O. *Measuring Polarization in an Online News Forum.* Tese de Doutorado, Wien, 2021.

[110] Rizoiu, M.-A.; Wang, T.; Ferraro, G.; Suominen, H. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829* (2019).

[111] Rosseti, I.; Viterbo, J., et al. On tweets, retweets, hashtags and user profiles in the 2016 american presidential election scene. In *Proceedings of the 18th Annual International Conference on Digital Government Research* (2017), ACM, pp. 120–128.

[112] Rothschild, D.; Malhotra, N. Are public opinion polls self-fulfilling prophecies? *Research & Politics 1*, 2 (2014), 2053168014547667.

[113] Ruder, S. *Neural transfer learning for natural language processing.* Tese de Doutorado, NUI Galway, 2019.

[114] Ruder, S. Neural transfer learning for natural language processing.

[115] Sagiroglu, S.; Sinanc, D. Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (2013), IEEE, pp. 42–47.

[116] Saif, H.; He, Y.; Fernandez, M.; Alani, H. Semantic patterns for sentiment analysis of twitter. In *International Semantic Web Conference* (2014), Springer, pp. 324–340.

[117] Sanders, E.; de Gier, M.; van den Bosch, A. Using demographics in predicting election results with twitter. In *International Conference on Social Informatics* (2016), Springer, pp. 259–268.

[118] Sanders, E.; van den Bosch, A. Optimising twitter-based political election prediction with relevance andsentiment filters. In *Proceedings of The 12th Language Resources and Evaluation Conference* (2020), pp. 6158–6165.

[119] Santos, J. S.; Bernardini, F.; Paes, A. Measuring the degree of divergence when labeling tweets in the electoral scenario. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2021)* (2021).

[120] Santos, J. S.; Bernardini, F.; Paes, A. A survey on the use of data and opinion mining in social media to political electoral outcomes prediction. *Social Network Analysis and Mining 11*, 1 (2021), 1–39.

[121] Santos, J. S.; Paes, A.; Bernardini, F. Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)* (2019), IEEE, pp. 455–460.

[122] Santos, J. S.; Paes, A.; Bernardini, F. Similarity-based dataset recommendation across languages and domains to sentiment analysis in the electoral domain. In *International IFIP Electronic Government Conference (EGOV)* (Linkoping, Sweden, 2022), EGOV.

[123] Sarkar, D.; Bali, R.; Ghosh, T. *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras.* Packt Publishing Ltd, 2018.

[124] Schultz, L. R.; Loog, M.; Esfahani, P. M. Distance based source domain selection for sentiment classification. *arXiv preprint arXiv:1808.09271* (2018).

[125] Shaban, T. A.; Hexter, L.; Choi, J. D. Event analysis on the 2016 us presidential election using social media. In *International Conference on Social Informatics* (2017), Springer, pp. 201–217.

[126] Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review 5*, 1 (2001), 3–55.

[127] Sharma, P.; Moh, T.-S. Prediction of indian election using sentiment analysis on hindi twitter. In *Big Data (Big Data), 2016 IEEE International Conference on* (2016), IEEE, pp. 1966–1971.

[128] Singh, P.; Dwivedi, Y. K.; Kahlon, K. S.; Pathania, A.; Sawhney, R. S. Can twitter analytics predict election outcome? an insight from 2017 punjab assembly elections. *Government Information Quarterly* (2020), 101444.

[129] Singh, P.; Sawhney, R. S.; Kahlon, K. S. Predicting the outcome of spanish general elections 2016 using twitter as a tool. In *Advanced Informatics for Computing Research.* Springer, 2017, pp. 73–83.

[130] Sokolova, K.; Perez, C. Elections and the twitter community: The case of right-wing and left-wing primaries for the 2017 french presidential election. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018), IEEE, pp. 1021–1026.

[131] Souza, F.; Nogueira, R.; Lotufo, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)* (2020).

[132] Srivastava, R.; Kumar, H.; Bhatia, M.; Jain, S. Analyzing delhi assembly election 2015 using textual content of social network. In *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015* (2015), ACM, pp. 78–85.

[133] Sturgis, P.; Kuha, J.; Baker, N.; Callegaro, M.; Fisher, S.; Green, J.; Jennings, W.; Lauderdale, B. E.; Smith, P. An assessment of the causes of the errors in the 2015 uk general election opinion polls. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*, 3 (2018), 757–781.

[134] Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research 2*, Nov (2001), 45–66.

[135] Torrey, L.; Shavlik, J., et al. Transfer learning. handbook of research on machine learning applications and trends: algorithms, methods, and techniques. *Information Science Reference* (2009), 22.

[136] Tsakalidis, A.; Papadopoulos, S.; Cristea, A. I.; Kompatsiaris, Y. Predicting elections for multiple countries using twitter and polls. *IEEE Intelligent Systems 30*, 2 (2015), 10–17.

[137] Tung, K.-C.; Wang, E. T.; Chen, A. L. Mining event sequences from social media for election prediction. In *Industrial Conference on Data Mining* (2016), Springer, pp. 266–281.

[138] Tunstall, L.; von Werra, L.; Wolf, T. *Natural language processing with transformers.* "O'Reilly Media, Inc.", 2022.

[139] Unankard, S.; Li, X.; Sharaf, M.; Zhong, J.; Li, X. Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering* (2014), Springer, pp. 1–16.

[140] Vepsäläinen, T.; Li, H.; Suomi, R. Facebook likes and public opinion: Predicting the 2015 finnish parliamentary elections. *Government Information Quarterly 34*, 3 (2017), 524–532.

[141] Wang, A. H. Don't follow me: Spam detection in twitter. In *2010 international conference on security and cryptography (SECRYPT)* (2010), IEEE, pp. 1–10.

[142] Wang, L.; Gan, J. Q. Prediction of the 2017 french election based on twitter data analysis. In *Computer Science and Electronic Engineering (CEEC), 2017* (2017), IEEE, pp. 89–93.

[143] Wang, L.; Gan, J. Q. Prediction of the 2017 french election based on twitter data analysis using term weighting. In *2018 10th Computer Science and Electronic Engineering (CEEC)* (2018), IEEE, pp. 231–235.

[144] Wang, M.-H.; Lei, C.-L. Boosting election prediction accuracy by crowd wisdom on social forums. In *Consumer Communications & Networking Conference (CCNC), 2016 13th IEEE Annual* (2016), IEEE, pp. 348–353.

[145] Wang, W.; Rothschild, D.; Goel, S.; Gelman, A. Forecasting elections with non-representative polls. *International Journal of Forecasting 31*, 3 (2015), 980–991.

[146] Warner, W.; Hirschberg, J. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (2012), Association for Computational Linguistics, pp. 19–26.

[147] White, K. Forecasting canadian elections using twitter. In *Canadian Conference on Artificial Intelligence* (2016), Springer, pp. 186–191.

[148] Wicaksono, A. J., et al. A proposed method for predicting us presidential election by analyzing sentiment in social media. In *Science in Information Technology (ICSITech), 2016 2nd International Conference on* (2016), IEEE, pp. 276–280.

[149] Woolley, S. C. Automating power: Social bot interference in global politics. *First Monday 21*, 4 (2016).

[150] Wu, F.; Huang, Y. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016), vol. 1, pp. 301–310.

[151] Wu, F.; Huang, Y.; Yuan, Z. Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources. *Information Fusion 35* (2017), 26–37.

[152] Xie, Z.; Liu, G.; Wu, J.; Wang, L.; Liu, C. Wisdom of fusion: Prediction of 2016 taiwan election with heterogeneous big data. In *Service Systems and Service Management (ICSSSM), 2016 13th International Conference on* (2016), IEEE, pp. 1–6.

[153] Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Hernandez Abrego, G.; Yuan, S.; Tar, C.; Sung, Y.-h.; Strope, B.; Kurzweil, R. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Online, July 2020), Association for Computational Linguistics, pp. 87–94.

[154] You, Q.; Cao, L.; Cong, Y.; Zhang, X.; Luo, J. A multifaceted approach to social multimedia-based prediction of elections. *IEEE Transactions on Multimedia 17*, 12 (2015), 2271–2280.

[155] Zeedan, R. The 2016 us presidential elections: What went wrong in pre-election polls? demographics help to explain. *Multidisciplinary Scientific Journal 2*, 1 (2019), 84–101.

[156] Zhang, Y.; Hu, X.; Li, P.; Li, L.; Wu, X. Cross-domain sentiment classification-feature divergence, polarity divergence or both? *Pattern recognition letters 65* (2015), 44–50.

[157] Zhang, Z.; Luo, L. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, Preprint (2018), 1–21.

[158] Zhong, E.; Fan, W.; Yang, Q.; Verscheure, O.; Ren, J. Cross validation framework to choose amongst models and datasets for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2010), Springer, pp. 547–562.