ARTHUR BUGARIN DE MELLO MARQUES
FERNANDA TIEMI KANAZAWA

SENTIMENT ANALYSIS APPLIED TO TRANSCRIPTS OF THE PORTUGUESE PARLIAMENTARY DEBATES

São Paulo
2021

ARTHUR BUGARIN DE MELLO MARQUES
FERNANDA TIEMI KANAZAWA

SENTIMENT ANALYSIS APPLIED TO TRANSCRIPTS OF THE PORTUGUESE PARLIAMENTARY DEBATES

Report presented to the Polytechnic School of the University of São Paulo as a partial requirement to obtain the title of Mechatronics Engineer.

Advisor:
Anna Helena Reali Costa

São Paulo
2021

ABSTRACT

Recent years have witnessed the development of several studies in the Sentiment Analysis field, a branch of Natural Language Processing. Most works had previously focused on the English language, but recent research in other languages, including Portuguese, has emerged more frequently. This work analyses the sentiment contained in the Portuguese Parliament's debates, whose transcripts are freely available on the internet, classifying the opinion of each politician as positive or negative towards the initiatives.

Keywords – Natural Language Processing, Sentiment Analysis, Portuguese Parliament, Debates.

LIST OF FIGURES

CONTENTS

1 INTRODUCTION

Portugal has been a democracy since 1974, when the dictatorship ended and the provisional government, Junta de Salvação Nacional (JSN), was installed. In 1975, JSN was replaced by Conselho da Revolução, a sovereign institution. In that year, the debates of the Portuguese Parliament started being transcribed, forming the parliamentary minutes, as illustrated in Figure 1, freely available to the public.

REPÚBLICA PORTUGUESA
SECRETARIADO DA ASSEMBLEIA CONSTITUINTE

DIÁRIO DA ASSEMBLEIA CONSTITUINTE

TERÇA-FEIRA, 3 DE JUNHO DE 1975 * PREÇO DESTE NÚMERO 2$00 * NÚMERO 1

SESSÃO INAUGURAL EM 2 DE JUNHO

Sob a presidência de S. Ex.ª o General Francisco da Costa Gomes, Presidente da República Portuguesa, que tinha à direita SS. Ex.as o Engenheiro Henrique Teixeira Queirós de Barros, Presidente Interino da Assembleia Constituinte, e o Vice-Almirante José Baptista Pinheiro de Azevedo, representante do Chefe do Estado-Maior-General das Forças Armadas, e à esquerda S. Ex.ª o General Vasco dos Santos Gonçalves, Primeiro-Ministro, efectuou-se na Sala das Sessões a sessão inaugural da Assembleia Constituinte.

Estavam presentes o Conselho da Revolução, o Governo, o Corpo Diplomático e altas autoridades civis e militares.

Às 16 horas e 10 minutos deu entrada na sala, onde se encontravam os Srs. Deputados, o cortejo presidencial, no qual se incorporaram, além do Presidente Interino da Assembleia Constituinte, do Primeiro-Ministro e do representante do Chefe do Estado-Maior-General das Forças Armadas, o Presidente da Comissão Nacional Instaladora da Assembleia Constituinte e o Chefe do Protocolo do Estado.

Às 16 horas e 12 minutos o Sr. Presidente Interino da Assembleia Constituinte declarou aberta a sessão, em nome de S. Ex.ª o Presidente da República, acrescentando:

Em nome de S. Ex.ª o Sr. Presidente da República, declaro aberta a sessão.

O Sr. Presidente da República vai ler a sua mensagem à Assembleia Constituinte.

S. Ex.ª o Chefe do Estado leu então a sua mensagem, que era do teor seguinte:

As dinâmicas revolucionárias avançam na resultante de miríades de actos espontâneos ou condicionados, individuais ou colectivos, mas que projectam os povos para fora do enquadramento das leis e normas sociais que os espartilhavam em situações políticas anteriores.

Revoluções são processos em que os povos, depois de tomarem um metabolismo social tornado intolerável, entram em marcha, mais ou menos acelerada, até atingirem um novo patamar de equilíbrio social, historicamente transitório.

Nos patamares do equilíbrio social dos povos compete aos legisladores prepararem legislação avançada, serem clarividentes, para que o enquadramento político permita crescente dimensão à dignidade humana e não crie as tensões sociais que conduzem às revoluções, fenómenos cujo saldo é sempre aleatório.

Nos períodos revolucionários quem legisla encontra-se em situação ainda mais complexa. Terá de reconhecer uma ou mais fases agudas em que o processo revolucionário se projecta com aceleração tal que muitas vezes uma lei, no momento em que se publica, é apenas a baliza de um trajecto ultrapassado.

O amadurecimento do fenómeno revolucionário cria fases de menor aceleração, em que os legisladores poderão já prever e implantar marcos legais suficientemente avançados, que não sejam balizas do passado mas pontos firmes de apoio, em que o processo revolucionário sente a segurança e o equilíbrio a que o ser humano aspira.

Retenhamos daqui a necessidade de promulgar marcos legais com seguro avanço.

Outra característica da vivência em clima revolucionário é uma constante flutuação no rumo do processo, o qual é a integração de vontades individuais, de fenómenos de massa, das correntes políticas empenhadas, das acções e reacções mais inesperadas.

Cada ponto de um processo revolucionário é o momento histórico da síntese convulsional entre a situação que se ultrapassa e um somatório dos anseios vitais das forças revolucionárias em luta.

Em revolução, na génese de uma lei avançada, interessa eliminar o risco de ver a dinâmica do processo torná-la inútil por passagem flanqueante.

[157]

Figure 1: Inaugural session transcript of the Portuguese parliamentary debate[1].

Recently, researchers from different fields started studying debates similar to the one presented in Figure 1 (ABERCROMBIE; BATISTA-NAVARRO, 2019). Those who analyse this data using Natural Language Processing, or more specifically, one of its branches, Sentiment Analysis, are among them.

Sentiment Analysis emerged on the 20th century when the analysis of public opinion

[1]Available on: <https://debates.parlamento.pt>. Accessed on May 3rd, 2021.

and text subjectivity started being made. However, most published studies date from 2004 because this field only spread after subjective texts were made available on the internet (MÄNTYLÄ; GRAZIOTIN; KUUTILA, 2018). As the authors explain in the article, traditionally, it has been used to understand if someone's opinion is positive, negative or neutral regarding a subject through textual data.

Through the analysis of political debates transcripts, it is possible to understand what subjects relevant to society, such as education, healthcare, infrastructure, among others, have been discussed and what politicians have said about these themes along with their positioning.

Furthermore there are far fewer Natural Language Processing tools available to researchers in Portuguese in comparison to more widely spoken languages, such as English. This justifies the creation of a tool that is able to analyse Portuguese text, so that speakers of this language may have an equal opportunity to look further into their country's politics with greater ease.

1.1 Objectives

An important part of the democratic process is to monitor what is discussed, approved, or rejected in the Parliament. This process is essential for citizens to see if parliamentarians are delivering on what they promised to better understand which candidate will receive their vote in the next elections and, in the end, to improve people's living conditions. However, accessing this information is time consuming once it is not presented in a friendly way to the population.

The objective of this study is to use sentiment analysis to analyse the Portuguese Parliament's debates transcripts, understand the politicians' position as positive or negative regarding the initiatives, which represent important themes facing society, and compare their position with their party's vote to analyse the coherence between them. Thus, we are interested not only in defining the position of each politician or party in relation to an initiative, but also in evaluating their coherence in approving or rejecting the initiatives.

It is important to note that, in the course of the development of this project, it became clear that parliamentarians rarely assume a neutral position in debates. The speeches that receive this classification are usually those that ask for clarification on a topic under debate, as exemplified below:

- **Speech:** "Sr. Presidente, Sr. Deputado Pedro Delgado Alves, o Partido Socialista, neste debate, coloca, fundamentalmente, dois problemas: um que tem a ver com a situação em que estamos, com a situação de pandemia em que nos encontramos e com o facto de termos eleições a muito curto prazo, as eleições para a Presidência da República, e outro que tem a ver com outras adaptações ao procedimento eleitoral a que me irei referir quando fizer a minha intervenção, em nome do Grupo Parlamentar do PCP. Mas a questão que lhe queria colocar é se os Srs. Deputados do Partido Socialista não entendem que haveria vantagem em separar estes dois problemas."

  - **Free translation to English:** "Mr. President, Mr. Deputy Pedro Delgado Alves, the Partido Socialista, in this debate, presents fundamentally two problems: one that is related to the situation we are in, with the pandemic context and with the fact we have elections soon, the presidential election, and the other that has to do with adaptations to the electoral procedure that I will refer to when I make my speech on behalf of the parliamentary group of PCP. But the question that I wanted to ask is if the Deputies of the Partido Socialista do not understand that there would be an advantage in separating these two problems."

As the speeches that are neutral to an initiative are not numerically relevant when compared to the frequency of statements with positive or negative sentiments, this project only classifies speeches in binary form, i.e., positive or negative.

A concern in the project was also related to the processing of new information. Since these discussions are constantly happening and new themes emerge according to the ongoing events, to provide the most up-to-date results possible, the tool needs to be able to perform a new analysis as soon as new transcripts and votes data become available.

For the purpose of facilitating the interpretation of the data by the population, the results are presented graphically according to the filters selected, available on the lists "Legislatura" (legislature), "Sessão legislativa" (legislative session), "Projeto de deliberação" (deliberation project), "Iniciativa" (initiative), and "Partido" (party), better explained in Section 4.1.

The target user is anyone who is interested in these discussions and who knows Portuguese, since the transcripts are in this language.

1.2 Background: Portuguese parliamentary debates

The Portuguese Parliament, Assembleia da República, is the legislative body of Portugal. It is formed by one chamber with 230 deputies (PARLAMENTO.PT, 2021), elected to a four-year term, called a "Legislatura", composed of one-year long subdivisions called a legislative session ("sessão legislativa").

The debates and each speech in Assembleia da República are introduced by the president of the Parliament, who conducts the discussion. A motion is a proposal put forward for debate or decision in the Parliament, presented by one of its members, and can be categorized into the following groups:

Parliamentary appraisal ("apreciação parlamentar") : initiative proposed by the parliamentarians to terminate or amend a government ordinance;

Referendum ("iniciativa popular de referendo") : initiative proposed by the Portuguese citizens to the Parliament;

Parliamentary inquiry ("inquérito parlamentar") : parliamentary oversight instrument, in which a commission of inquiry is created to inspect a specific matter;

Deliberation project ("projeto de deliberação") : proposals made by the parliamentarians that are discussed and voted on;

Bill ("projeto de lei") : legislative initiative proposed by any parliamentarian or citizens groups;

Law proposal ("proposta de lei") : legislative initiative proposed by the government or any Legislative Assembly of the autonomous regions;

Resolution project ("projeto de resolução") : initiative proposed by any parliamentarian to regulate the matters within the exclusive competence of the Parliament and those of a political, procedural, legislative or administrative nature;

Resolution proposal ("proposta de resolução") : initiative proposed by the government to the Parliament for the approval of treaties or agreements;

Ratification ("ratificação") : action that can be taken by the President of the Republic, within the scope of international relations regarding international treaties, after being duly approved in the Parliament;

Constitution revision project ("projeto de revisão constitucional") : revision of the text of the Constitution by the parliamentarians;

Regiment project ("projeto de regimento") : amendments to the Rules of Procedure of the Parliament.

Following this, members of the Parliament request to speak and may do so when invited during a delimited time. All the parties can speak out in the discussions but not every one does. The voting may or may not happen in the same day the motion was discussed and is registered as "a favor" to support, "contra" to oppose, or "abstenção" when they abstain. Only the party's vote is registered, the individual vote is not available. The existing parties in the Parliament and the number of parliamentarians in each one in the current legislature are listed below (PARLAMENTO.PT, 2021).

- Partido Socialista (PS): 108 parliamentarians
- Partido Social Democrata (PSD): 79 parliamentarians
- Bloco de Esquerda (BE): 19 parliamentarians
- Partido Comunista Português (PCP): 10 parliamentarians
- CDS - Partido Popular (CDS-PP): 5 parliamentarians
- Pessoas–Animais–Natureza (PAN): 4 parliamentarians
- Partido Ecologista "Os Verdes" (PEV): 2 parliamentarians
- Chega (CH): 1 parliamentarian
- Iniciativa Liberal (IL): 1 parliamentarian
- Livre (L): 1 parliamentarian
- Unregistered deputies: the parliamentarians that leaves the party, but continue to be a member of Assembleia da República

The data related to what each person said during these debates and the votes are publicly available on the Portuguese Parliament website *parlamento.pt* in both PDF and Plain Text formats, which makes it feasible to automatically download them to provide an analysis that is frequently updated, besides using this data in a Sentiment Analysis environment to classify the speeches of the parliamentarians as positive or negative towards the motions and compare with their party's vote.

1.3 Project requirements

To achieve the objectives detailed in section 1.1, the project requirements are:

PR1 : the tool must be able to extract the transcripts from the Parliament website, *parlamento.pt*, where all of the minutes are publicly and freely available;

PR2 : after the release of new voting data, the tool must perform the analysis and make the results available to the public within one day;

PR3 : the tool must have a sentiment classification algorithm with accuracy rate of at least 89% so there is a tool for the Portuguese language as good as there is for English, as shown by (ABERCROMBIE; BATISTA-NAVARRO, 2018);

PR4 : the user interface, detailed in Section 4.1, must be easily understandable, with clear instructions to guide the user, so they know what actions are required from them.

1.4 Organization of the manuscript

The remainder of this document is organized as follows. Chapter 2 outlines some recent works done by others in the same area with similar objectives, and compares the technologies used by them and the results they obtained. Chapter 3 briefly describes the theory involved in the tools used in the project. In Chapter 4 our proposal is detailed, and in Chapter 5 we describe how the project was executed. Chapter 6 presents our results and finally in Chapter 7 we present our conclusions and suggest future work.

2 RELATED WORK

Existing approaches to quantify the emotion contained in texts can be divided into three groups: lexicon approach, machine learning approach, and a hybrid approach that combines both previous approaches (PEREIRA, 2020). The first one aims to measure sentiment by comparing each word in the input text to an entry in a lexicon, which consists of a list of words in a language and their sentiment value, attributing a sentiment to each individual word. Thus, this approach depends on a good linguistic resource to achieve good results. On the other hand, instead of assigning emotions to individual words, the second approach can identify sentiment by the presence of keywords, the proximity between them, and the frequency that these words appear together. This means that different sentiment values may be assigned to a single word, depending on the context in which it was used. However, since a large amount of training data is required and it generally struggles to extract deep meaning from isolated words, the hybrid approach aims to combine the two to exploit the strengths of each one, being the preferable one for this study (PEREIRA, 2020).

In the domain of political debates, Sentiment Analysis has already been used to analyse transcripts. In this application, most previous research has been done in English, and although many articles generally use techniques such as Support Vector Machine (SVM) (THOMAS; PANG; LEE, 2006; BURFOOT, 2008), Neural Networks became popular as a powerful method for Sentiment Analysis (PEREIRA, 2020) and have recently been shown to perform better than SVM when applied to transcripts of debates by (ABERCROMBIE; BATISTA-NAVARRO, 2018). The research that led to this conclusion used a type of Neural Network called Multilayer Perceptron and, as such, it is the preferable method to execute this task.

In Portuguese, however, we couldn't find any previous work that performs Sentiment Analysis on political debates, and there are few linguist resources available (PEREIRA, 2020). However, for other tasks, seven Portuguese sentiment lexicons were built: OpLexicon, Brazilian Portuguese LIWC Dictionary (LIWC-PT), Reli-Lex, LexReLi, SentiStrength,

synsets with polarity of Onto.PT, and SentiLex.

OpLexicon and LIWC-PT were built based on Brazilian Portuguese corpora, Reli-Lex and LexReli are smaller lexicons built from a book review corpus (ReLi), SentiStrength was created to analyse short social web texts in English, except political texts, being adapted later to the Portuguese language, and synsets with polarity of Onto.PT and SentiLex were based on European Portuguese corpora.

Collecting more information about the last two lexicons, since they are closer to the context of this work, synsets with polarity of Onto.PT were automatically created from three electronic dictionaries[1], while SentiLex was mostly manually labeled[2], being more reliable. Also, this first one requires word stemming and lemmatization, as explained in Section 3.2, which depends on another dictionary and could decrease the general performance of the model.

Besides, SentiLex was already used by (BENITES-LAZARO et al., 2020) in a similar study, where the authors examined policy debates related to land, food and water by analysing the sentiment on governmental and business documents, Brazilian newspapers, and the bulletins of non-governmental organizations.

Thus, based on the analysis made in the literature in the area, we define some aspects of the project, namely:

- To use Neural Networks, more specifically the Multilayer Perceptron, in a hybrid approach;
- To use the SentiLex lexicon in hopes of achieving the same or better results than those expressed in (ABERCROMBIE; BATISTA-NAVARRO, 2018), since its study is the most similar approach to this work and its results were the best in the context of political debates.

In the next chapter we provide explanations of the fundamental concepts involved in the methods used in the project.

[1]Information on the READ ME section available after the download on ontopt.dei.uc.pt/index.php?sec=projecto. Accessed on November 21st, 2021.

[2]Available on http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3. Accessed on September 25th, 2021.

3 THEORETICAL FOUNDATION

As described in (MOHAMMAD et al., 2016a), sentiment analysis is a technique used to determine whether a piece of text data is positive, negative or neutral, or to infer the speaker's opinion based on the text.

If someone reads a product review on an online store, it is often obvious whether the experience associated with that product is positive or negative. This is something we can infer from a text through a subjective meaning behind the words: in the example of the product review, we naturally associate words such as "expensive" and "regret" with bad experiences, while "economic" and "good" are associated with pleasure.

This is the objective of sentiment analysis, to classify the text as a whole in predefined categories related to the hidden meaning behind its words. As an example, (WU; PAO, 2012) show that it is possible to predict whether a film critic thinks positively or negatively about a given film through their review.

In order to perform this classification, models called "Classifiers" need to be tuned to fit the data of interest. These concepts will be explained further in the following section. Then more details about the lexicon SentiLex will be explained, followed by the text preprocessing methods used and the performance metrics taken into account.

3.1 Neural Networks

Classification is the statistical problem of identifying to which group or category a piece of data belongs to. In machine learning, the objective of a classifier is to take in a set of attributes and their values in order to categorize it into a group.

Training a classifier means to feed it data so it has patterns to base its classifications on. In general, the output of a classifier gets more accurate as more data is fed into it to train it, since there are more patterns for it to analyse.

Neural networks are a type of classifier. They are made of artificial neurons, with

their inputs and outputs connected together to form a network. Artificial neurons, which can also referred to as *Perceptrons*, consist of mathematical functions created to model biological neurons, with multiple inputs and an output, as shown in Figure 2.

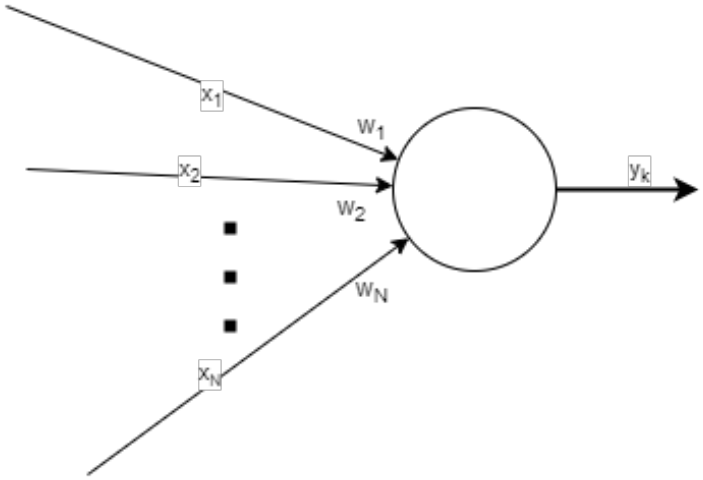

Figure 2: Model of an artificial neuron.

The output y_k of these neurons are related to the inputs x_i via an activation function f, that takes in a linear combination of the inputs with weights w_i (ALZAHRANI; PARKER, 2020),

$$y_k = f\left(\sum_{i=0}^{n} w_i x_i\right). \tag{3.1}$$

The activation function may take many forms, one of them being a Rectified Linear Unit function (ReLU), i.e. a function that is 0 when the input is less than zero, and equal to the input otherwise:

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x >= 0, \\ 0 & \text{if } x < 0. \end{cases}$$

In order to maintain non-linearity throughout the network to mimic human neurons more closely, the input may be modified to include a bias term. In other words, the neuron receives an extra input x_{n+1} with a constant value.

Training the model translates to adjustments of the weights w_i associated with each input x_i, seeking to minimize the error between the expected output y_k (given in the training data and referred to as *label*) and the one currently estimated by the model, $\hat{y}_k$. The function that determines this error is called the Loss Function. One possible Loss Function is the Mean Squared Logarithmic Error:

$$\text{Mean Squared Logarithmic Error}(y_k, \hat{y_k}) = \frac{1}{n}\sum_{i=1}^{n}(\log(y_k + 1) - \log(\hat{y_k} + 1))^2. \quad (3.2)$$

The procedure that determines how the weights are adjusted through the Loss Function is called the Learning Algorithm. Backpropagation with gradient descent is a popular choice for the Learning Algorithm, and it is explained in greater detail in Section 3.2.

The learning may occur through an iterative process by feeding the same training data to the Network multiple times, in order to further reduce the Loss Function. The number of iterations is called the number of Epochs.

A Multilayer Perceptron (MLP) is a type of Neural Network that consists of layers of artificial neurons with no cycle between them through their connections. The Network is built of at least three layers:

- One input layer;
- One output layer;
- Any amount of intermediate layers, called hidden layers.

3.2 Backpropagation

Backpropagation is a Learning Algorithm that determines the adjustment in weights for connections between neurons closer to the end of the net, and propagates back as adjustments are made to tune the entire network.

The adjustment in the weight w_{ij} of the connection between neurons i and j is determined by the following equation:

$$\Delta w_{ij} = -\alpha \frac{\partial L}{\partial w_{ij}}, \quad (3.3)$$

where L is the Loss Function and α is called the Learning Rate. This algorithm functions based on gradient descent: a technique used to find local minimums. In this case, the objective is to find the local minimum of the Loss Function. Since $\frac{\partial L}{\partial w_{ij}}$ gives the direction of greatest differential value, multiplying it by a scaling value, $-\alpha$, and adding the result to w_{ij} allows w_{ij} to get closer to a local minimum.

Tuning of the value of α can have a big impact on how the model functions: if it is too large, the model may oscillate around the local minimum or overshoot it entirely and never reach it; but if it is too small, the amount by which each weight in the network changes may be so little that it requires too much training data for any significant results to become apparent.

The partial derivative can be calculated using the chain rule:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial (\text{net input})_j} \frac{\partial (\text{net input})_j}{\partial w_{ij}}. \tag{3.4}$$

Since only one term of the net input depends on w_{ij}, and considering that $y_j = f((\text{net input})_j)$, where f is the Activation Function, this equation can be changed to:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial y_j} \frac{\partial f((\text{net input})_j)}{\partial (\text{net input})_j} y_j. \tag{3.5}$$

The first term is the derivative of the chosen Loss Function with respect to one of its arguments. The second term is the derivative of the neuron's activation function.

There are also different algorithms to perform this tuning. Optimizers such as the Limited-Memory-Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS), for example, don't use Learning Rates, and they calculate the Jacobian matrix to determine the direction of steepest gradient and reach minimum levels by approximating the Hessian matrix and using it to estimate a step size, with the search direction given by $z = -H_k g_k$, where g_k is the current value of the gradient and H_k is the inverse of the Hessian matrix.

3.3 SentiLex

Sentiment lexicons are a set of words in which a sentiment score is attributed to each one of them. SentiLex is a lexicon that links the words to one number out of three, representing their sentiment: 1 (positive), -1 (negative) or 0 (neutral).

It is composed by 7,014 lemmas and 82,347 inflected forms, describing 4,779 (16,863 inflected forms) adjectives, 1,081 (1,280 inflected forms) nouns, 489 (29,504 inflected forms) verbs, and 666 (34,700 inflected forms) idiomatic expressions.

Lemmas can be interpreted as the most basic form of a word, which is usually found at the head of a definition of a dictionary. For example, "eaten", "eating" and "eat" all have the same lemma: "eat". Inflected forms, on the other hand, include all possible

variations of a word, so “eaten”, “eating” and “eat” are all inflected forms of “eat”.

The majority of the sentiment attributes was manually labeled, but some adjectives were automatically classified by a software developed exclusively for the creation of this lexicon. The inflected forms of the verbs and idiomatic expressions were extracted from the lexicon for Portuguese LABEL-LEX in a semi-automatically way (CARVALHO P.; SILVA, 2021).

3.4 Text Preprocessing

Text preprocessing is used to transform the input data into a format that the model can understand. Below there is a list of the most common text preprocessing methods. All of them are going to be used, except word stemming and lemmatization.

- Punctuation removal: in some models the removal of all of the punctuation must be done. Some lexicons, such as SentiLex, lack punctuation interpretation, so the removal is necessary.
- Lower-Upper case conversion: this conversion is done to normalize and minimize the text data, otherwise the model could understand that a word written with upper case and lower case (for example, *parliament* and *Parliament*) are two different ones.
- Stop words removal: stop words are the ones that link one word to another (for example, *and*, *the*, *a*). They don’t interfere in the sentiment of the text and usually are removed to minimize the text data.
- Word stemming: it refers to the transformation of words to their base form by cutting the inflections. For SentiLex, word stemming is not necessary, since it contains inflected forms of the words.
- Word lemmatization: it refers to the transformation of words to their base form, but, differently from word stemming, word lemmatization does not just cut inflections, it relies on databases to find the correct base form of the words. For SentiLex, it is not needed.
- Word tokenization: it consists of attributing numeric values, called tokens, to every word in a piece of text. The numeric value and meaning of the word don’t necessarily have to be connected, the purpose of this preprocessing technique is only to convert

textual data into numerical data, which is required for feeding text into a Neural Network.

- One-hot encoding: it is used to represent categorical features numerically by creating binary features that represent each category. For example, if a database contains entries with a "Political party" feature that may have values PS and BE, this feature is broken down into two new ones: "Political party_PS" and "Political party_BE". Entries that had PS as their "Political party" will now have a 1 value in the "Political party_PS" feature, and a 0 value in the "Political party_BE", and vice-versa.

3.5 Performance Metrics

Performance metrics are used to evaluate the performance of a model. Suppose the interest is to know if the model is classifying a class A correctly, the metrics are based on the analysis of true positives (TP, right classification as A), true negatives (TN, right classification as not A), false positives (FP, wrong classification as A), and false negatives (FN, wrong classification as not A). The metrics that are going to be used in this study are explained below.

3.5.1 Accuracy

Accuracy is the percentage of right classifications, as shown by its formula below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.6}$$

This is an easily understandable measure and works well for balanced datasets. For unbalanced datasets, accuracy may provide misleading information. For example, suppose that there are more data with a positive sentiment. In this case, the classifier that predicts the positive sentiment would have high accuracy score, but it does not represent the ability of the model to learn from the data.

3.5.2 F1-score

F1-score is a balance between precision and recall. Returning to the initial example, precision measures the percentage of right classifications as A among all the classifications as A made by the model. Its formula is given below.

$$Precision = \frac{TP}{TP + FP}. \tag{3.7}$$

Recall measures how many of the actual positives (actual A) our model capture through labeling it as positive (as A). Its formula is given below.

$$Recall = \frac{TP}{TP + FN}. \tag{3.8}$$

The formula of F1-score is then given by:

$$F1 = 2\frac{precision * recall}{precision + recall}. \tag{3.9}$$

4 PROPOSAL

Along debates, the parliamentarians' speeches show mainly support or disapproval towards a target. Therefore the words used are in agreement with the sentiment they wanted to express. However, to determine if a speech supports or disapproves a motion, it is important to analyse not only the speech, but also the characteristics of the motion.

Contextual features, like the speaker's party and the party that proposed the motion, are indicators of the speech sentiment towards the initiative, since people are more likely to agree with it if their party is the one that proposed the motion. The authors in (ABERCROMBIE; BATISTA-NAVARRO, 2018) analyse the effects of including these contextual features along with the speeches, obtaining better results.

Furthermore, as shown by the same study, when a motion has a negative connotation, a parliamentarian may use negative words to support the initiative or use positive words to demonstrate negative sentiment towards it, as illustrated below.

Motion with negative connotation: "Elimina o banco de horas grupal e por acordo de grupo, a adaptabilidade individual e grupal"

- **Free translation to English:** "Eliminates the time bank group and by group agreement, individual and group adaptability"

Speech to support the motion with negative sentiment: "Sr. Presidente, Sras. e Srs. Deputados: Os horários de trabalho e o seu cumprimento, o respeito pelos tempos de descanso e de lazer, a articulação da vida profissional com a vida pessoal e familiar, o cumprimento e o pagamento das devidas compensações previstas na lei continuam a ser hoje, acentuando-se até no atual contexto, alvo de fortes ataques, colocando-se em causa os direitos dos trabalhadores. Em pleno século XXI, persistem imposições de longas jornadas de trabalho, trabalho suplementar que não é pago, uma profunda desregulação dos horários de trabalho, com consequências na vida quotidiana dos trabalhadores, das suas famílias e na sua saúde também. Alastra, cada vez mais, o abuso da laboração contínua, do trabalho noturno e por

turnos, que, ainda na passada semana, aqui discutimos por proposta do PCP, mas também o prolongamento ilegal de horários de trabalho, os horários concentrados, os atropelos ao descano semanal, as adaptabilidades, os bancos de horas individuais e grupais. Mais horário de trabalho e menos salário — é disto que falamos."

- **Free translation to English:** "Mr. President, Mrs. and Mr. Deputies: working hours and the compliance with them, the respect of rest and leisure time, the work-life balance, the compliance and payment of due compensation guaranteed by law continue to be today, even in the current context, the target of strong attacks, calling into question the rights of workers. In the 21st century, there are still impositions of long working hours, overtime work that is not paid, a profound deregulation of working hours, with consequences in the daily lives of workers, their families and their health as well. The abuse of continuous work, night work and shift work, which we discussed here last week as proposed by the PCP, are spreading, but also the illegal extension of working hours, concentrated hours, the abuse of weekly rest, adaptability, individual and group hour banks. More working hours and less payment — that's what we're talking about."

Speech to oppose the motion with positive sentiment: "Sr. Presidente, Sr.as e Srs. Deputados: Quem já trabalhou fora desta Assembleia sabe que a atividade laboral de uma empresa não é regular ao longo do tempo. Há picos de trabalho que alternam com alturas de atividade mais reduzida. A bem da sua sustentabilidade, as empresas têm de se organizar e de organizar o trabalho em função dessas variações. E uma forma de o fazerem é recorrendo ao pagamento de horas extraordinárias aos trabalhadores, para que estes exerçam a sua função para além do horário normal de trabalho. [...] Para responder a esta realidade é da maior utilidade permitir que o empregador e o trabalhador acordem de antemão, e dentro de certos limites, acréscimos semanais e anuais nas horas trabalhadas, bem como a respetiva forma de pagamento. É isso que o banco de horas pretende atingir e que este nosso projeto de lei contempla."

- **Free translation to English:** "Mr. President, Mrs. and Mr. Deputies: Those who worked outside of this Parliament know that the labor activities of a company are not regular throughout time. There are work peaks that alternate with reduced activities. In favor of sustainability, companies have to organize themselves and organize their work in function of these variations. And one way to do this is to resort to paying extraordinary hours to their

> workers, so that they may exert their functions outside of normal work hours. [...] In order to respond to this reality, it is of greater utility to allow the employer and the worker to agree beforehand, and within certain limits, to weekly and annual increases in work hours, as well as the corresponding form of payment. This is what the bank of hours intends to achieve and that this project of ours project contemplates."

As a solution, the authors compare three separated models, using two methods to label data. As for the latter, one is to use the votes as the sentiment of the speeches and the other is to manually label them. In this study, manually labeled data are going to be used to train the model, being a more reliable method and since the votes are already going to be used at the end in a comparison with the sentiment of the speeches.

Regarding the three models compared by the authors in (ABERCROMBIE; BATISTA-NAVARRO, 2018):

- Model 1: a model that only analyzes and classifies the speeches, without considering the connotation of the motions;
- Model 2: a model that classifies the motion's title and the speeches related to it, and the output is the combination of both classifications;
- Model 3: a model that separates motions by whether the author belongs to the government or the opposition party, classifying opposition as negative and government as positive, and combining the output with the speech classification in the same way as the model 2.

Even though the performance achieved through model 1 is good, it only happened when details about the motion's data were included with the text. When only textual information was considered, model 2 gave better results in comparison to model 1. This suggests that the classifier may be identifying a stronger correlation between the voting results and the motion's context rather than what was discussed about the motion, then the text loses its importance. Therefore, instead of analyzing sentiment, the classifier trained through model 1 is actually predicting votes mostly by analyzing motion information (e.g. making comparisons between the speaker's party and the author's party), which is not the original objective of this study of understanding the sentiment of the speeches and comparing to the parties' votes.

Since there are no reliable sources of information that show the parties that support and oppose the government in an organized and up-to-date structure, this study uses the second model as basis, as shown in the figure 3. In this architecture the user selects the desired initiative and, as the output, the sentiment of the parliamentarian's or party's speeches during the debate of this initiative is shown along with the parties' vote, as explained in section 4.1.

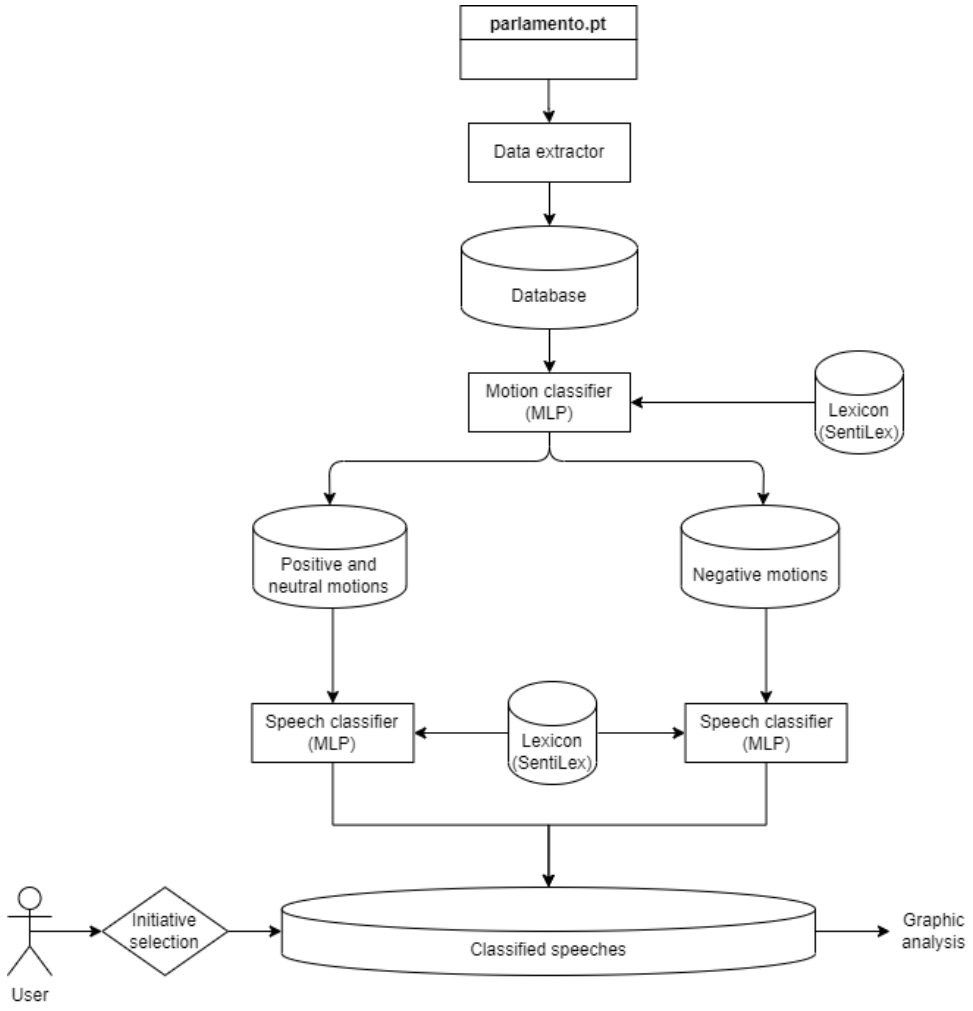

Figure 3: Project architecture. Adapted from (ABERCROMBIE; BATISTA-NAVARRO, 2018).

For both the motion and the speech classifiers, the hybrid approach of MLP plus SentiLex is used. The final sentiment of the text is calculated by the average of the sentiment of your words, given by the lexicon SentiLex.

Due to the limited time available to manually label a sufficient amount of model training data, this study only considers the debates of the current legislature and the first and second legislative sessions, using recent data.

4.1 User interface

There is a large amount of transcriptions of motions and, to present the results in an organized way, the user will be able to filter their search according to the lists "Legislatura" (legislature), "Sessão legislativa" (legislative session), "Projeto de deliberação" (deliberation project), "Iniciativa" (initiative), and "Partido" (party). The proposed user interface is illustrated in Figure 4.

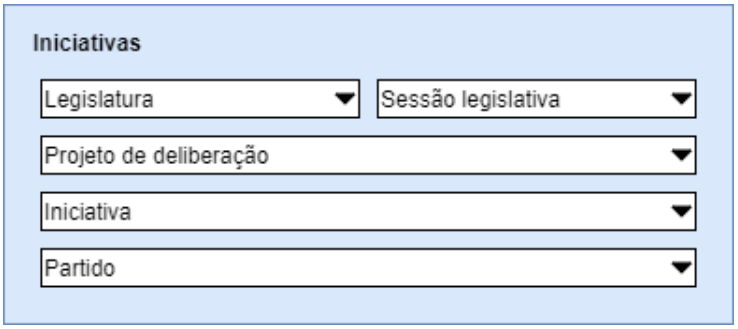

Figure 4: Proposed user interface.

In the "Legislatura" and "Sessão legislativa" lists, the user will be able to narrow the time period, choosing the parliamentary legislature and legislative session, respectively. The "Projeto de deliberação" list enables the person to select the type of the motion (parlamentary appraisal, referendum, bill etc.) and, in the "Iniciativa" list, the user can choose the motion according to its title.

Finally, in the party list, the user will be able to narrow the granularity of the results by selecting a specific party, which will provide a comparison among the position of parliamentarians within that party along with the party's vote, or the option "Todos os partidos" (all of the parties), which will provide a comparison among positions of the parties and their votes, as shown in figures 5 and 6.

In Figure 5, the party's members ("Membros do partido") are represented by the letters W, X, Y, and Z; the position of the parliamentarians ("Sentimentos") is measured by the blue column; and the party's vote ("Voto do partido") is shown by the dotted line, which, in this example, is in favor ("A favor").

In Figure 6, the parties ("Partidos") are represented by the letters A, B, C, and D; the position of the parties ("Sentimentos") is measured by the green column; and the parties' votes ("Votos dos partidos") are shown by the green circle. Here, A and C have positive positions ("A favor"), B has negative position ("Contra"), and D is abstention ("Abstenção"). The position of each party is calculated through the average

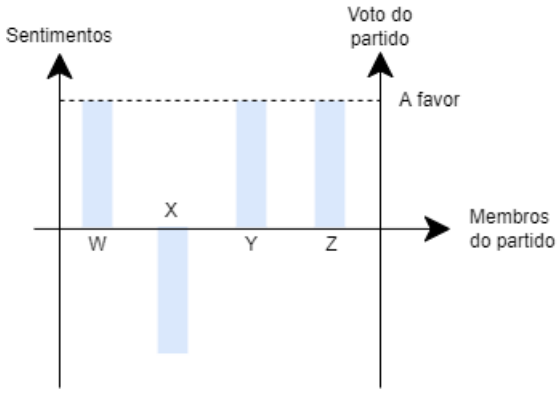

Figure 5: Proposed graph for the specific party selection.

of the positions of its members.

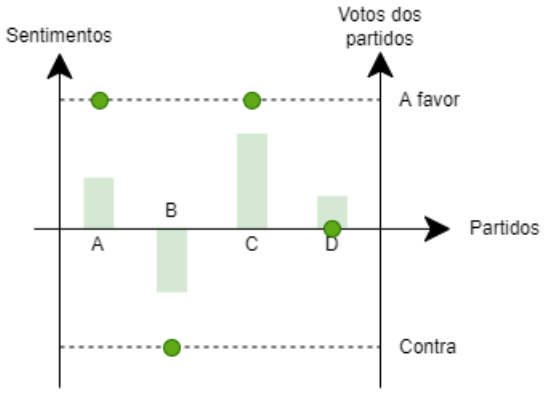

Figure 6: Proposed graph for the option “all of the parties”.

5 PROJECT EXECUTION

The implementation of the project began with the construction of the database by extracting the votes, motions and speeches from the Portuguese Parliament's website. After that, the data was manually labeled and then fed into the lexicon and the MLP classifier. The results of each classifier and their combinations were compared amongst themselves, and the user interface was built to interact with the finalized model.

5.1 Building the database

To see if the members of the Parliament are aligned with their party, this study compares the sentiment of the member speeches with the party's vote. Therefore it is necessary to extract the voting data from the website.

As explained in the chapter 4, the voting data within the current legislature in Portugal was collected, which started on October 25th, 2019 through the Parliament's website (parlamento.pt), illustrated in the figures 7 and 8. The data was divided in:

Title : the motion's title;

Type : the type of the motion being discussed, which can be a parliamentary appraisal ("apreciação parlamentar"), referendum ("iniciativa popular de referendo"), parliamentary inquiry ("inquérito parlamentar"), deliberation project ("projeto de deliberação"), bill ("projeto de lei"), law proposal ("proposta de lei"), resolution project ("projeto de resolução"), resolution proposal ("proposta de resolução"), ratification ("ratificação"), constitution revision project ("projeto de revisão constitucional"), or regiment project ("projeto de regimento"). For the purpose of this project, the last two types will not be considered once they only make changes in government texts, not presenting parliamentarians' opinions about themes impacting the society;

Number : the ID of the motion. The number in roman numerals is the number of the legislation;

Authorship : the party that elaborated the motion;

Session : the legislative session;

URL : part of the url that must be preceded by "https://www.parlamento.pt";

Voting subject : the date the motion was voted and the number of the plenary session;

Status : if the motion was approved or rejected;

Voters : the parties that voted in favor, against, or abstained.

Part of the collected data can be seen in Figure 9.

Figure 7: Motion details page.

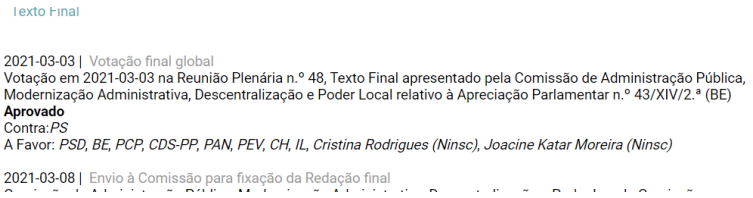

Figure 8: Voting results within motion details page.

	TYPE	NUMBER	SESSION	AUTHORSHIP	TITLE	URL	Voting subject	Status	Voters
0	Apreciação Parlamentar	47/XIV	2	PSD, CDS-PP, CR	Decreto-Lei n.º 25-A/2021, de 30 d	/ActividadeParlar	[]	[]	[]
1	Apreciação Parlamentar	46/XIV	2	PCP	Decreto-Lei n.º 25-A/2021, de 30 d	/ActividadeParlar	[]	[]	[]
2	Apreciação Parlamentar	45/XIV	2	PSD	Decreto-Lei n.º 25-A/2021, de 30 d	/ActividadeParlar	[]	[]	[]
3	Apreciação Parlamentar	44/XIV	2	BE	Decreto-Lei n.º 14/2021, de 12 de f	/ActividadeParlar	[]	[]	[]
4	Apreciação Parlamentar	43/XIV	2	BE	Decreto-Lei n.º 10-A/2021, de 2 de	/ActividadeParlar	['Votação em 2021	['Aprovad	['Contra:PSA Favor: PSD, BE, P(
5	Apreciação Parlamentar	42/XIV	2	BE	Decreto-Lei n.º 6-E/2021, de 15 de	/ActividadeParlar	['Votação em 2021	['Aprovad	['Contra:PS, PSDA Favor: BE, P(
6	Apreciação Parlamentar	41/XIV	2	PCP	Decreto-Lei n.º 8-B/2021, de 22 de	/ActividadeParlar	['Votação em 2021	['Aprovad	['Contra:PS, PSDAbstenção:ILA
7	Apreciação Parlamentar	40/XIV	2	PCP	Decreto-Lei n.º 6-E/2021, de 15 de	/ActividadeParlar	['Votação em 2021	['Aprovad	['Contra:PS, PSDA Favor: BE, P(
8	Apreciação Parlamentar	39/XIV	2	BE	Decreto-Lei nº 8-B/2021, de 22 de	/ActividadeParlar	['Votação em 2021	['Aprovad	['Contra:PS, PSDAbstenção:ILA
9	Apreciação Parlamentar	38/XIV	2	BE	Decreto-Lei n.º 102-D/2020, de 10	/ActividadeParlar	[]	[]	[]
10	Apreciação Parlamentar	37/XIV	2	PCP	Decreto-Lei n.º 102-D/2020, de 10	/ActividadeParlar	[]	[]	[]
11	Apreciação Parlamentar	36/XIV	2	PSD	Decreto-lei n.º 102-D/2020, de 10	/ActividadeParlar	[]	[]	[]

Figure 9: Voting data extracted from the parliament website.

After extracting the voting data, the text relating to each motion must also be extracted. This information is available on the motion's details page on the Parliament's own website, by clicking on any name in the "Intervenções" section, as in figure 10. This will specify exactly which pages of each meeting's minutes contain each person's speeches relating to that motion, as seen in figure 11.

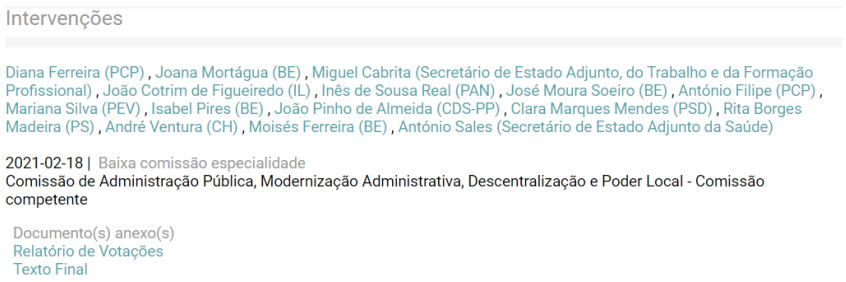

Figure 10: List of speeches related to a motion in its details page.

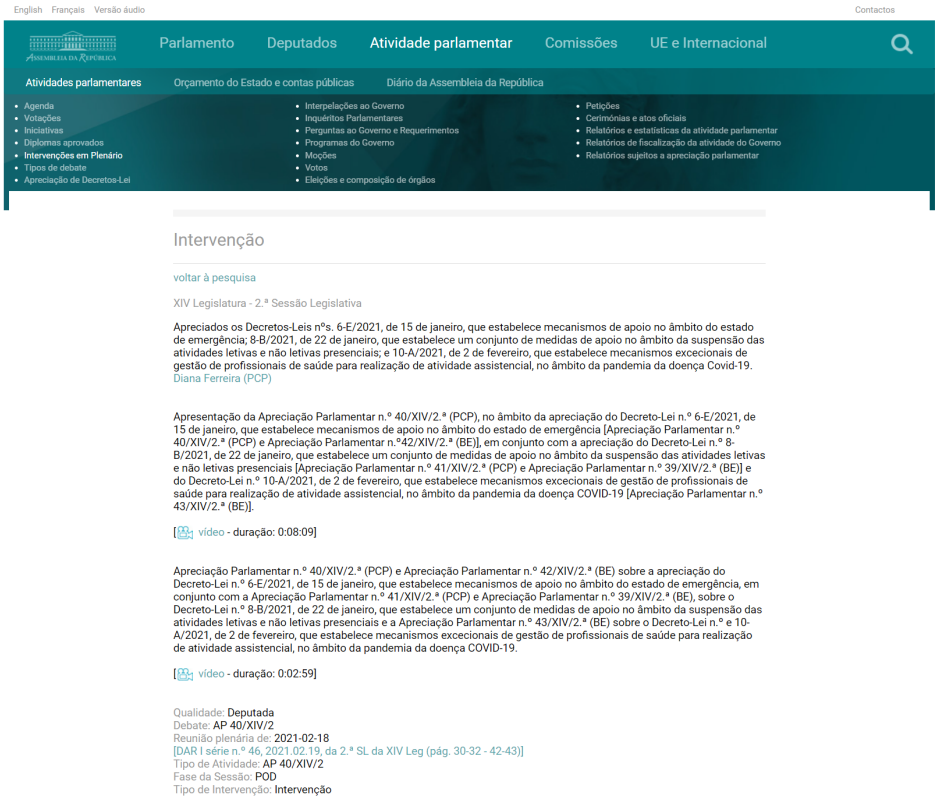

Figure 11: One of the speeches related to a motion, accessed through the links in figure 10. Note the reference to the minutes and its pages at the bottom of the web page.

So, in order to retrieve the textual data, the minutes must be downloaded as well. They can be found in debates.parlamento.pt to be downloaded in PDF or TXT format, as shown in figure 12.

Figure 12: Minutes available in debates.parlamento.pt. At the top of the PDF viewer, it is possible to see the download button for these minutes.

Once this project performs a comparison between the votes and the speeches, only the motions that have already been voted on remain in the database.

Organizing the speeches by parliamentarian, a dataset is formed by grouping the title of the motion ("Title"), its type ("Type"), its ID ("Number"), its author ("Authorship"), the legislative session ("Session"), part of the url that must be preceded by "https://www.parlamento.pt" ("URL"), the party of the speaker ("Party"), the speaker ("Person"), besides the speech ("Speeaches"), as shown in the figure 13.

Title	Type	Number	Session	Authorship	URL	Party	Person	Speeches
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	?	Presidente	Sr.ª Deputada, tem de concluir.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PS	Edite Estrela	Termino já, Sr. Presidente. É injusto e inconstitucional que as mulheres ganhem menos
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	?	Presidente	Chegámos, assim, ao fim deste ponto. Do quarto ponto da nossa ordem do dia consta a
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	?	Presidente	Boa tarde, Srs. Deputados. Vamos, então, passar à apreciação do Decreto-Lei n.º 92/202
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PSD	Bruno Coimbra	Sr. Presidente, Sr.ª Secretária de Estado, Sr.as e Srs. Deputados: Os resíduos, a sua gestã
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Para apresentar a apreciação parlamentar n.º 34/XIV/2.ª, do PCP, tem a palavra a Sr.ª De
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PCP	Paula Santos	Sr. Presidente, Sr.as e Srs. Deputados, Srs. Membros do Governo: O PCP pediu a apreciaç
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Antes de passarmos à fase de encerramento, tendo em conta que ainda há tempos disp
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PSD	Bruno Coimbra	Sr. Presidente, peço a palavra.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Sr. Deputado, o PSD ainda dispõe de tempo no período de encerramento do debate. Ter
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PSD	Bruno Coimbra	Sim, Sr. Presidente.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Tem, então, a palavra, para uma intervenção, o Sr. Deputado Bruno Coimbra.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PSD	Bruno Coimbra	Sr. Presidente, Srs. Deputados: Não deixa de ser interessante ouvir o Partido Socialista fa
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Sr. Deputado, tem de concluir.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PSD	Bruno Coimbra	Concluo já, Sr. Presidente. I SÉRIE — NÚMERO 31 Na verdade, o que é que o Partido Soc
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Inscreveu-se, também para uma intervenção, o Sr. Deputado Nelson Peralta, do Bloco d
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	BE	Nelson Peralta	Sr. Presidente, Srs. Deputados: Quero dar conta de que a proposta do Bloco de Esquerd
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Para uma intervenção, tem, ainda, a palavra a Sr.ª Deputada Paula Santos, do PCP.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PCP	Paula Santos	Sr. Presidente, Srs. Deputados: Nestes poucos segundos que restam, quero dizer que, de
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Tem de concluir. Sr.ª Deputada.
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	PCP	Paula Santos	É aqui, sim, que está a chave desta questão, a questão central no que diz respeito às pol
Decreto-Lei n.º 92/2020, d	Apreciação Parla	34/XIV	2	PCP	/Actividade	António F	Presidente	Srs. Deputados, vamos passar à fase de encerramento do debate.

Figure 13: Motion, speeches by parliamentarian and contextual data extracted from the parliament website.

5.2 Manual labeling

In order to have a more reliable classification of the data, (ABERCROMBIE; BATISTA-NAVARRO, 2018) manually labeled roughly 1,200 speeches to train and test the models. Following that procedure, the same was done in this work, classifying the title of the motion, the speeches, and the positioning of the speeches towards the motion as 1 (positive) or −1 (negative), as illustrated in figure 14. In red it is the president's speeches.

Title	Title sentiment	Type	Number	Session	Authorship	URL	Party	Person	Speeches	Speech sentiment	Speech_Motion sentiment
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Srs. Deputados, concluímos es	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	CDS-PP	Ana Rita Bessa	Sr.ª Presidente, Srs. Deputados	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	BE	Moisés Ferreira	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	PEV	José Luís Ferreira	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	PAN	Bebiana Cunha	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	IL	João Cotrim de Fi	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	PS	Susana Correia	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	Edite Estr	Presidente	Para uma intervenção, tem a p	0	
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	PSD	Rui Cristina	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Repõe a atribuição da bonifi	1	Projeto de Lei	642/XIV	2	IL	/Actividade	PSD	Rui Cristina	Sr.ª Presidente, Sr.as e Srs. Dep	1	1
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	José Man	Presidente	Queira terminar, Sr. Deputado.	0	
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	CDS-PP	João Pinho de Aln	Vou terminar, Sr. Presidente. D	0	0
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	José Man	Presidente	Terminámos assim este ponto	0	
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	CDS-PP	João Gonçalves P	Sr. Presidente, Sr.as e Srs. Dep	1	1
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	José Man	Presidente	Sr. Deputado, queira concluir.	0	
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	CDS-PP	João Gonçalves P	Termino já, Sr. Presidente. Este	1	1
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	José Man	Presidente	Para uma intervenção, tem a p	0	
Apoio aos sócios-gerentes e	1	Projeto de Lei	635/XIV	2	CDS-PP	/Actividade	IL	João Cotrim de Fi	Sr. Presidente, Sr.as e Srs. Dep	1	1

Figure 14: Examples of manual classification of the motions, speeches and the positioning of the speeches towards the motion.

A point to be highlighted is that, during these analyses, the speeches by the president of the sessions stood out due to the large amount of data that have a neutral sentiment, since the president only conducts the debates and does not take a stand for or against. For example, speeches such as "Sr.ª Deputada, tem de concluir" (*Mrs. Deputy, you must conclude*), "Para apresentar a apreciação parlamentar n.º 34/XIV/2.ª, do PCP,

tem a palavra a Sr.ª Deputada Paula Santos" (*To conclude the parliamentary apreciation number 34/XIV/2, from PCP, the word is given to Mrs. Deputy Paula Santos*), or "Para uma intervenção, tem, ainda, a palavra a Sr.ª Deputada Paula Santos, do PCP" (*For an intervention, word is given, still, to Mrs. Deputy Paula Santos, of PCP*) do not add any relevant information regarding the positioning towards the motion in discussion, and only serve to designate the current speaker or topic of discussion and to notify speakers of their time limit to finish their argumentation. For this reason, the president's speeches were removed from the database.

5.3 SentiLex implementation

By utilizing the Corpus SentiLex-PT02, the textual data was classified according to the frequency of positive and negative words in each speech. The classifications were generated by averaging the sentiments of the words contained in the entire speech according to the lexicon, resulting in a value that ranges from -1 to $+1$. Words that were not contained in the lexicon were not accounted for in the average. An example of this application can be seen in the following sentence, said by Manuel Azenha during the discussion of Motion number 612/XIV:

> Pelo que se conclui que a integração da CPAS na segurança social é a garantia de que os direitos sociais de dezenas de milhares de pessoas serão efetivamente respeitados, como devem ser numa democracia, sem discriminação.

- **Free translation to English:** "As can be concluded that the integration of the CPAS in social security is the guarantee that the rights of dozens of thousands of people will be respected effectively, as they must be in a democracy, with no discrimination."

The words "segurança", "direitos", "respeitados" and "democracia" are all classified by the lexicon as being positive. "discriminação", on the other hand, is classified as negative. Thus, the classification of the sentence is $(+1 + 1 + 1 + 1 - 1)/5 = 0.6$.

5.4 Multilayer Perceptron (MLP) Neural Networks implementation

As illustrated in Figure 3, there are two classifiers that have been implemented with MLP: the motion classifier and the speech classifier. Of the two MLPs that had to be implemented, the one that classifies the motion titles did not need further tuning compared to the values obtained in (ABERCROMBIE; BATISTA-NAVARRO, 2018), which means that all of the parameters were conserved:

- Hidden layers: 1 layer with 100 neurons
- Optimization algorithm: L-BFGS
- Number of epochs: Maximum of 200, although if convergence is achieved within a tolerance of 10^{-4} the iterations stop, so the actual number of epochs may be lower
- Activation function: ReLU on every node except for the output layer, which uses a generalization of the sigmoid function called the softmax function to determine the probabilities of each classification, and picking the classification with highest probability
- Loss function: Mean Squared Logarithmic Error
- The rest of the parameters use the default values available in https://scikit-learn.org

Contrary to the title classifier, the speech classifier needed further tuning in order to achieve optimal performance. Through manual tuning, the amount of neurons in the hidden layer was increased to 200, while also using a sigmoid function in the output layer. Changing any other parameters did not improve the best performance scenario for this network. The results obtained with these classifiers is exposed in greater detail in the next section, along with the interface generated for interactivity with the model.

6 RESULTS

In order to achieve the best results from the entire pipeline, the motion classifier was analysed first. The scores obtained are presented in Table 1. In both cases where the lexicon was used on its own and in combination with the MLP, the performance is noticeably worse than when the tokenized text was classified only by the Neural Network. This is due to the fact that the titles of the motions tend to be short, and they use more technical terms that are generally devoid of sentiment on their own, relying on context to define an emotion. Contrary to the lexicon, the Neural Network does not classify isolated words only, and is capable of using context to infer sentiment, which is why it performed better and, thus, the decision to use only an MLP as the motion classifier was made, removing the lexicon from this step of the pipeline. As is evident on Table 2, the same issue does not appear on the speech classifier, since its texts are constructed in the format of dialogue or declarations which are meant to convey intention more clearly throughout the entire speech, so they generally consist of words that don't need much context to be understood, which can be classified by the lexicon.

Table 1: Scores obtained for the motion classifier.

Score \Classifier	Lexicon only	Lexicon + MLP	MLP only
F1-Score	31%	26%	**94%**
Accuracy	31%	64%	**96%**

After it was determined that the motion classifier would use an MLP without the lexicon, the same combinations were tested on the speech classifier, and the results are exhibited in Table 2. The scores presented are those obtained throughout the entire pipeline, by using only an MLP as the motion classifier, and by experimenting combinations of the lexicon and the MLP for the speech classifier. The highest scores were obtained when the Neural Network and the Lexicon were combined in the speech classifier, confirming the theory presented in (PEREIRA, 2020) and discussed in Section 2 that lexicon-based and machine learning-based approaches perform best when used together, instead of separately.

The best combination of features only included information about the title, the speeches, and the parties of both the speaker and the author of the motion. The inclusion of the party of the author provided a noticeable improvement in accuracy and F1-Score, suggesting that the opinion of a speaker may be as dependent on the author of a motion as much as the on the motion itself. This correlation, however, does not occur with the type or ID of the motion being discussed, since the inclusion of this data only hindered the model's performance.

Table 2: Scores of the entire pipeline obtained from combinations of classifiers for the speech classifier.

Features \Classifier	Lexicon only	Lexicon + MLP	MLP only
Speeches and Title only	F1-Score: 71% Accuracy: 82%	F1-Score: 65% Accuracy: 71%	F1-Score: 66% Accuracy: 73%
Speeches, Title and Speaker Party	F1-Score: 71% Accuracy: 82%	F1-Score: 81% Accuracy: 83%	F1-Score: 81% Accuracy: 82%
Speeches, Title, Speaker Party and Author	F1-Score: 71% Accuracy: 82%	**F1-Score: 92%** **Accuracy: 93%**	F1-Score: 89% Accuracy: 89%
Speeches, Title, Speaker Party, Author and Motion Type and ID	F1-Score: 71% Accuracy: 82%	F1-Score: 71% Accuracy: 74%	F1-Score: 68% Accuracy: 73%

The results can be visualized by a web interface, implementing the elements presented in Section 4.1. Figures 15 and 16 show an implementation that allows the user to filter the results by Legislative Session, Party of authorship, Legislature, Type of initiative and Title. There is also a checkbox to toggle between visualization of data separated by parties and by all participants.

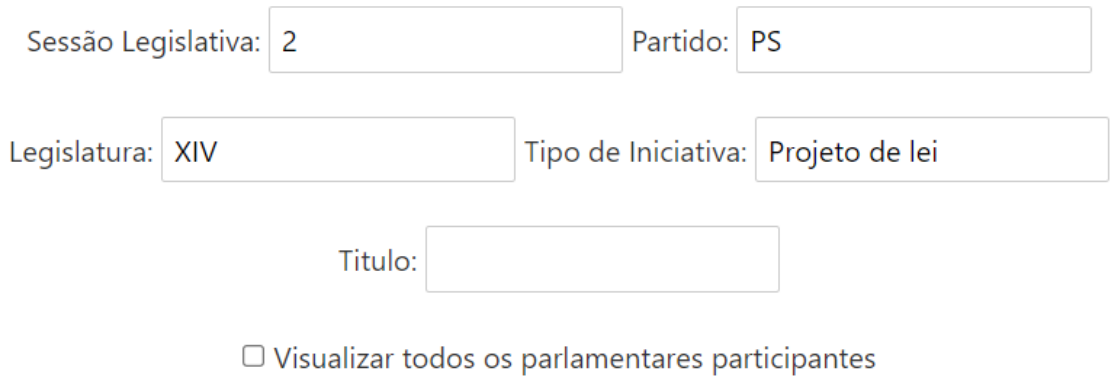

Figure 15: Fields for filtering the motion results.

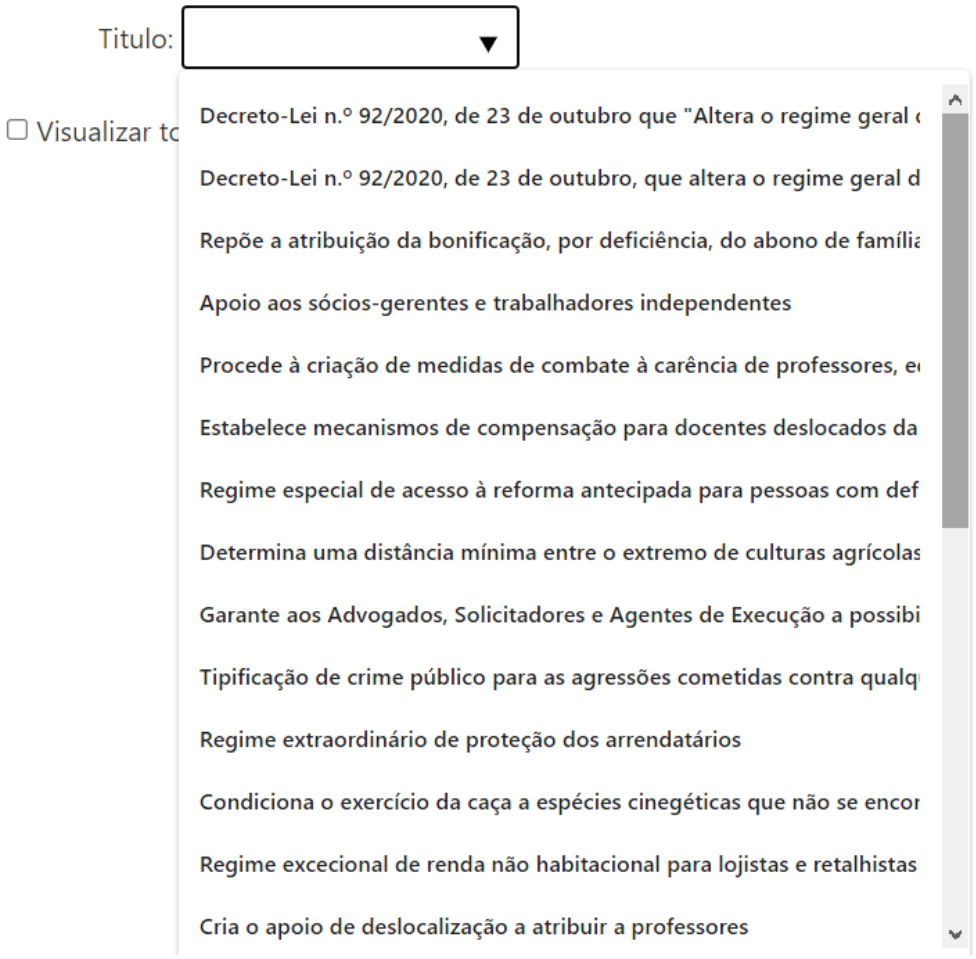

Figure 16: Selection available for the title of the motion shown on screen, which functions analogously to other filtering fields.

Figures 17 and 18 show the visualizations of data separated by all participants and by political parties. The sentiment of a party or person is represented by a vertical bar, and their vote is represented by a dot. Abstention of vote is represented by a dot on the center line. With this layout, it is easy to identify incoherences between speeches and votes, either by abstention of vote or discrepancy between vote and sentiment.

For the visualization of sentiment by parties, the sentiment of each member of each party are averaged to represent the sentiment of the entire party, although this average usually only contains one entry since, on most occasions, only one person speaks on behalf of the entire party.

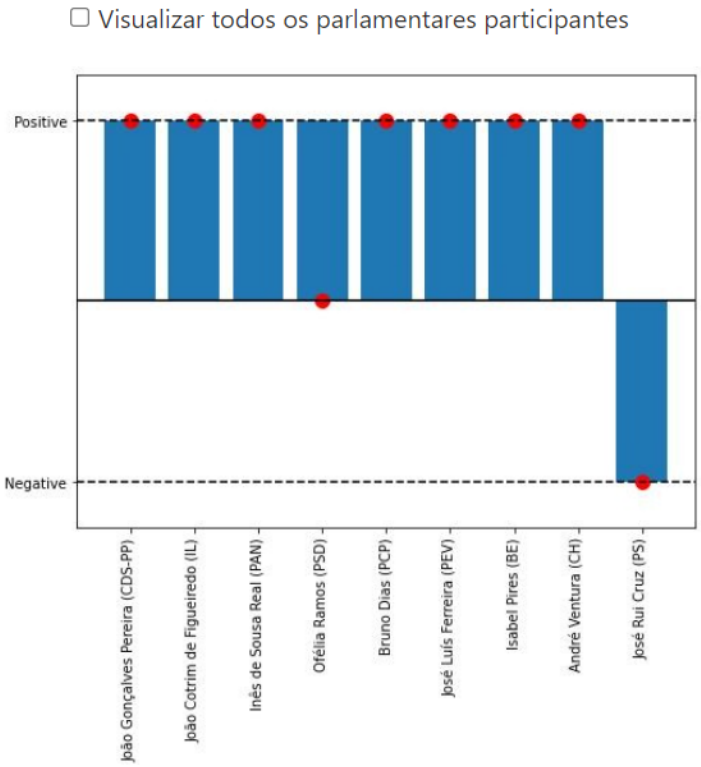

Figure 17: Visualization of data related to an initiative, divided by all of a debate's participants.

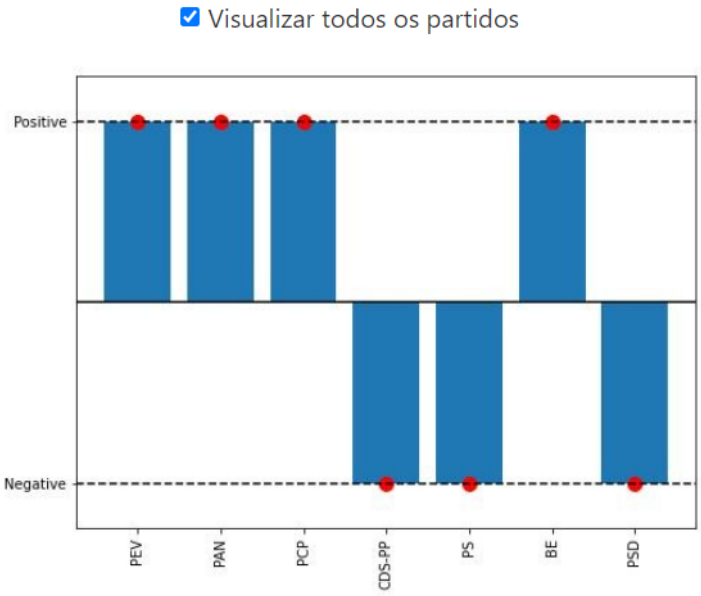

Figure 18: Visualization of data related to an initiative, divided by parties included in the debate.

7 CONCLUSION AND FUTURE WORK

The goal of obtaining an accuracy greater than 89% was successfully achieved, and new insights into the Portuguese Parliamentary Debates were able to be gathered through the analysis of the results obtained, even with limited amount of data. From Table 2, it is possible to see that, by using the same methods, the same scores obtained in previous works could be achieved, and by incorporating different approaches to sentiment analysis these values were improved even further, as was stated in (PEREIRA, 2020). However, it is important to note that the weaknesses of each approach does not disappear, as was observed in Table 1, which suggests that, in order to achieve better performance, both lexicon and machine learning-based approaches need to be able to perform well separately.

Given enough time and resources, more data could be gathered and manually classified in order to continue training these models to increase their accuracy, and the methods utilized in this project could be generalized to other languages to extract information and create better tools to integrate more people into the political scene in the same way as it was done in this work. The combination of lexicons and statistical classification methods can also be explored further to understand the limits of these combinations, and how to optimize their usage so that their combined strength can overcome their individual downsides.

REFERENCES

ABERCROMBIE, G.; BATISTA-NAVARRO, R. 'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts. *European Language Resources Association (ELRA)*, 2018.

ABERCROMBIE, G.; BATISTA-NAVARRO, R. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *arXiv*, 2019. ISSN 23318422.

ALZAHRANI, R. A.; PARKER, A. C. Neuromorphic Circuits With Neural Modulation Enhancing the Information Content of Neural Signaling. In: *International Conference on Neuromorphic Systems 2020*. New York, NY, USA: ACM, 2020. p. 1–8. ISBN 9781450388511. Disponível em: ⟨https://dl.acm.org/doi/10.1145/3407197.3407204⟩.

BENITES-LAZARO, L. L. et al. Land-water-food nexus of biofuels: Discourse and policy debates in Brazil. *Environmental Development*, Elsevier Ltd, v. 33, n. May 2019, p. 100491, 2020. ISSN 22114645. Disponível em: ⟨https://doi.org/10.1016/j.envdev.2019.100491⟩.

BURFOOT, C. Using multiple sources of agreement information for sentiment classification of political transcripts. *Proceedings of the Australasian Language Technology Association Workshop 2008*, p. 11–18, 2008.

CARVALHO P.; SILVA, M. *SentiLex-PT 02*. 2021. URL http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3, Accessed: December 1st, 2021.

MÄNTYLÄ, M. V.; GRAZIOTIN, D.; KUUTILA, M. The evolution of sentiment analysis – A review of research topics, venues, and top cited papers. *Computer Science Review*, v. 27, p. 16–32, feb 2018. ISSN 15740137.

MOHAMMAD, S. M. et al. SemEval-2016 Task 6: Detecting Stance in Tweets. *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, v. 16, 2016a.

PARLAMENTO.PT. *Composição dos Grupos Parlamentares/Partidos. Legislatura XIV, 2019. Accessed: November 30th, 2021*. 2021. Disponível em https://www.parlamento.pt/DeputadoGP/Paginas/GruposParlamentaresI.aspx.

PEREIRA, D. A. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, Springer Netherlands, v. 54, n. 2, p. 1087–1115, 2020. ISSN 15737462. Disponível em: ⟨https://doi.org/10.1007/s10462-020-09870-1⟩.

THOMAS, M.; PANG, B.; LEE, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, p. 327–335, 2006.

WU, J. Y.; PAO, Y. *Predicting Sentiment from Rotten Tomatoes Movie Reviews*. [S.l.], 2012. URL http://nlp.stanford.edu/courses/cs224n/ 2012/reports/WuJean_PaoYuanyuan_224nReport.pdf.