UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS
FACULDADE DE LETRAS
FACULDADE DE MEDICINA
FACULDADE DE PSICOLOGIA

Recognizing Emotions in Short Texts

Iolanda Mafalda Dias Pastor Vieira

Mestrado em Ciência Cognitiva

Dissertação orientada por:
Doutor João Ricardo Martins Ferreira da Silva
Prof.ª Doutora Patrícia Paula Lourenço Arriaga Ferreira

2021

Agradecimentos

Escrever esta dissertação durante uma pandemia, em isolamento, grávida e com crises de ansiedade constantes não foi uma tarefa fácil. Foram muitos os contratempos e muitas as vezes em que achei que não ia conseguir concluir este trabalho. Teria sido, sem dúvida, impossível fazê-lo sem apoio e, por esse motivo, considero esta uma das partes mais importantes desta dissertação.

Em primeiro lugar, quero agradecer aos meus orientadores, Professor Doutor António Branco, Professor Doutor João Silva e Professora Doutora Patrícia Arriga. O apoio, a disponibilidade, a paciência e a compreensão foram maiores do que poderia pedir. Infelizmente, não me foi permitido colocar o nome dos três como orientadores. Apesar de ter informado o Gabinete de Estudos Pós-graduados da minha intenção de ter três orientadores no momento do registo da presente dissertação, em maio de 2021, apenas me foi dito que tal não era possível após a conclusão e entrega da mesma. Assim sendo, o Professor Doutor António Branco voluntariou-se para que o seu nome fosse retirado deste documento como orientador, apesar de todo o seu trabalho e tempo dispensado a orientar-me, os quais foram fundamentais para a realização desta dissertação. Muito obrigada aos três!

Agradeço também a oportunidade de ter feito parte do grupo NLX e toda a ajuda que me foi dada pelas pessoas que dele fazem parte. Foi, sem dúvida, fundamental para a realização deste trabalho.

Não posso deixar de agradecer também aos meus companheiros de quatro patas, que estiveram sempre ao meu lado, a inspirarem-me e animarem-me, na realização deste trabalho em tempos de isolamento, Leonardo, Pitágoras, Ísis, Kali e Nix.

À minha tia Sandra, agradeço por ter estado sempre presente e por, na ausência da sua irmã, me ter dado todo o apoio que ela não teve a possibilidade de dar.

Ao meu amor, meu companheiro, meu melhor amigo, Nuno, nunca poderei agradecer o suficiente por ter escolhido dividir a sua vida comigo e por ter estado ao meu lado em todos os momentos em que dele precisei.

Ao Xavier, o mais lindo bebé, agradeço por me fazer sorrir todos os dias, mesmo quando tudo parece difícil e me dar a motivação extra de que precisava.

Resumo

O reconhecimento automático de emoções em texto é uma tarefa que mobiliza as áreas de processamento de linguagem natural e de computação afetiva, para as quais se pode contar com o especial contributo de disciplinas da Ciência Cognitiva como Inteligência Artificial e Ciência da Computação, Linguística e Psicologia. Visa, sobretudo, a deteção e interpretação de emoções humanas através da sua expressão na forma escrita por sistemas computacionais.

A interação entre processos afetivos e cognitivos, o papel essencial que as emoções desempenham nas interações interpessoais e a crescente utilização de comunicação escrita online nos dias de hoje fazem com que o reconhecimento de emoções de forma automática seja cada vez mais importante, nomeadamente em áreas como saúde mental, interação pessoa-computador, ciência política ou marketing.

A língua inglesa tem sido o maior alvo de estudo no que diz respeito ao reconhecimento de emoções em textos, sendo que ainda existe pouco trabalho desenvolvido para a língua portuguesa. Assim, existe uma necessidade em expandir o trabalho feito para a língua inglesa para o português.

Esta dissertação tem como objetivo a comparação de dois métodos distintos de aprendizagem profunda resultantes dos avanços na área de Inteligência Artificial para detetar e classificar de forma automática estados emocionais discretos em textos escritos em língua portuguesa.

Para tal, a abordagem de classificação de Polignano *et al.* (2019) baseada em redes de aprendizagem profunda como Long Short-Term Memory bidirecionais e redes convolucionais mediadas por um mecanismo de atenção será replicada para a língua inglesa e será reproduzida para a língua portuguesa. Para a língua inglesa, será utilizado o conjunto de dados da tarefa 1 do SemEval-2018 (Mohammad *et al*., 2018) tal como na experiência original, que considera quatro emoções discretas: raiva, medo, alegria e tristeza. Para a língua portuguesa, tendo em consideração a falta de conjuntos de dados disponíveis anotados relativamente a emoções, será efetuada uma recolha de dados a partir da rede social Twitter recorrendo a hashtags com conteúdo associado a uma emoção específica para determinar a emoção subjacente ao texto de entre as mesmas quatro emoções presentes no conjunto de dados da língua inglesa que será utilizado. De acordo com experiências realizadas por Mohammad & Kiritchenko (2015), este método de recolha de dados é consistente com a anotação de juízes humanos treinados.

Tendo em conta a rápida e contínua evolução dos métodos de aprendizagem profunda para o processamento de linguagem natural e o estado da arte estabelecido por métodos recentes em tarefas desta área tal como o modelo pré-treinado BERT (*Bidirectional Encoder Representations from Tranformers*) (Devlin *et al*., 2019), será também aplicada esta abordagem para a tarefa de reconhecimento de emoções para as duas línguas em questão, utilizando os mesmos conjuntos de dados das experiências anteriores.

Enquanto a abordagem de Polignano *et al.* teve um melhor desempenho nas experiências que realizámos com dados em inglês, com diferenças de F1-score de 0.02, o melhor resultado obtido nas experiências com dados na língua portuguesa foi com o modelo BERT, obtendo um resultado máximo de F1-score de 0.6124.

Palavras-chave: reconhecimento de emoções, processamento de linguagem natural, aprendizagem automática, redes neuronais

Abstract

Automatic emotion recognition from text is a task that mobilizes the areas of natural language processing and affective computing counting with the special contribution of Cognitive Science subjects such as Artificial Intelligence and Computer Science, Linguistics and Psychology. It aims at the detection and interpretation of human emotions expressed in the written form by computational systems.

The interaction of affective and cognitive processes, the essential role that emotions play in interpersonal interactions and the currently increasing use of written communication online make automatic emotion recognition progressively important, namely in areas such as mental healthcare, human-computer interaction, political science, or marketing.

The English language has been the main target of studies in emotion recognition in text and the work developed for the Portuguese language is still scarce. Thus, there is a need to expand the work developed for English to Portuguese.

The goal of this dissertation is to present and compare two distinct deep learning methods resulting from the advances in Artificial Intelligence to automatically detect and classify discrete emotional states in texts written in Portuguese.

For this, the classification approach of Polignano *et al*. (2019) based on deep learning networks such as bidirectional Long Short-Term Memory and convolutional networks mediated by a self-attention level will be replicated for English and it will be reproduced for Portuguese. For English, the SemEval-2018 task 1 dataset (Mohammad *et al*., 2018) will be used, as in the original experience, and it considers four discrete emotions: anger, fear, joy, and sadness. For Portuguese, considering the lack of available emotionally annotated datasets, data will be collected from the social network Twitter using hashtags associated to a specific emotional content to determine the underlying emotion of the text from the same four emotions present in the English dataset. According to experiments carried out by Mohammad & Kiritchenko (2015), this method of data collection is consistent with the annotation of trained human judges.

Considering the fast and continuous evolution of deep learning methods for natural language processing and the state-of-the-art results achieved by recent methods in tasks in this area such as the pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al*., 2019), this approach will also be applied to the task of emotion recognition for both languages using the same datasets from the previous experiments. It is expected to draw conclusions about the adequacy of these two presented approaches in emotion recognition and to contribute to the state of the art in this task for the Portuguese language.

While the approach of Polignano *et al*. had a better performance in the experiments with English data with a difference in F1 scores of 0.02, for Portuguese we obtained the best result with BERT having a maximum F1 score of 0.6124.

Keywords: emotion recognition, natural language processing, machine learning, neural networks

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Written language is a means of preserving and transmitting information across time. The amount of available written communication has been exponentially increasing since the advent of the Web and more markedly so since the emergence of social networks, driving the need for automated methods for efficiently handling these vast amounts of language data. The expression of emotions is central to human communication and language is one of the greatest vehicles for achieving this. The sharing of ideas and emotions is particularly easy on social networks and encouraged by their own inherent characteristics. For these reasons, analysing written language becomes increasingly important for the study of emotion expression and social networks are an especially apt domain to do it in.

Natural Language Processing (NLP) is an area that has been receiving a lot of attention and in which a great amount of research and progress has been done in recent years thanks to the advances in Artificial Intelligence. In **Emotion Recognition**, as an NLP task, it should be no different, especially when considering the many applications of its results and the benefits that it can bring to other areas.

In this chapter, we will begin by presenting the motivation for the research reported in this dissertation as well as the objectives that we plan to achieve, followed by the main contributions of this work and its challenges and limitations. Finally, the structure of the thesis will be described.

1.1. Motivation and Objectives

Emotion Recognition is a topic of research that connects and takes advantage of knowledge from different scientific areas related to Cognitive Science, mainly **Artificial Intelligence**, **Computer Science**, **Linguistics** and **Psychology**. More specifically, Emotion Recognition lies at the intersection of **NLP** and **Affective Computing**. NLP aims to develop computational systems that can process written or spoken language data and extract meaning from it, while learning how to deal with the complexity of human language, and Affective Computing is concerned with the recognition, understanding and simulation of emotional states by machines.

The identification and classification of the underlying sentiment in a given text in terms of the affective valence (i.e., as having positive, negative, or neutral polarity) is called **Sentiment Analysis**. However, emotions are complex and emotions with the same polarity may be characterised by different reactions and cause distinct behaviours. Therefore, there is a need of finer discrimination between positive and negative emotions. The task of Emotion Recognition in text emerges from this need, and it

is a natural evolution of sentiment analysis. Emotion recognition in text can then be defined as a multi-class classification task with the goal of classifying emotional states held by people in written utterances using computational systems, more specifically applying NLP techniques.

And why is the study of emotion expression so important? Feeling and expressing emotions play an important role in our lives. Our emotional responses can provoke mental changes and make us act (Plutchik, 2001), affecting our cognitive mechanisms and abilities and having great impact in the way we perceive things or make decisions. Thus, taking into consideration that affective processes are extremely relevant in interpersonal relationships and in our own cognitive processes, the analysis of people's emotional responses, and in particular the ones shared in social network texts, has great interest to several different areas such as human-computer interaction (the recognition of emotions by the computer leads to a better interaction between the computer and the human), mental healthcare (*e.g.* in automatic detection of increased risk of mental disorders such as depression or eating disorders), political science (*e.g.* predicting voting intentions and election results by analysing people's emotional reactions to certain subjects) or marketing (*e.g.* using consumers' emotional reactions to improve marketing strategies). The importance of the study of emotion expression and its many applications is one of the motivations for this work.

As mentioned earlier, investment in Artificial Intelligence has been increasing rapidly and NLP techniques are constantly evolving. The application of recent deep learning techniques that have obtained state-of-the-art results in other NLP tasks and the comparison with other approaches that have previously obtained good results is also one of the motivational components and objectives of this thesis. Although there is a lot of work in Emotion Recognition for the English language, unfortunately it is not possible to say the same for the Portuguese language. In particular, no previous work on this subject for Portuguese and resorting to more advanced Machine Learning and NLP techniques was found. Therefore, another main motivation for writing this thesis is to apply techniques with good results for English in this task, as well as more recent techniques and thus being able to compare both, and to contribute to the state of the art of this task for Portuguese.

1.2. Contributions

The main contributions of this work are the following:

(1) an overview of the different theories of emotion, emotion classification models, and techniques for emotion recognition,

(2) the replication of the Emotion Recognition experiment of Polignano *et al.* (2019) for the SemEval-2018 Task 1 data set using their model that combines a Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Networks (CNN) and a Self-Attention mechanism, since this is the data set that allows an easier adaptation of the data collection methodology for Portuguese,

(3) the implementation of a current state-of-the-art model Bidirectional Encoder Representations from Transformers (BERT) for the same task using the same data set,

(4) the creation of a Portuguese data set composed by emotionally labelled tweet, following the same process used to create the SemEval-2018 Task 1 data set, which consists of automatic data extraction from Twitter and the automatic labelling of tweets using the presence of hashtags and

(5) the application and comparison of both approaches used for the English data set to the Portuguese data set, exploring methods that have never been used before for this language and contributing to the state of the art of emotion recognition in Portuguese.

1.3. Challenges and Limitations

Even when they have emotional content, texts may not always contain words associated with a specific emotional meaning. Emotions can then be expressed explicitly, that is, using emotional-related words, or implicitly, without resorting to words with these characteristics. For example, in the sentence *I feel sad to be away from home* the author is expressing sadness in an explicit way given the presence of the word *sad*. However, emotions are often expressed **implicitly**. For instance, the sentence *It has been a long time since I have been home* can also express sadness as well as longing or nostalgia, but without reference to any words associated with that specific emotion. In this case, although the emotional content of the sentence may not be obvious, the emotional state of the author can have an influence in what and how the sentence is written. One of the interesting challenges of emotion recognition is to correctly identify the underlying emotion to a text not only when it is expressed explicitly, but also when it is expressed implicitly. This makes the task of emotion recognition more difficult, but also more relevant as implicitness is frequent in human language.

In addition to this challenge, there are other factors that make it difficult to recognize emotions in text. Regarding difficulties imposed by language, we can also refer to **ambiguity** (*e.g.*, *He behaved badly* vs *I want to buy a house so badly* – lexical ambiguity of the word *badly*, which can have a negative or a positive meaning), to the use of metaphors (*e.g.*, *I am dying to meet you later – dying* in this sentence does not have its literal sense, having a positive meaning) or to irony and sarcasm (*e.g.*, by stating the opposite of what is meant), just to give some examples.

Another important factor that influences the ability to recognize emotions in text is the **absence of prosodic cues,** such as pauses, intonation or rhythm which are present in spoken communication, and non-verbal cues, since people express emotion not only through verbal communication, but also through non-verbal communication such as facial expressions and body language or even physiological signals (Metri *et al.*, 2011).

However, we do not find difficulties when recognizing emotions just because of the complexity of human language. It is also important to mention the complexity of human emotions and their expression. One of the first challenges that we face is in the very definition of *emotion*. In fact, the scientific definition of emotion is not consensual.

Another challenge is regarding the subjective aspect intrinsic to emotions. Different people can have different interpretations of what a certain emotion is and what it is to experience it. More than that, the same emotion can have different meanings for the same person depending on the situation. Sometimes we may not even know exactly what we are feeling, or we may experience more than one emotion at the same time.

However, all these difficulties are not exclusively encountered by computational systems. The recognition of emotion in text by humans presents the same difficulties. It is therefore important to accept and try to deal with these difficulties rather than to find ways to avoid them.

In particular, regarding the usage of machine learning techniques on the social media domain, another challenge concerns the quantity and quality of the available data. Data annotated with emotions is scarce and, when communicating in social networks, and in Twitter in particular, abbreviations are

often used, and it is not uncommon to find spelling mistakes. These factors can make recognizing emotions more difficult.

Time and cost constraints can also make this task more challenging. One of our goals is to create a Portuguese data set as similar as possible to the data set used by Polignano *et al.* (2019), since there was no previously available emotion annotated and similar data set. Despite having a similar data collection method, the SemEval-2018 task 1 data set were later annotated manually, which unfortunately was not possible to accomplish for Portuguese in the scope of the current work.

These limitations are addressed later in the dissertation when discussing the results.

1.4. Structure of the Thesis

The structure of this thesis is provided below.

Chapter 2 – Background.

Cognitive Science is a multidisciplinary field. This chapter provides the theoretical foundations of Psychology, Artificial Intelligence and NLP necessary for this work.

Chapter 3 – Related Work.

This chapter presents an overview of relevant work done in the field of emotion recognition, as well as current state-of-the-art approaches.

Chapter 4 – Data Sets and Models.

This chapter addresses the data and the tools used in this work. In the first part, it describes the data sets and the data collection methodology that was used. The second part describes the approaches used for this work, namely the BiLSTM, CNN and self-attention mechanism used in the Polignano *et al.* (2019) approach and the BERT language model.

Chapter 5 – Results and Discussion.

In this chapter, we provide the results of the experiments and the discussion of these results.

Chapter 6 – Conclusion.

This chapter concludes this dissertation presenting some final remarks and discusses possible directions for future work.

Chapter 2
Background

This chapter presents the background from the areas of Psychology and Artificial Intelligence that are relevant to the understanding of this work. The chapter is divided into two sections, the first describing relevant techniques for natural language processing (NLP) tasks and the second dealing with emotion models.

2.1. Natural Language Processing and Deep Learning

Two different models have been implemented for the task of emotion recognition in text: 1) a model that combines Bidirectional Long Short-Term Memory, Convolutional Neural Networks and Self-Attention (Polignano *et al.*, 2019) and 2) the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al*., 2019).

In this section, the key aspects of Natural Language Processing and Deep Learning that are relevant to our work will be described. First, neural networks such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), Bidirectional LSTMs and Convolution Neural Networks (CNNs) and the concepts of Attention and Self-attention will be introduced as they are essential elements for the approach of Polignano *et al*. (2019).

Another important element of this approach that will be described in this chapter is the process in which the text data is transformed into distributed representations of words and phrases that contain syntactic and semantic information – word embeddings.

Finally, the Transformer architecture and BERT will be presented.

2.1.1 Recurrent Neural Networks

Human languages expressions are hierarchical and sequential in nature and the word order and the context in which they occur are important to establish syntactic dependencies and to encode meaning. Artificial neural networks with feedforward connections where the information flows forward in the network until the output nodes without feedback connections do not consider the impact that time steps have on language processes.

Unlike this type of neural networks, **Recurrent Neural Networks** (RNNs) proposed by Elman (1990) can handle sequential input, like language expressions, by having memory which allows the extraction of temporal dependencies. RNNs contain feedback connections that allow to capture information from previous states and feeding outputs of the network back to itself. So, useful past information that was learned by the network will be incorporated in subsequent states and will influence predictions.

To calculate the vector of a state at time t s_t, the network will receive as input both the vector of the previous state s_{t-1} and the input vector at time t x_t. The vector s_t is given by the following equation:

$$s_t = F(s_{t-1}, x_t, \theta) \tag{2.1}$$

$$s_t = W_s \sigma(s_{t-1}) + W_x x_t + b \tag{2.2}$$

where W_s is recurrent weight matrix, W_x is the input weight matrix, b is the bias term and σ is a nonlinear activation function. The state s0 is defined by the user.

To train a neural network, we need to minimize an objective function called **cost function** or **loss function** to improve its performance by using a gradient-based method called gradient descent. The gradient descent algorithm will reduce loss by calculating the gradient of the loss function and moving in the direction of the negative error gradient with a fixed-size step called the **learning rate**, by adjusting the network weights. Backpropagation through time (BPTT) (Werbos, 1990) is one of the algorithms used to backpropagate the error of the loss function through the network to calculate weight partial derivatives of RNNs that has proven to be efficient.

Although RNNs can store past information providing them some sort of memory, they still face problems to establish long-term dependencies. When a RNN backpropagates the error gradients, partial derivatives will be continuously multiplicated and error gradients will tend to vanish. This is called **the problem of vanishing gradients** and it means that for long sequences, the gradient changes happening later in the sequence will tend to be very low and thus not affecting the weights earlier in the sequence. The following section will describe a variation of RNNs that solves this problem.

2.1.2 Long Short-Term Memory and Bidirectionality

As previously stated, the context in which words occur has a great importance in assessing their meaning. However, RNNs face the problem of vanishing gradients, as mentioned in the previous section. Given this problem, a gating-based network called **Long Short-Term Memory** (LSTM) was proposed by Hochreiter and Schmidhuber (1997) that can deal with long-term dependencies and use previous contextual information even if it is far away in a sentence. LSTMs can learn these dependencies due to the addition of **memory cells** and **gate units** in each layer of the network.

The memory cell stores information that can later be passed on if it is relevant, providing memory of information in long sequences to the network in addition to the short-term memory that RNNs already have. The gate units have an important role in avoiding perturbation by irrelevant inputs. More specifically, an **input gate** is used to decide when to protect the information inside the memory cell or when to add new information and an **output gate** is used to access the information kept in the memory cell when necessary and to protect other units from being affected by it otherwise. The **forget gate** was a later addition (Gers *et* al., 2000) and it allows the memory cell to forget unnecessary past information by resetting itself.

Until now, we described how to deal with contextual dependencies of previous words. However, a given word or phrase can be dependent on following context. To include contextual information that appears later in text, bidirectionality was proposed by Graves and Schmidhuber (2005). A bidirectional network concatenates the results of two independent networks, one fed with the original sequence and another fed with the reverse of that sequence.

2.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are artificial neural networks proposed by LeCun (1989) and inspired by the visual cortex. As such, they have great success in tasks that involve detecting visual features like image classification, but they can be used in other fields, in particular in NLP tasks.

CNN architectures include an input layer, an output layer, and hidden layers such as convolution layers, pooling layers, and fully connected layers. In a **convolutional layer** a filter or kernel is applied to the input by a convolution operation to obtain a feature map which represents the presence of certain features in the input. A **pooling layer** is used to reduce the dimensionality of the feature map and thus reducing computational costs. Finally, the output is predicted by feeding the result of convolutional and pooling layers to the **fully connected layer**.

2.1.4 Word Embeddings

Word Embeddings are used to represent meaning of words in natural language processing. This is a distributed semantic representation of words which means that it is obtained from the context in which the word occurs.

According to the Distributional Hypothesis, words that have similar meanings tend to have similar distributions, that is, tend to appear in similar contexts (Harris, 1954). This representation of words is made in the form of vectors of real numbers capturing the similarity of meaning of two words by the proximity of the two vectors that represent those two words in the vector space.

In Chapter 4, we will describe the word embeddings used in this work.

2.1.5 Bidirectional Encoder Representations from Transformers

BERT is a language model introduced by Devlin *et al*. (2019) that achieved state-of-the-art results in several natural language processing tasks, in general, and promising results in emotion recognition, in particular, as will be discussed in Chapter 3.

BERT uses a bidirectional transformer pre-training model architecture as shown in Figure 2. 1. The **Transformer** (Vaswani *et al*., 2017) has an encoder-decoder architecture that abandons the use of recurrent and convolutional networks, and it is the first model to rely solely on **attention** mechanisms introduced by Bahdanau *et al*. (2015) that were meant to improve contextual awareness. The attention mechanism will allow the network to focus on relevant contextual information to capture contextual relationships between words in a sentence.

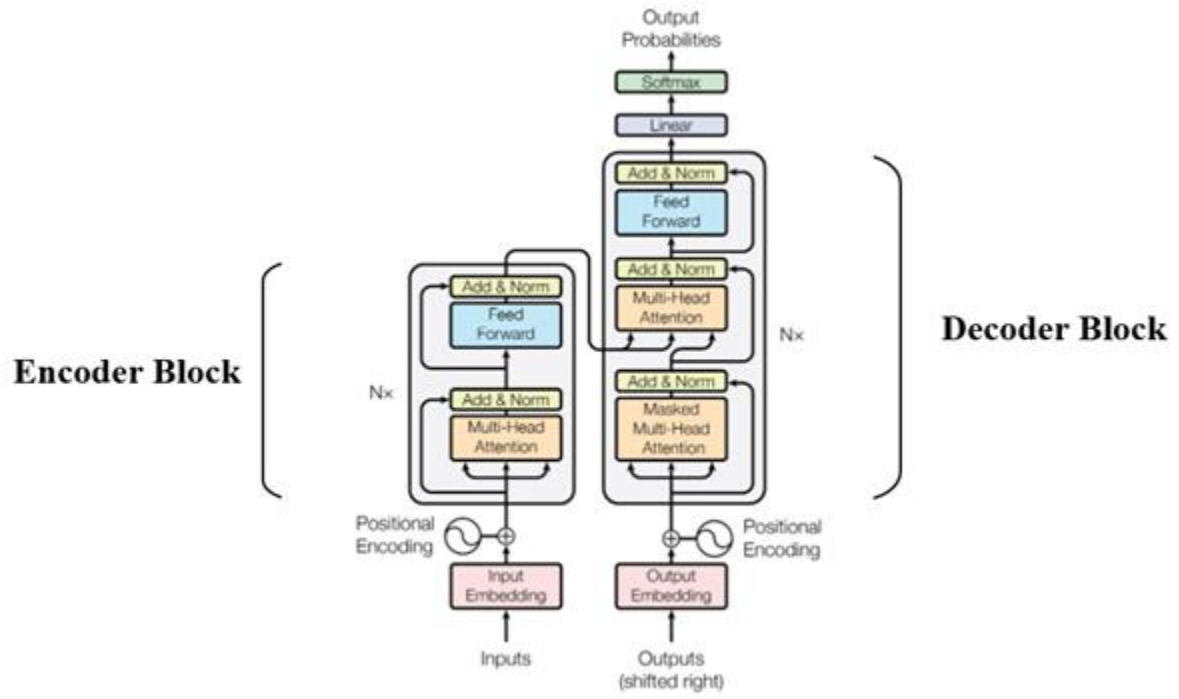

Figure 2. 1 Transformer architecture (Vaswani *et al.*, 2017)

BERT's model architecture is a stack of layers of Transformer encoders. Each encoder layer is composed by two sub-layers: a **multi-head self-attention** layer and a feed forward neural network. In the multi-head self-attention layer, several attention layers run in parallel creating different representations of the input that will be combined based on the scaled dot-product attention.

Regarding input representations, BERT uses pre-trained **WordPiece embeddings** (Wu *et al.*, 2016) and combines them with segment embeddings and positional embeddings as shown in Figure 2. 2. Segment embeddings give specific information about the sentence which the words belong to, and the positional embeddings give information about the location of a word within the sentence. Two special tokens are also introduced: a classification token [CLS] at the beginning of every sequence that will be used for classification by accumulating the language information in the sequence and a token to separate different sequences [SEP].

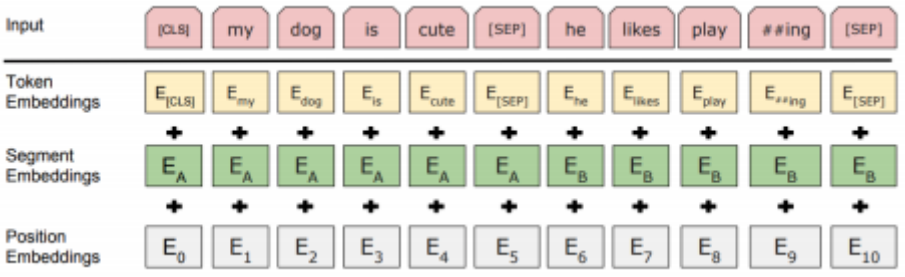

Figure 2. 2 BERT embeddings (Devlin *et al.*, 2019)

There are two phases of training BERT: pre-training and fine-tuning (Figure 2. 3). The **pre-training** phase will help BERT to learn the language and the context by training on two different unsupervised tasks simultaneously. These two tasks are: 1) a masked language modelling where random

words are masked in a sentence and the goal of this task is to guess what the masked words are which will help BERT to capture the bidirectional context within a sentence and 2) a next sentence prediction task where two sentences are given, and the goal is to predict if the second sentence follows the first which will help in learning context across different sentences. With these two different tasks, BERT gets a greater understanding of context.

In the **fine-tuning** phase, the model is initialized with the pre-trained parameters, and they will be fine-tuned to specific NLP tasks by performing supervised training using a labelled data set for the specific task.

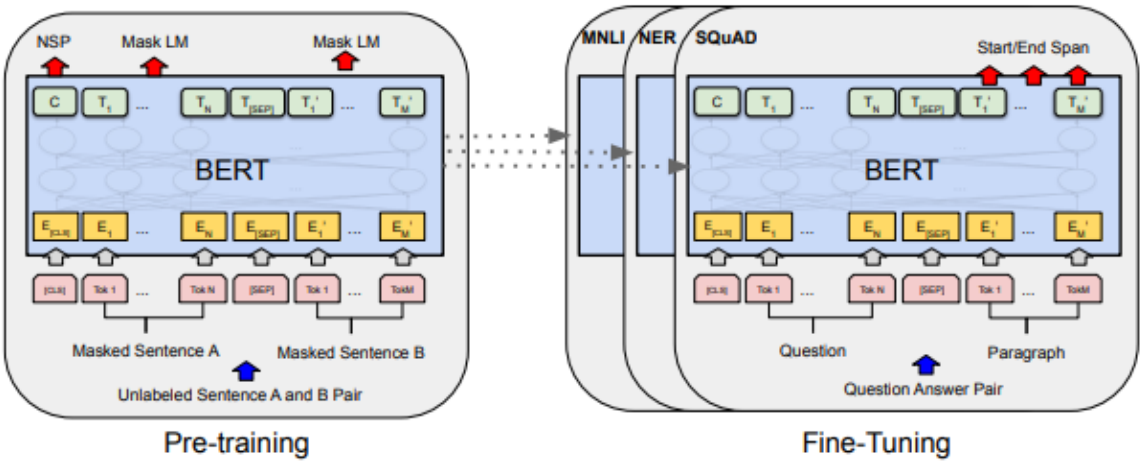

Figure 2. 3 Pre-training and fine-tuning phases (Devlin *et al*., 2019)

2.1.6 Evaluation and Metrics

To evaluate the performance of a given model, it is necessary to have a way to measure it. The definition of the metrics used to measure performance of the classification models will be presented in this section.

Accuracy is defined as the percentage of correct guesses. Though it is an intuitive measure, in a multi-class classification task, it does not show how well the model is predicting each class, which may be important to analyse the results and draw conclusions. Besides, in an unbalanced test data set, it is influenced by the accuracy of the largest classes. For example, if we have a test set with 3 classes A, B and C composed by 80% samples belonging to class A, 10% belonging to class B and 10% belonging to class C and a model that predicts class A for all samples, we will have an accuracy of 80%, but not a very good model. For these reasons, the use of other metrics that can give information about the performance of the model on each class is essential such as Precision, Recall and the F1-score.

When the model correctly classifies an occurrence of a given class, this is called a true positive result for that class. On the other hand, when the model classifies an occurrence of class A as class B, this is called a false negative for class A and a false positive for class B.

Precision reflects how the model can identify occurrences of other classes as non-members of a specific class. It gives us the proportion of correct guesses (true positives) in all predictions made that belong to that class (true and false positives).

Recall is a measure that allows us to see how well a model can correctly classify a given class, that is the percentage of correct guesses for a specific class (true positives) among all the samples of that class in the test set (true positives and false negatives).

F1-score is defined as a balanced harmonic mean of precision and recall, that considers both false positives and false negatives. With this metric, it is possible to obtain a more balanced measure to evaluate the performance of the model for each class.

These metrics are given by the following formulas for each class:

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2.3}$$

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.4}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.5}$$

2.2. Emotion Models

This section discusses different theories of emotion while seeking to define the concept of emotion. It also presents different theoretical models of emotion classification divided into two main approaches: categorical and dimensional.

2.2.1 Definition of Emotion

Emotions are a central element in our lives that influence our actions and interactions with others and therefore the concept of emotion may seem familiar to us. However, it may be difficult to find a definition of emotion if we are asked to do so. In fact, although emotions and their causes have been studied for a long time, its scientific definition has been a subject of debate and it is not consensual.

One of the first to address the study of emotions in a manner that was grounded in physiology, granting them evolutionary justification, was Darwin (1872), who argued that emotions had adaptive functions and evolved over time with an important role in evolution, namely in survival and reproduction. Also in the 19th century, James (1884) and Lange (1885) proposed physiological theories of emotion arguing that physiological processes and specific physical reactions to external stimuli cause the experience of certain emotions, instead of being the consequences of those experiences. On the other hand, Cannon (1927) and Bard (1928) claimed that physiological changes and the experience of emotions are rather simultaneous, and both are a result of responses of the central nervous system to an external stimulus.

Cognitive appraisal theories (Arnold, 1960; Lazarus, 1966) are another important family of approaches in the study of emotion. In particular, the cognitive-motivational-relational theory (Lazarus, 1991) is a cognitive appraisal theory that states that emotion arises from three main processes: (1) the appraisal, which is the cognitive process that evaluates the significance of the environment for the person's life, (2) the motivation of the person considering their intentions and goals and (3) the relationship between the person and the external events.

Including both physiological and cognitive factors, the two-factor theory (Schachter & Singer, 1962) considers that, although essential, physical arousal is not sufficient in inducing emotional responses and it is the cognition that assigns an emotional label to that arousal.

In order to find a consensual definition of emotion, Kleinginna & Kleinginna (1981) compiled 92 definitions of this concept from the existing literature and categorized them into eleven different approaches namely affective, cognitive, external emotional stimuli, physiological, emotional/expressive behaviour, disruptive, adaptive, multiaspect, restrictive, motivational and sceptical, based on the aspects of emotions that were highlighted in those definitions. The authors then proposed their definition that tried to encompass several aspects of emotion but that could still make the discrimination between emotion and other psychological processes: "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labelling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behaviour that is often, but not always, expressive, goal-directed, and adaptive." (p. 355). This is the definition adopted in this work.

2.2.2 Models for Emotion Classification

Regarding the study and the classification of emotions, there are two major theoretical models: categorical and dimensional.

Categorical models are based on discrete emotional categories. More specifically, there are evolutionary models of emotions that are based on a limited number of specific emotions considered as basic emotions. However, it is not consensual which emotions should constitute this set.

Ekman (1992) believed that there are six specific emotions that are essentially different from other emotions and suggests that the basic emotions are **anger, fear, sadness, enjoyment, disgust,** and **surprise**. Ekman argued that these six emotions are universal to all cultures based on the facial expressions of people and their recognition across cultures (Ekman, 1993) and that they "evolved for their adaptive value in dealing with fundamental life-tasks" (Ekman, 1992, p. 171).

Regarding **dimensional models**, the most used model in emotion recognition is the **Circumplex Model of Affect** (Russell, 1980) that categorizes emotions in a two-dimensional space of **valence** and **arousal** (pleasure-displeasure axis and arousal-sleep axis) with eight different categories in a circular order - **arousal**, **excitement**, **pleasure**, **contentment**, **sleepiness**, **depression**, **misery,** and **distress** (Figure 2. 4). For the creation of this two-dimensional space, thirty-six subjects performed two tasks: (1) a category-sort task, in which they associated twenty-eight emotional words to one of the eight categories presented above and, (2) a circular ordering task, in which they were asked to place the words in a circular order so that the distance between words in the circle would translate the proximity or opposition of each category in terms of affect (Figure 2. 5).

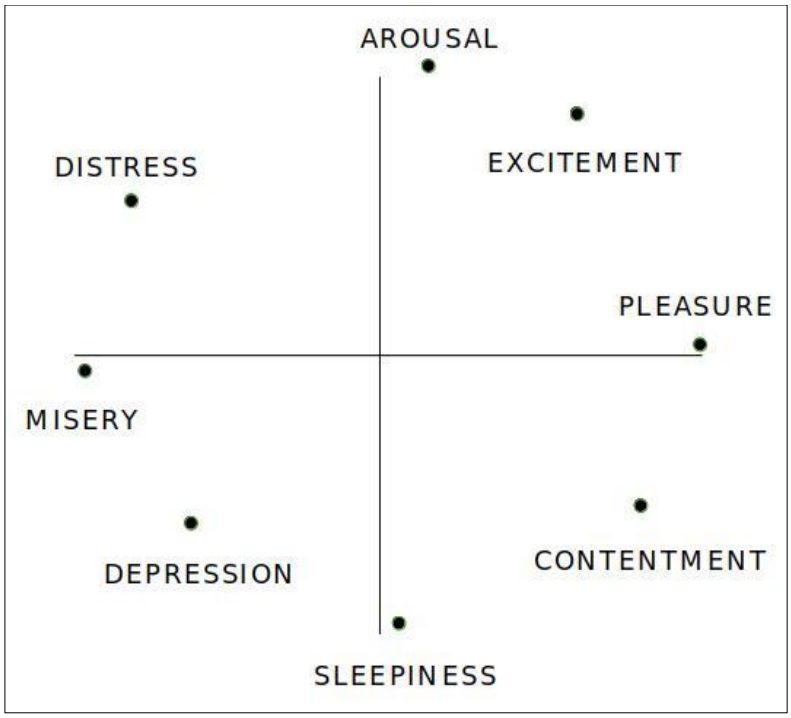

Figure 2. 4 Russell's two-dimensional circumplex model of affect, with 8 categories[1]

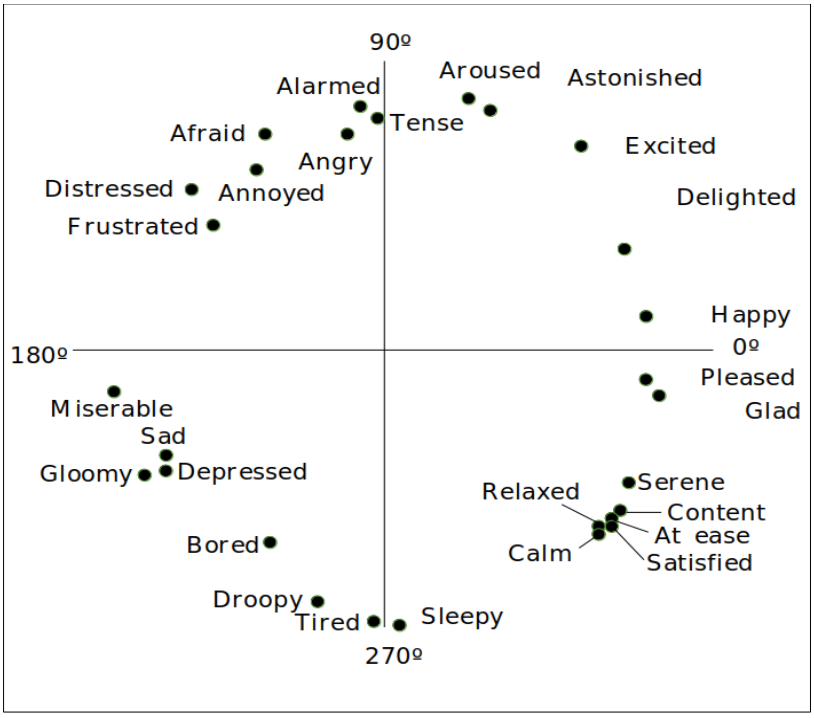

Figure 2. 5 Russell's two-dimensional circumplex model of affect, with 28 affect words[1]

Plutchik's **Wheel of Emotions** (1980) is a hybrid model that describes four pairs of contrasting primary discrete emotions: **joy** vs. **sadness**, **trust** vs. **disgust**, **fear** vs. **anger** and **anticipation** vs. **surprise**. According to Plutchik's psycho-evolutionary theory, these are the eight primary emotions given that they are an evolution of emotions to which the physiological responses allowed a greater

[1] Adapted from Russell (1980).

survival chance to the species. Other emotions could be originated by combining or by changing the level of intensity of primary emotions giving rise to a three-dimensional (similarity, polarity, and intensity) circumplex (Figure 2. 6).

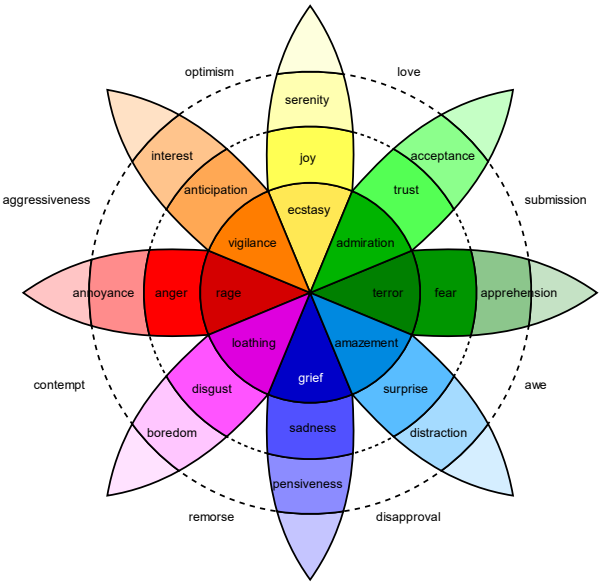

Figure 2. 6 Plutchik's circumplex[2]

In this dissertation, we will address the following emotions: **anger**, **fear**, **joy**, and **sadness** - **joy** is an emotion with positive valence, while **anger**, **fear** and **sadness** are emotions with negative valence. These are the emotions used by Polignano *et al.* (2019) in the experiment that we will reproduce in this work. Furthermore, these emotions are generally considered to be basic emotions by most authors, and some even suggested that there are only four basic emotions – the four emotions chosen in this work (Jack *et al.*, 2014).

[2] https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg

Chapter 3
Related Work

This chapter presents relevant work and the current state of the art approaches for emotion recognition, with a special focus on the English and the Portuguese languages. Since this task requires annotated data but there are few data sets available especially for the Portuguese language, this chapter also includes a section dedicated to relevant emotionally labelled data sets and addresses important work regarding automatic methods for labelling emotions, especially those resorting to social media platforms.

3.1. Emotionally Labelled Data Sets

Data is a fundamental part of any Machine Learning task. Specifically, in supervised learning, the data used to train the algorithm (training data), that is, labelled examples from which algorithms will learn to process information and learn patterns to make right predictions, is extremely important to get a good performance.

Data sets used in emotion recognition can differ in several ways, such as the method used for data collection, the type of text to be classified or the emotion classification model used to label the data. This section identifies some of the most relevant emotionally annotated data sets in the literature and describes their data annotation process.

The International Survey on Emotion Antecedents and Reactions (ISEAR) data set (Scherer & Wallbott, 1994) was built from the results of a cross-cultural questionnaire with 1,096 participants from 37 countries around the world that were asked to report situations in which they had strongly experienced the emotions presented. For this questionnaire seven emotions were chosen: joy, fear, anger, sadness, and disgust as five basic emotions and shame and guilt as self-reflexive emotions that are usually considered specific to the human species. The data set is composed by 7,666 statements and it has approximately the same number of examples for each emotional category.

The Affect Dataset collected by Alm (2008) consists of approximately 176 children's stories and fairy tales from Beatrix Potter, Brothers Grimm, and Hans Christian Andersen. Every sentence of each story was independently annotated by a pair of native US English speakers who attended a literary course on fairy tales or from library science for eight categories: anger, sadness, fear, disgust, happiness, positive surprise, negative surprise and neutral.

SemEval-2007 Task 14 - Affective Text data set (Strapparava & Mihalcea, 2007) is composed of a total of 1,250 news headlines. According to the authors, news headlines were chosen because they

are short, and they are usually written with the purpose of drawing people's attention and inducing emotions which can make them rich in affective and emotional features. The data set is labelled for Ekman's (1992) six basic emotions. The annotation of the headlines was made by six annotators that evaluated each headline for each emotion from 0 to 100, where 0 represents the absence of the corresponding emotion and 100 means that there is a maximum emotional load for that emotion, and for valence in an interval from -100 (highly negative headline) and 100 (highly positive headline).

The Affect in Tweets Dataset (Mohammad *et al*., 2018; Mohammad & Kiritchenko, 2018) was built for the SemEval-2018 Task 1. This task was divided into five subtasks: Emotion Intensity Regression (EI-reg), Emotion Intensity Ordinal Classification (EI-oc), Valence Regression (V-reg), Valence Ordinal Classification (V-oc) and Emotion Classification (E-c). For the EI-reg and EI-oc tasks the same data is used considering four basic emotions: anger, fear, sadness, and joy. However, the labels of the data are different for each task since that EI-reg task aims to classify the intensity of emotion, having real-valued intensity scores as labels, while the EI-oc task aims to classify the tweets into ordinal classes (no emotion, low emotion, moderate emotion, and high emotion). The V-Reg and V-oc tasks data set is composed of 1,200 tweets associated with negative emotions (anger, fear, and sadness) and 1,200 tweets annotated for the positive emotion (joy) taken from the previous mentioned data set and 200 new tweets with hashtags such as #sarcastic, #sarcasm, #irony, or #ironic to study valence of ironic and sarcastic tweets as well. Finally, the E-c task uses the same data as EI-reg and EI-oc, but with different annotation. For this task, the data was manually annotated not for four, but eleven emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust).

To build the Affect in Tweets data set, 50 to 100 emotionally related terms for each of the four emotions were selected and categorized into different intensity levels (*e.g.*, for anger, the authors used terms such as *angry*, *annoyed* or *fury*). These terms included hashtags made of words chosen from *Roget's Thesaurus* with the emotion words and their close synonyms as head words, emojis associated with one of the four emotions (anger, sadness, fear, and joy) and emoticons associated with joy (*e.g*. ":D") and sadness (*e.g*. ":("). These terms were used as query terms in data collection using the Twitter API between June and July of 2017 and a total of more than sixty million tweets were collected. In order to avoid the repetition of examples, all retweets, which are repostings of a tweet and that always starts with "RT", were removed. Tweets with URLs were also removed from the data set. For annotation of emotion intensity, the authors randomly selected 1,400 tweets for joy and 200 tweets for each of the three negative emotions. The 600 tweets associated to negative emotions were annotated regarding the intensity of all three negative emotions. To this, 800 tweets corresponding to only one emotion were added for each of the negative emotions. Tweets from the EmoInt data set (Mohammad & Bravo-Marquez, 2017) collected in 2016 using the same method were also added to the data set. The data set is composed of total of 10,983 tweets (3,091 tweets for anger, 3,627 tweets for fear, 3,011 tweets for joy and 2,905 tweets for sadness). The data set is divided into three subsets: a training set with 6,838 tweets, a development set with 886 tweets and a test set with 3,259 tweets. A subset of this data set is used on this work and more details about the subset creation and the strategy of data pre-processing will be given in section 4.1.

The SemEval-2019 Task 3 - EmoContext (Chatterjee *et al*., 2019) is a task with the goal of classifying the last turn of a three-turn dialogue as one of four classes: happy, sad, angry or others. The three emotion classes were chosen because of their popularity in conversations. The data set is then composed of dialogues with three turns - the utterance of the user, the response by an agent and the user's utterance in response to the agent. The data set is divided into three subsets: a training set with 30,160 dialogues and two test sets for two different phases of the task, one with 2,755 dialogues and the other with 5,509 dialogues. The training set was created in the following way: 1) 2 million dialogues were randomly sampled from dialogues collected over a period of one year, 2) human judges annotated a randomly selected sample to create a small set of approximately 300 dialogues per emotion class, 3)

the authors used dialogue embedding to identify potentially similar dialogues from the 2 million collected dialogues to the ones that were annotated, 4) from these potentially similar dialogues some were eliminated based on heuristics such as the presence of opposite emotions (*e.g.* ":)") in a dialogue classified as *sad,* and finally 5) the resulting set is shown to human judges that decide if the dialogues are correctly classified. The method used for data collection reduced the number of human judgements while creating a large data set.

EmotionLines (Chen *et al*., 2018) is a data set of dialogues collect from two different sources: transcripts from the TV show Friends (Friends data set) and conversations obtained from Facebook messenger logs (EmotionPush data set). It is important to note that since Facebook messenger conversations are private, named entities in the corpus were replaced using the Stanford Named Entity Recognizer (Manning *et al*, 2014). Utterances were labelled using Ekman's (1992) basic emotions as well as two other categories: neutral and non-neutral (utterances containing more than one emotion) by paid crowd workers. Each dialogue was assigned to 5 workers that classified each utterance of the dialogue considering the context by selecting an answer among 7 labels (neutral, joy sadness, fear, anger, surprise, and disgust) and the gold label was set as the answer with the highest number of votes. If an utterance is classified by the crowd workers into more than two categories, the label was set as *non-neutral*. Examples of these dialogues are shown in Table 3. 1.

Table 3. 1 Dialogues from Friends (top) and EmotionPush (bottom)[3]

Speaker	Utterance	Emotion
Phoebe	"Ohh, that's too bad."	sadness
Ross	"No, I-I'm saying I liked her."	non-neutral
Phoebe	"Yeah, y'know what, there are other fish in the sea."	neutral
Ross	"Pheebs, I think she's great. Okay? We're going out again."	non-neutral
Phoebe	"Okay, I hear you! Are you capable of talking about any thing else?"	disgust
1167038771	"why is that poster so bad"	disgust
1220662692	"what do you mean its great"	surprise
1167038771	"oh"	neutral
1167038771	"Can you Really Trust Anyone"	non-neutral
1220662692	"lol"	joy
1167038771	"Can You?"	neutral
1220662692	"I guess not"	sadness

3.2. Automatic Methods for Emotion Labelling

Although people use social media and microblogs with various motivations, they are often used to express opinions and emotions, both explicitly and implicitly. Therefore, social media can be very helpful in collecting data with information about the emotional state of their users. Java *et al*. (2007) analysed user intentions in Twitter messages (tweets) collected over a period of two months. The reported results indicate that Twitter users mostly use this social media platform to talk about their daily lives followed by conversations with other users. Since emotions are an intrinsic part of our daily lives and our interactions with other people it is expected that tweets with this kind of intentions are

[3] Adapted from EmotionLines (Chen et al., 2018).

emotionally rich. Therefore, social media such as Twitter, can be very useful when it comes to collecting data for the task of emotion recognition in text.

Hasan *et al.* (2014b) and Mohammad & Kiritchenko (2015) pointed out the challenges of manual labelling Twitter messages which include the large amount of time to execute this task and its underlying tedium, the difficulty of detecting the emotions of other people from written text due to semantic ambiguity, the different styles of writing of Twitter users that may include spelling mistakes and slang words, the difficulty of creating manually labelled data able to grasp all of the chosen emotions taking into consideration the various topics discussed by Twitter users and, finally, the inconsistency of human annotator's judgements that may vary regarding the same tweet due to the subjectivity involved in the task of assigning labels.

Automatic data extraction from microblogs such as Twitter can help prevent some of the challenges of manual data labelling thanks to features common to a large number of tweets posted by users such as emoticons, emojis or hashtags, that allow a greater perception of the user's emotional state to a certain degree. Hashtags, emoticons and emojis can help identifying the emotion of the user, especially when this information is not explicitly present in the rest of the message. Additionally, Twitter API allows to collect huge amount of data searching by query terms or hashtags, that can be filtered by several parameters such as the user id, the language in which the tweet is written or the geolocalization of the user.

Hasan *et al*. (2014b) studied the use of hashtags containing words associated to specific emotions for automatic labelling of tweets for the task of emotion recognition based on the Circumplex Model of Affect (Russell, 1980) by using four different categories corresponding to all the possible pairs of values of valence and arousal: Happy-Active (e.g. *#overjoyed, #superhappy)*, Happy-Inactive (e.g. *#calm, #relaxing)*, Unhappy-Active (e.g. *#anxious, #furious)* and Unhappy-Inactive (e.g. *#verysad, #depression)*. The authors compared automatic labelling with manual labelling by both psychology non-experts and experts. Their results show that there was a low level of agreement in classifying tweets according to emotional labelling by non-experts which would mean that manual annotation by non-experts is not sufficiently reliable. However, the level of agreement by experts is high and their classification of tweets in the emotion classes is more consistent. The automatic labels were compared with the manual labels of experts, and it was found that they matched in 87% of tweets showing that hashtags can be used for automatic labelling of tweets regarding emotions. For automatic labelling, they used data collected from Twitter API for a period of three weeks. They filtered these tweets by hashtags from a list of keywords corresponding to each of the four classes and removing all tweets found with hashtags belonging to more than one category or with contradictory emoticons (Hasan *et al.*, 2014a).

Mohammad & Kiritchenko (2015) built the Hashtag Emotion Corpus with tweets collected using Twitter API containing hashtags that corresponded to Ekman's (1992) six basic emotions. To evaluate the consistency of the Hashtag Emotion Corpus, the authors run two different automatic emotion classification experiments. The first experiment was an emotion hashtag prediction task, and the results indicate that the hashtags used to label tweets are sufficiently consistent and that this method can be employed to detect emotion hashtags in other tweets. The second experiment was built to discover if the created corpus could help improve emotion classification when using another domain and the results showed that it can help in this task to some extent and that automatic labels are consistent with labels of expert judges in general.

3.3. Emotion Recognition from Text

Emotion recognition from text is a task that brings together natural language processing and affective computing and can be done resorting to natural language processing techniques. It is important to refer the evolution of emotion recognition from text and the methods used as well as the current state of the art.

Most of the recent work is on contextual emotion recognition, where the emotional content of an utterance is detected within a discourse or conversation taking advantage of the context given by previous utterances.

3.3.1 Traditional Machine Learning

The work of Alm *et al.* (2005) had the goal of detecting and classifying emotions in children's fairy tales manually annotated for the six basic emotions established by Ekman (1992). However, given the ambivalence of the emotion *surprise*, the authors considered *positively surprised* and *negatively surprised* as distinct categories. For this purpose, the authors used SNoW (Sparse Network of Winnows) (Carlson *et al*., 1999), a supervised machine learning architecture. Although the data set was annotated for six emotions, recognition of specific emotions was not the goal of this work. Two different experiments were performed: classifying sentences as neutral or emotional and categorizing sentences according to the valence of the expressed emotion. Results for neutral sentences were better than for emotional sentences, with an F1 score of 0.70 versus an F1 score of 0.47 for the first experiment. In the second experiment, neutral sentences had also obtained a higher F1 score (0.69) than the negative valence sentences (0.32) and the positive valence sentences (0.13).

Strapparava & Mihalcea (2008) used five different knowledge-based and corpus-based methods for emotion recognition in text for the six basic emotions using the SemEval-2007 Task 14 - Affective Text (Strapparava & Mihalcea, 2007) data set composed by news headlines in English and annotated for these six emotions. These five methods were WordNet-Affect Presence, Latent Semantic Analysis (LSA) for single specific words denoting each emotion, LSA Emotion Synset using the synonyms from the WordNet synset of the emotion word, LSA with all emotion words found in WordNet Affect and Naïve Bayes classifier trained on blogs. The authors found that while the WordNet-Affect Presence method had the highest precision and the worst recall, the LSA method behaved in an opposite way having a low precision and the highest recall.

Amam & Szpakowicz (2007) also used a knowledge-based approach to detect if a sentence has emotional content, that is, if a sentence is emotionally neutral or not, regardless of the expressed emotion. They built a data set by extracting blog posts that contained emotion-related words for the six basic emotions and that was manually annotated for these six emotions as well as two other categories: *mixed emotion* and *no emotion*. They used emotion words retrieved from the General Inquirer (Stone *et al.*, 1966) and from WordNet-Affect (Strapparava & Valitutti, 2004) as lexical features and the classifiers Naïve Bayes and Support Vector Machines (SVM), achieving a maximum accuracy of 73.89% with SVM when using both lexical features.

Given the lack of large emotion-annotated data sets for training and the restrictions about the number of emotional-labels present in the available data sets, Agrawal & An (2012) used a new unsupervised context-based approach to detect the emotional content of sentences without fixing the number of emotional categories. NAVA words (nouns, adjectives, verbs, and adverbs) were extracted from each sentence and syntactic dependencies between them were captured for the purpose of obtaining

contextual information. Semantic relatedness between words and emotion concepts was calculated to allow the computation of emotion vectors for words by using Pointwise Mutual Information (PMI) which measures the similarity between two words according to the probability of co-occurrence in the different corpora used. The use of semantic relatedness avoids one of the weaknesses of calculating emotional vectors by directly matching words to an affect dictionary. The vectors were then fine-tuned considering the contextual information from the syntactic dependencies used (adjectival complement, adjectival modifier, and negation modifier). The proposed model was evaluated on three data sets: Children Fairy tales (Alm *et al.*, 2005), ISEAR (Scherer & Wallbott, 1994) with seven emotions (joy, fear, anger, sadness, disgust, shame, and guilt) and blog posts (Amam & Szpakowicz, 2007). The results obtained were comparable to some supervised methods and outperformed other unsupervised models.

Focusing on the Portuguese language for emotion recognition in text, Duarte *et al*. (2019) extracted data from Twitter using emojis as emotional labels to build a Portuguese data set annotated for the six basic emotions. They used Naïve Bayes and SVM classifiers to detect emotions as well as for the task of emoji prediction. Tweets were represented as a bag-of-words using unigrams, bigrams, and trigrams as features as well as the number of words, number of characters in the longest word, number of elongated words, average number of characters in the words and total count of negation words. They also extracted features from information about the emotion content of words obtained from affective lexicons such as the sentiment, dominance, valence, and arousal. For the task of emotion recognition, the best result was obtained using the Naïve Bayes classifier with a F1 score of 0.7. When analysing the results for each emotion, the model achieved a F1 score 0.86 for *happiness*. However, the F1 scores for other emotions are lower with a result of 0.527 for *sadness* and results lower than 0.153 for the remaining categories. The authors explain these results given that the collected data set is not balanced and the proportion of tweets of the *happiness* category was very high (68.4%).

3.3.2 Deep Learning

Much of the current work on emotion recognition in text uses deep learning techniques to detect and classify the emotional states expressed by people. In particular, recurrent neural networks and their variants such as LSTMs or **Gated Recurrent Neural Networks** (GRNNs) (Cho et al., 2014; Chung et al., 2015) have been providing very good results in this task.

Abdul-Mageed & Ungar (2017) used the extended version of Plutchik's model of emotions (1980), the three-dimensional circumplex, to classify a total of twenty-four discrete emotions. The authors collected data in English from Twitter using hashtags as emotional labels. They used GRNN to classify the tweets in emotions and obtained accuracies between 80% and 92%.

Gupta *et al.* (2017) considered four different classes of emotions (happy, sad, angry and others) for contextual emotion recognition in a data set composed by 3-turn conversation (Figure 3. 1) created from Twitter data. They proposed a model called **Sentiment and Semantic LSTM** (SS-LSTM), by feeding the input to an LSTM layer using a sentiment word embedding and another using a semantic word embedding. Their results show that the SS-LSTM model provides better results compared with other deep learning models and to traditional machine learning techniques, although they have found an unexpected low accuracy for the *happy* category (59.68%).

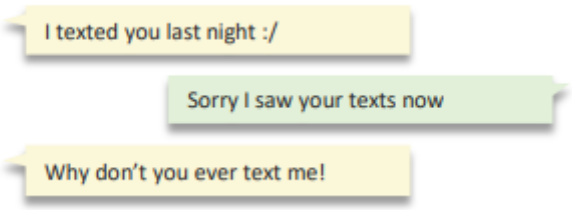

Figure 3. 1 Example of a 3-turn conversation (Gupta *et al.*, 2017)

Batbaatar *et al*. (2019) proposed a new model to address emotion recognition in text called Semantic-Emotion Neural Network. This model consists of two subnetworks: a BiLSTM to incorporate contextual and semantical information using word2vec (Mikolov *et al.*, 2013), GloVe (Pennington *et al*., 2014) and fastText (Bojanowski *et al*., 2017) as semantic word embeddings and a CNN to capture emotional features and how words related to one another regarding their emotional content fed by emotion-enriched word embeddings (EWE) (Agrawal *et al*., 2018) learned on product reviews. The authors used ten different emotion-annotated data sets including data sets composed by dialogues, fairy tales, tweets, or new headlines on their experiments. All ten data sets contain the emotional labels *sadness* and *joy*, and most of them also contain the labels *fear*, *anger*, *surprise*, and *disgust*. They compared their model to several traditional machine learning and deep learning models such as Naïve Bayes, Random Forest, SVM, Logistic Regression, CNN, LSTM, Gated Recurrent Unit and combinations of LSTM and CNN. In nine of the ten data sets used in the experiments, the proposed model outperforms all the compared models. Furthermore, in five of these nine data sets, the best results of this model are achieved using fastText embeddings.

Polignano *et al.* (2019) used a combination of Bidirectional LSTM and CNN mediated by a level of self-attention to detect and classify emotions in text for English. They compared the use of three pretrained word embeddings for this task. The word embeddings used were GloVe, word2vec and fastText. Their experiments were made using three different data sets – ISEAR (Scherer & Wallbott, 1994), SemEval 2018 task 1 (Mohammad et al., 2018) and SemEval 2019 task 3 (Chatterjee et al., 2019). The results that they obtained in their experiments showed that their model performed better than traditional machine learning approaches, with the exception for the "sadness" category in the ISEAR data set. Their results also showed that there was a statistically significant difference at p-value < 0.05 when considering the use of fastText Embedding for the two most recent data sets, with an improvement of 2% on average of the results comparing to the other two embeddings. This work was chosen to be replicated in this dissertation and it will be explained in more detail in section 4.2.

Artificial Intelligence is advancing rapidly and new techniques of natural language processing that can be applied to emotion recognition are emerging. In particular, the Transformer, that was introduced in Chapter 2, was the basis for later developed models such as the **Generative Pre-training model** (GPT) (Radford *et al*., 2018) and BERT (Devlin *et al*., 2019) that obtained new state-of-the-art results in several natural language processing tasks as previously mentioned.

In the EmotionX 2019 Challenge (Schmueli & Ku, 2019), a competitive shared task of detecting the emotional content of an utterance within a conversation, there were used two different data sets Friends and EmotionPush. According to the competition results, models based on BERT architectures achieved the best results (Huang *et al*., 2019; Luo & Wang, 2019; Yang *et al.*, 2019) for both data sets. Li *et al*. (2020) proposed a hierarchical transformer (**HiTransformer**) network using the pretrained language BERT as a lower-level transformer for modelling word-level input and an upper-level transformer to obtain the contextual relationship of utterances in a dialog. They also added speaker embeddings as a variation of the proposed model (**HiTransformer-s**). These embeddings model the

interaction between different speakers in a dialog. To perform the task, the authors used three different data sets: Friends, EmotionPush and EmoryNLP (Zahiri & Choi, 2018) for seven categories (neutral, sad, mad, scared, powerful, peaceful, and joyful). Both models outperformed state-of-the-art models and the HiTransformer-s obtained better results than the HiTransformer without speaker embeddings.

While traditional machine learning approaches have been present in the early works on emotion recognition in text, deep learning approaches in this field have only emerged in the last few years. However, they are becoming increasingly popular for detecting and classifying emotions in text thanks to their better performance.

For the Portuguese language, no work exploring deep learning models for emotion recognition in text was found to this date, making this dissertation the first to address this issue.

Chapter 4
Data Sets and Models

The main objective of the experimental part of this work is to recognize which emotion is present in a short text from a set of four emotions, for text in English and for text in Portuguese. In order to achieve this, we compare the performance of two different approaches: (i) a model based on BiLSTM, CNN and self-attention proposed by Polignano *et al*. (2019) which obtained encouraging results on three different emotion recognition data sets and (ii) BERT, a language representation model that, over the past few years, has come to support state-of-the-art results for many natural language processing tasks.

In this chapter, we describe the work done. We will start by presenting the data sets used for this work. In specific, we will explain our choice for the English data set and describe its composition and we will present the methodology of data collection for Portuguese as well as the resulting data set. Then we define our pre-processing strategy. We proceed to describe the methodology of our work by first explaining the architecture of the model of Polignano *et al*. (2019) and, finally, we present the BERT models that we will use and the hyperparameters used for fine-tuning these models.

4.1. Data Sets

This section describes the data set used for the English language and the adaptations made to the original data set and the creation of the Portuguese data set. The pre-processing strategy for both data sets will also be presented.

4.1.1 English Data Set

As previously mentioned in section 3.3.2, Polignano *et al*. (2019) experimented with three data sets in the paper presenting their emotion recognition model, namely ISEAR (Scherer & Wallbott, 1994), SemEval-2018 Task 1 Affect in Tweets data set (Mohammad *et al*., 2018; Mohammad & Kiritchenko, 2018) and SemEval-2019 Task 3 data set (Chatterjee *et al*., 2019).

One of the objectives of the present work is, on the one hand, the replications of the Polignano approach for English and, on the other hand, applying the same approach for Portuguese. Accordingly, the data sets used should be as similar as possible, apart from the language difference. The data set chosen for this work was the Affect in Tweets data set. This choice is due to the need of creating a

similar data set for the Portuguese language, since no publicly emotion annotated data set for Portuguese was found[4], and the Affect in Tweets data set is the only data set from those used by Polignano *et al.* that was created and annotated in an automated way that can be more easily adapted for the collection of Portuguese data, given the limitations of this work in terms of time and lack of annotators.
The Affect in Tweets data set was already described in section 3.1. However, following the approach of Polignano *et al.* (2019), only a subset of the data set is used on this work and the creation of this subset is described next.

For their experiments, Polignano *et al*. (2019) decided to only use tweets with an emotional intensity higher than 0.5 according to the annotation of Mohammad *et al*. (2018). Thus, both training and test sets were reduced. The training set used in their experiment has a total of 2761 (63.6%) tweets of which 933 (33.8%) are annotated for anger, 442 (16%) annotated for fear, 707 (25.6%) annotated for joy and 679 (24.6%) for sadness. Regarding the test set, it is composed of 389 (24.6%) of anger tweets, 229 (14.5%) fear tweets, 578 (36.6%) joy tweets, and 384 (24.3%) sadness tweets, with a total of 1580 (36.4%) tweets. Since a development set for fine-tuning BERT pre-trained model is required to evaluate the model during this phase, we used the development set of Mohammad *et al*. (2018) that consists of 766 tweets – 202 anger tweets, 210 fear tweets, 199 joy tweets and 155 sadness tweets. Examples of collected tweets are presented in Table 4. 1.

Table 4. 1 Examples of collected English tweets and their emotional labels.

Tweet	**Emotional Label**
I think about you when I'm drunk but can't stand the thought of you when I'm sober.	Sadness
Music is so empowering. it can literally bring people tears, smiles, laughter, and so many emotions	Joy
Nothing fuels my daily anger and hatred like a bus driver who stops at a yellow light	Anger
I am having anxiety right now because I don't know it's gonna happen	Fear

4.1.2 Portuguese Data Set

The Portuguese data set was collected using a methodology as similar as possible, given the scope of this work, to the one used for the English Affect in Tweets data set. For this reason, the emotions considered for the creation of the data set were the same as those used in the English data set: sadness, joy, anger, and fear. However, it was not possible to follow the exact same methodology due to time and cost constraints, specifically with regard to the annotation of emotion intensity.

Data collection was performed between January and July of 2020 and all data was collected in real time from Twitter using the Twitter Streaming API and Tweepy, a Python library to access the Twitter API. Hashtags composed of emotion related words or expressions were used as query terms as well as emotional labels. Tweets were filtered by language, which is one of the standard streaming API request parameters, and only tweets that were identified as being written in Portuguese were collected.[5]

[4] There are publicly annotated data sets regarding the sentiment of the text as, for example, the TweetSentBR data set (Brum & Nunes, 2018). These are appropriate for Sentiment Analysis, but not for Emotion Recognition.
The data set used by Duarte *et al*. (2019) was not publicly available. Furthermore, given the imbalance of the data set, with 68.4% of the tweets belonging to the same emotional class, it was not considered for this work.

[5] Despite cultural and linguistic differences among varieties of the Portuguese language, this is a language that unites people from different Portuguese-speaking countries. Additionally, an attempt was made to obtain the

A list of emotion related words in Portuguese (Table 4. 2) was created based on the items to evaluate specific emotions from Arriaga *et al*. (2010) and their primitive (e.g., "desgosto" (heartbreak) as primitive of "desgostoso" (heartbroken)) or derivative words (e.g., "felicidade" (happiness) as a derivative of "feliz" (happy)) in both grammatical genders, when they exist and are different (e.g., "animada" as the feminine of "animado" (cheerful). To obtain a larger dataset, synonyms obtained from Dicionário Priberam da Língua Portuguesa and Infopédia were also included. To avoid ambiguity and to obtain data with the best possible quality, lexically ambiguous words were not included (e.g., "irado", which means "irate", can also mean something good in an informal use of the word in Brazilian Portuguese).

Table 4. 2 List of emotion related words in Portuguese used to collect tweets.

Anger - Raiva	**Fear - Medo**	**Joy - Alegria**	**Sadness - Tristeza**	
encolerizado	amedrontada	alegrar	desconsolo	desespero
encolerizada	amedrontado	alegre	angústia	desgosto
enraivecido	amedrontar	alegremente	angustiada	desolação
enraivecida	apavorada	alegria	angustiado	desolada
cólera	apavorado	animada	angustiante	desolado
colérico	aterrorizador	animado	decepcionada	devastada
colérica	aterrorizadora	animar	decepcionado	devastado
enfurecer	assustada	diversão	dececionada	devastador
enfurecida	assustado	divertida	dececionado	devastadora
enfurecido	assustador	divertido	decepcionante	entristece
fúria	assustadora	engraçada	dececionante	entristecedor
furiosa	aterrorizada	engraçado	depressão	entristecedora
furioso	aterrorizado	excitação	deprimente	entristecer
indignação	aterrorizante	êxtase	deprimida	entristecida
indignada	aterrorizar	felicidade	deprimido	entristecido
indignado	empânico	feliz	deprimir	fúnebre
irritação	medo	muitofeliz	desanimada	inconsolável
irritada	medos	satisfação	desanimado	infeliz
irritado	pânico	satisfeita	desanimador	infelizmente
irritar	pavor	satisfeito	desanimadora	sofrer
odiando	petrificada	contente	desanimar	sofrimento
ódio	petrificado		desânimo	triste
raiva	receio		desapontada	tristeza
raivoso	receosa		desapontado	
raivosa	receoso		desapontamento	
revolta	susto		desapontar	
revoltante	sustos		desconsolada	
ultrajante			desconsolado	
zangada			desencorajada	
zangado			desencorajado	

greatest possible diversity of tweets written in Portuguese. Therefore, data collection was carried out for Portuguese, regardless of the language variety or location of the user.

Using the same procedure adopted when creating the English data set, retweets were removed from the data set. Tweets with hashtags belonging to more than one emotion category were also removed to avoid the presence of ambiguity regarding the tweet's emotional content.

A total of 11219 tweets were collected. Specifically, the data set is divided into 3170 tweets expressing joy, 3439 tweets for sadness, 1195 tweets for anger, 3415 tweets for fear. Table 4. 3 presents some examples of tweets that were collected and the corresponding emotional label.

Table 4. 3 Examples of collected Portuguese tweets and their emotional labels.

Tweet	**Emotional Label**
Na verdade eu não sei se quero descobrir toda a verdade...	Fear
amanhã recebo meu primeiro salário	Joy
Que pronunciamento ridículo do presidente da República!	Anger
Quando uma pessoa que você gosta só sabe te alfinetar.	Sadness

The data set was then divided into a training set with 6579 tweets, a development set with 1683 tweets and a test set with 2957 tweets, using the same split as the English data set.

Table 4. 4 Distribution of tweets per set and class for both languages.

	English				**Portuguese**			
	Training	**Test**	**Development**	**Total**	**Training**	**Test**	**Development**	**Total**
Anger	933	389	202	1524	713	297	185	1195
Fear	442	229	210	881	1954	925	536	3415
Joy	707	578	199	1484	1927	810	433	3170
Sadness	679	384	155	1218	1985	925	529	3439
Total	2761	1580	766	5107	6579	2957	1683	11219

4.1.3 Pre-Processing

The strategy adopted for data pre-processing was the one used by Polignano *et al.* (2019) and it was the same for the data sets in both languages, except when it was not possible to adapt some procedure for the Portuguese language.

The ekphrasis pre-processor library (Baziotis *et al.*, 2017) was also used to make annotation of mentions, URL, e-mail addresses, numbers, dates, and currency, which were replaced by specific tags (<mention>, <url>, <email>, <number>, <date> and <money>, respectively). In the English data set, it was possible to make some word spelling correction and hashtag segmentation into words when necessary, using the available corpus 'twitter' for word segmentation and spell correction. However, for the Portuguese data set, this option was not available. This may influence the results for Portuguese since data cleaning is very important in text classification tasks and it will be discussed in Chapter 5.

Since the objective of the present work is not only to recognize emotions expressed explicitly, but also implicitly, all hashtags used to collect data, *i.e.*, all hashtags composed of emotion related words or expressions used as query terms, were removed from both data sets. This was not a procedure adopted by the authors of the Affect in Tweets data set nor by Polignano et al. (2019) since hashtags were only removed from a small part of the tweets but is necessary given the goals of this work of recognizing

implicit emotions. Nonetheless, in Chapter 5 we present a comparison of the results of the reproduction of Polignano's approach using the original data set used by the authors and the results of the same model using the data set without the hashtags.

The next step in pre-processing is tokenization. We need to transform each short text into vectors of word embeddings to use Polignano's approach. Tokenization is the process of separating text into a list of tokens, which can be a very useful step in several natural language processing tasks. For this, we used the TweetTokenizer class of NLTK and tokenize the text of each tweet so that they can then be transformed into vectors of word embeddings.

Table 4. 5 Example of pre-processing steps.

Original Tweet
@jairbolsonaro Uma das maiores empresas de transporte do Brasil demitindo por causa do decreto de governadores socialista #triste https://t.co/49rKDC0qWz
Pre-processed Tweet
<mention> Uma das maiores empresas de transporte do Brasil demitindo por causa do decreto de governadores socialista <url>
Tokenized Tweet
['<mention>', 'Uma', 'das', 'maiores', 'empresas', 'de', 'transporte', 'do', 'Brasil', 'demitindo', 'por', 'causa', 'do', 'decreto', 'de', 'governadores', 'socialista', '<url>']

4.2. Polignano's Approach

The Polignano *et al.* (2019) model is based on two deep learning classification approaches described in Chapter 2, BiLSTM and CNN, mediated by a self-attention layer. The authors compared three different word embeddings to transform the input sentences into vectors of word embeddings, namely word2vec, GloVe and FastText. For the reproduction of their approach, we have selected only fastText (Bojanowski *et al.*, 2017), as Polignano *et al.* found this to be the best performing word embedding. Figure 4. 1 shows the architecture of this model. Next, we present the description of the architecture.

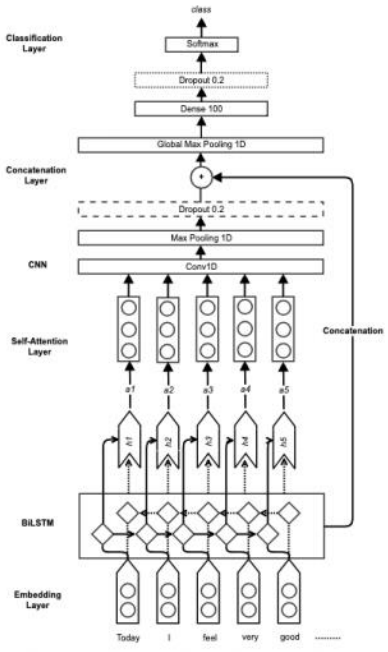

Figure 4. 1 BiLSTM, CNN and self-attention model's architecture (Polignano *et al*., 2019)

The first layer of their model is the word embedding layer. The fastText pretrained embeddings are a vocabulary with 2-million-word vectors with a dimensionality of 300 that take morphological information into consideration. Each word is represented by the sum of the n-grams of its characters, making the model able of capturing information regarding prefixes and suffixes. They were obtained by training on Common Crawl, a massive 600 billion token corpus crawled from the Web.

For Portuguese, we used the LX-DSemVectors version 2.2b (Rodrigues *et al*., 2016; Rodrigues & Branco, 2018) pretrained embeddings, with a vocabulary of more than 1-million-word vectors of dimensionality 500, obtained from a 2.2 billion token training corpus gathered from newspaper articles.

As is common in this kind of neural architectures, the input size needs to be fixed. Polignano *et al*. (2019) defined the maximum number of terms as being 80 words. For sentences with more than 80 terms only the first 80 terms will be considered and for sentences with less than 80 terms a padding operation with zeros will be applied until the desired dimension is reached. Words not found in the vector space were transformed into word embeddings through a vector selected from the entire collection.

The following layer is a BiLSTM network with 200 hidden units and a dropout value of 0.3. The authors chose this value to reduce the overfitting effect to the training data and to reduce the computational cost. The hyperbolic tangent function (tanh) was chosen as the activation function to speed up convergence when compared to the sigmoid function since it produces a larger gradient.
After the BiLSTM layer, a mechanism of self-attention was added that allows the model to consider relevant contextual relationships between neighbouring tokens by weighing the vectors of single words differently according with their similarity.

The CNN layer is added on top of the attention mechanism. More specifically, a 1D Convolutional Network with 400 filters and 5x5 kernel. The ReLu was chosen as activation function since it is faster to calculate than the hyperbolic tangent function. After this, to reduce dimensionality and computational load, the values were subsampled by adding a Max Pooling function, using a 2x2 kernel. A dropout function was then applied as a regularization technique.

The next layer concatenates the model obtained so far with the output of the BiLSTM. Another max pooling layer and a dense layer were then added to reduce dimensionality. Lastly, the probability distribution of each one of the four emotional classes will be estimated by a softmax activation function. The authors chose the categorical cross entropy function as loss function and Adam optimizer for 100 epochs.

To reproduce this approach, we used the code publicly available by the authors[6].

4.3. BERT

First, let us note that, as previously mentioned, BERT uses pre-trained WordPiece embeddings and therefore the pre-processing strategy is different than the one described in section 4.1.3, since the downloaded pre-trained model includes the relevant pre-processing strategy, and the tokenization step is not needed beforehand.

For English, we used the BERT-Base[7] uncased model, which has 12 encoder layers with 12 attention heads with a maximum of 768-length word embeddings and 110 million parameters and with the BooksCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia (2,500M words) as the pre-training corpus.

For Portuguese, we used BERTimbau Base[8] (Souza *et al.*, 2019; Souza *et al.*, 2020) which is a pretrained BERT model for the Portuguese language and that has, just like BERT-Base 12 encoder layers, 12 attention heads, embedding size of 768 and 110 million parameters. This model was trained on the BrWaC (Brazilian Web as Corpus) (Wagner Filho *et al.*, 2018) composed by 2.7 billion tokens. The results of this model on three NLP tasks (Sentence Textual Similarity, Recognizing Textual Entailment and Named Entity Recognition) were compared by the authors with the results of Multilingual BERT (mBERT) (Devlin *et al.*, 2019) and the performance of BERTimbau Base was better than mBERT for all three tasks.

For fine-tuning, the model is initialized with the same hyperparameters as in the pre-training phase, except for the learning rate, batch size and number of epochs. The pre-trained hyperparameters are fine-tuned to our task, that is, they will be updated during the training of our task which will allow the incorporation of specific characteristics of our task. During this phase, the model will be evaluated using the development set. With BERT, as with other Deep Learning models, is important to choose the best hyperparameters possible to use in training and in the architecture, since this choice can have a great impact on the results. For this work, we started by choosing as hyperparameters a learning rate of 2e-5, a batch size of 32 and 3 training epochs as in Devlin *et al.* (2019). From there, we vary each of the hyperparameters one at a time trying to find the combination of values that allow us to obtain the best performance. Our best results were obtained with a learning rate of 2e-4, a batch size of 4 and 5 training epochs.

[6] https://github.com/marcopoli/emofinder
[7] https://github.com/google-research/bert
[8] https://github.com/neuralmind-ai/portuguese-bert

Chapter 5
Results and Discussion

In this chapter, we will start by presenting the results obtained in the experiments that we carried out. Finally, we will discuss these results and attempt to provide a possible explanation for them.

5.1. Results

Table 5. 1 summarizes the results for the experiments on the Semeval-2018 data set for the original model of Polignano *et al.* (2019) as reported on their paper, for our replication of that work, and for the BERT model. For each model, the table shows precision, recall and F_1 score under each emotion class, as well as the overall micro-average F_1. The highest value in each column is in shown in bold.

Table 5. 1 Results for the original Semeval-2018 data set.

	Anger			**Fear**			**Joy**			**Sadness**			
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**μF_1**
Polignano *et al.* paper	0.77	**0.83**	**0.80**	**0.77**	0.70	**0.73**	0.97	**0.97**	**0.97**	**0.74**	**0.74**	**0.74**	**0.836**
Polignano *et al.* replication	**0.79**	0.80	0.79	0.76	0.71	**0.73**	0.97	**0.97**	**0.97**	0.73	**0.74**	**0.74**	0.8348
BERT	0.78	0.80	0.79	0.68	**0.75**	0.71	**0.98**	0.93	0.95	0.70	0.69	0.70	0.8139

Comparing the first two lines of the table, the results show that we have successfully replicated the model of Polignano *et al.* (2019) with the original data, having only a minor 0.001 difference in micro-average F_1 that can be attributed to the randomness in training, such as in network initialization and dropout. This replication was one of the goals of the current work and it ensures that the Polignano model used in the following experiments faithfully represents the original model.

The Polignano approach performs marginally better than BERT in all emotional categories in terms of micro-average F_1. **Joy** is the most easily recognized emotion by either approach, while **fear** and **sadness** are the hardest to correctly identify.

In the English data set, the emotion-word hashtags were only removed from a small part of the tweets. As explained in Chapter 4, the goal of this work is to recognize emotions expressed both explicitly and implicitly in texts. Since we are using hashtags to classify the text of the tweet regarding its emotional content, the presence of these hashtags will make the task of emotion recognition easier as the emotions will be expressed explicitly through the presence of emotion-indicative terms. Hence, we removed all the emotion-word hashtags used to classify the texts and obtained a new "cleaned" data set for English. The results for both approaches in the new data set are shown in Table 5. 2.

Table 5. 2 Results for the Semeval-2018 data set without emotion-word hashtags.

	Anger			**Fear**			**Joy**			**Sadness**			
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**μF1**
Polignano *et al.*	**0.70**	0.65	**0.67**	0.26	0.79	0.39	**0.92**	**0.86**	**0.89**	**0.65**	0.53	**0.59**	**0.7044**
BERT	0.63	**0.67**	0.65	**0.49**	0.63	**0.55**	**0.92**	0.80	0.86	0.60	**0.58**	**0.59**	0.6892

The results achieved in both approaches using the cleaned data set are considerably lower than when using the original dataset. This is expected since we are making the task more difficult by removing explicit emotional terms that would be used as obvious cues by the models. This drop in performance is particularly noticeable for the **fear** category, especially when using Polignano's approach (from a F1 score of 0.73 to 0.39). However, the difference in performance in the **fear** category in both approaches is not enough to translate into a better micro-averaged F1 score for BERT. BERT achieved a micro-averaged F1 score of 0.6892, just below the micro-averaged F1 score obtained using Polignano et al. (2019) model of 0.7044. **Joy** is still the best predicted emotion for both approaches, and it is the category with the lower drop in performance when compared with the experiments using the original data set.

In Table 5. 3, we show the results for Polignano's approach and BERT for the Portuguese data set.

Table 5. 3 Results for the Portuguese data set.

	Anger			**Fear**			**Joy**			**Sadness**			
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**μF1**
Polignano *et al.*	0.28	**0.41**	**0.33**	0.50	**0.61**	0.55	**0.82**	0.65	0.72	**0.55**	0.53	0.54	0.5972
BERT	**0.37**	0.08	0.13	**0.62**	0.57	**0.59**	0.78	**0.77**	**0.77**	0.51	**0.69**	**0.59**	**0.6124**

For the Portuguese data set, both approaches had worse results than for the English language. As described in Chapter 4, the methodology to collect the Portuguese data was as similar as possible to the one used for the English data set, but not the same. In particular, due to time and cost constraints it was

not possible to manually correct the automatic annotation. We expect that there might be more labelling inconsistencies in the Portuguese data set than in the English one, Thus, the difference in performance when compared to the experiments using the English data set is expected.

Unlike the two other experiments, BERT performs better on this data set with a micro-averaged F1 score of 0.6124 than the approach of Polignano *et al*. (2019) with a micro-averaged F1 score of 0.5972. However, it was more difficult for BERT to correctly predict **anger** emotions than it was for Polignano's approach. This is the category with worse results for both approaches and it can be explained due to the imbalance found in the data set since there is less data regarding the **anger** category compared to the other categories – between 35 and 38% of tweets from other categories. As in the other two data sets the **joy** category is the category with better results for both approaches.

5.2. Discussion

In this section, we will discuss the results in more detail. This will be done using theoretical foundations, analysing confusion matrices to see incorrect predictions and uneven distribution and by looking at some example text and the respective predicted output probabilities for each emotion class.

5.2.1 English Experiments

A clear result is the best performance using either approach for all the data sets in the **joy** category compared to the other three categories. This can be explained by the fact that **joy** is an emotion with positive polarity, as opposed to **anger**, **fear** and **sadness**. Either approach can more easily predict whether a text has a positive or a negative sentiment, for all the used data sets. It is more difficult to recognize negative emotions because, while there is a single positive emotion, the classifier needs to discriminate between the three negative emotions.

BERT's confusion matrix for the original English data set is presented in Table 5. 4 and it shows that it never predicts the category **joy** for texts that belong to **fear** category, and it almost never predicts **joy** for texts that belong to **anger** or **sadness** categories. We can also see that **anger** and **fear** texts are more incorrectly predicted as belonging to **sadness** than to other categories.

Table 5. 4 BERT's confusion matrix for the original Semeval-2018 data set.

	True							
Predicted	**Anger**		**Fear**		**Joy**		**Sadness**	
Anger	313	80.5%	13	5.7%	10	1.7%	66	17.2%
Fear	27	6.9%	171	74.7%	11	1.9%	43	11.2%
Joy	3	0.8%	0	0%	536	92.7%	9	2.3%
Sadness	46	11.8%	45	19.7%	21	3.6%	266	69.3%
Total	389	100%	229	100%	578	100%	384	100%

Table 5. 5 Some tweets correctly classified by BERT with the original Semeval-2018 data set.

Text	true emotion	predicted emotion	anger	fear	joy	sadness
some people just need to learn how to smile 😁 and laugh 😂 😂 live life	joy	joy	9.390e-05	5.876e-05	**0.999798**	4.891e-05
Tuesday night and the wine is coming out . Just got home from work if that explains it . 🤦 🏻 ♀ 🤦 🏻 ♀ 🤦 🏻 ♀ 🤦 irritated	anger	anger	**0.999472**	0.000133	1.649e-05	0.000377
we mourn the death of our hopes today james	sadness	sadness	0.000142	0.000950	0.000316	**0.998590**
Living downtown for the first time while tiff is happening . Schedule is out today! excitement the children act Stanley Tucci Emma Thompson	joy	joy	9.850e-05	6.944e-05	**0.999785**	4.640e-05
<mention> My dad ordered my tickets for the show in Hamburg, his name is now printed on the tickets, is the same surname enough? panic	fear	fear	0.000166	**0.9991043**	0.000140	0.000588

In Table 5. 5, we present some examples of correctly classified texts by BERT when using the original English data set. Under the columns for each emotion are the probabilities assigned by the model to that emotion, with the highest in bold. Note that, in all cases, the probability is overwhelmingly assigned to the correct class. However, it is more important to look at the misclassified texts since they can provide some information and give us some clues as to why the model is making incorrect predictions. The misclassified texts by BERT of the original data set are shown in Table 5. 6.

Table 5. 6 Some tweets misclassified by BERT with the original Semeval-2018 data set.

Text	true emotion	predicted emotion	anger	fear	joy	sadness
<mention> Man people wont mourn the <number> chaps beheaded by terrorists . But they will mourn a corrupt killer politician	anger	sadness	0.000237	0.000772	0.000192	**0.998796**
I am so done with all the despair I am having	fear	sadness	0.000200	0.013114	0.000247	**0.986437**
<mention> is coming for DePaul's welcome week and I will be in Cali 😒 😒 😒 pissed	sadness	anger	**0.999549**	9.427e-05	2.141e-05	0.000334
<mention> I think I will tomorrow . I ain't ready for all those feels though . 😔	sadness	fear	0.000690	**0.9747093**	0.000218	0.024382
My main concern are the children and his wife thats if she is still stuck with him and if she is, then she's strong af . Jeez 😐	fear	sadness	0.002440	0.0180094	0.00263	**0.976913**
<mention> I'd never leave the couch again! excited and weary ♡	joy	sadness	0.002144	0.0207837	0.000259	**0.976812**
I dreamt that my dog, Snoopy, came back to life . Man I miss that dog 😐	sadness	joy	2.338e-05	0.0004443	**0.998989**	0.000542

The first misclassified example shown in Table 5. 6 includes the word “mourn” two times. This word has an emotional charge that can be associated with the emotion sadness. In fact, the third example in Table 5. 5, which also includes the word “mourn” is correctly classified as sadness. It is possible to say the same regarding the presence of specific words in other examples in Table 5. 6 such as “despair” in example 2 (also associated with sadness) or “pissed” in example 3 (associated with anger). As with

the previous examples, for each tweet the probability is overwhelmingly assigned to a single class, which indicates that the model was strongly mislead by the presence of certain words.

In Table 5. 7, we present BERT's confusion matrix for the cleaned English data set (without the emotion-word hashtags). It is possible to see that while **joy** is rarely predicted for texts in other categories, it is wrongly predicted more often than when using the original data set. As with the original data set, **anger** and **fear** texts are also more incorrectly predicted as belonging to **sadness** than to the other categories.

Table 5. 7 BERT's confusion matrix for the Semeval-2018 data set without hashtags.

	True							
Predicted	**Anger**		**Fear**		**Joy**		**Sadness**	
Anger	259	66,6%	30	13,1%	37	6,4%	85	17,2%
Fear	52	13,4%	145	63,3%	38	6,6%	61	11,2%
Joy	14	3,6%	8	3,5%	463	80,1%	16	2,3%
Sadness	64	16,5%	46	20,1%	40	6,9%	222	69,3%
Total	389	100%	229	100%	578	100%	384	100%

Table 5. 8 presents some examples of text that were misclassified by BERT for the cleaned data set. However, to better assess the impact of hashtag removal, these are examples that BERT classified correctly in the original data set. As such, we can say that for these cases the misclassification was caused by the removal of the hashtags. The hashtags that were removed from each text are presented in the last column named **hashtag removed**.

Table 5. 8 Some tweets misclassified by BERT with Semeval-2018 data set without hashtags that were correctly classified with the original data set.

Text	true emotion	predicted emotion	anger	fear	joy	sadness	hashtag removed
<mention> My dad ordered my tickets for the show in Hamburg, his name is now printed on the tickets, is the same surname enough?	fear	sadness	0.203850	0.1922438	0.233639	**0.370265**	#panic
<mention> My Big Mac was cold!	anger	sadness	0.002769	0.1712892	0.001187	**0.824754**	#fuming
For many years I have despised olives, my thoughts on them have now changed.	joy	anger	**0.917404**	0.0013237	0.000753	0.080519	#delightful
I am a terrible person . I have a viscerally negative reaction to Big Bird's new voice on Sesame Street . nope	fear	sadness	0.002064	0.014032	0.000436	**0.983467**	#shudder
There's no better way to relax than putting on a face mask, having a bath while drinking tea and listening to John Mayer	joy	sadness	0.004335	0.024992	0.022727	**0.947945**	#bliss
I made a joke & my boyfriend laughed but when I looked over he was actually laughing at a video on his phone <number> broken heart	sadness	joy	0.005448	0.0008071	**0.99319**	0.000554	#sad #hurting
Surprisingly even though I am not a Trump supporter I simply do not get the vibe that all we have heard is going to be his undoing.	sadness	fear	0.026257	**0.8305618**	0.001777	0.141402	#pessimism

When looking at some of the examples in Table 5. 8, we can find some emotional words that may steer the model in the wrong direction such as "despised" (associated with **anger**) in the third example or "laughing" and "laughed" (associated with **joy**) in sixth example. Except for the first

example, where there is a relatively flat distribution of probabilities over the possible emotions due to a lack of emotional words in the text, there is always a high probability assigned to the chosen emotion. Removing words with a strong emotional content (the words present in the hashtags) was enough to make BERT misclassify these examples.

However, there are also cases (although very few) in which the hashtag can be misleading and lead to misclassifications as in the examples below (Table 5. 9). In these examples, the classification was correct after removing the hashtag.

Table 5. 9 Some tweets misclassified by BERT with the original Semeval-2018 data set that were correctly classified when the hashtags were removed.

Text	**Hashtag**	**Predicted emotion – original data set**	**Predicted emotion with the hashtag removed**	**True emotion**
Having a bad day <mention> nightmare	#nightmare	fear	sadness	sadness
Beyond disappointed with <mention> clothing line and service Lulu never asks questions about returns Not getting money back horrible	#horrible	fear	sadness	sadness

5.2.2 Portuguese Experiments

BERT's confusion matrix for the Portuguese data set is shown in Table 5. 10. It is possible to see that almost all the texts that belong to the **anger** category are being classified as **sadness** (58.2%) and as **fear** (23.9%). This may have happened due to the low number of examples belonging to the **anger** category when compared to the other three categories. Texts belonging to the **fear** category are also often misclassified as belonging to the **sadness** category (34.9%). In fact, **sadness** is the category that is most often wrongly predicted.

Table 5. 10 BERT's confusion matrix for the Portuguese data set.

	True							
Predicted	**Anger**		**Fear**		**Joy**		**Sadness**	
Anger	23	7,7%	13	1,4%	11	1,4%	16	1,7%
Fear	71	23,9%	525	56,8%	64	7,9%	187	20,2%
Joy	30	10,1%	64	6,9%	622	76,8%	81	8,8%
Sadness	173	58,2%	323	34,9%	113	14,0%	641	69,3%
Total	297	100%	925	100%	810	100%	925	100%

Table 5. 11 Correctly classified texts by BERT with the Portuguese data set.

Text	true emotion	predicted emotion	anger	fear	joy	sadness
<mention> É sério que você quer demitir o Mandetta pq ele não concorda com vc Um bom líder lapida seus liderados e não os dispensa	sadness	sadness	0.282637	0.0673757	0.007735	**0.642253**
você foi uma falta de respeito com os meus sonhos que bom que acordei	sadness	sadness	0.055048	0.0183268	0.025230	**0.901395**
ainda por cima estão querendo convocar um juiz para eedir um mandato para verificar nossos equipamentos vocês cometem os erros de vocês lançando update com problemas que causam defeito em periféricos e agora ELE É CULPADO	anger	anger	**0.524749**	0.1231865	0.009808	0.342256
Dá uma olhada nessa reclamação contra Unidas Aluguel de Carros no <mention> <url>	anger	anger	**0.63674**	0.1415495	0.070742	0.150968
acabei de tonalizar meu cabelo sozinha pela primeira vez	fear	fear	0.009955	**0.948341**	0.013546	0.028158
<mention> Já comprei paracetamol e um xarope pra tosse Não tem muito o que fazer No máximo vou procurar fazer o exame	fear	fear	0.096779	**0.5424345**	0.011766	0.349021
A live hoje foi um verdadeiro transbordo para essa geração Conta ai o que mais aprendeu Vamos gostar de te ouvir <url>	joy	joy	0.028093	0.0441094	**0.651066**	0.276732
Comemorando aniversário do nosso futuro Vereador Danilo Leandro com seus familiares <url>	joy	joy	0.015001	0.0113998	**0.930647**	0.042953

In Table 5. 11, we present some examples that were correctly classified by BERT for the Portuguese data set. Note that, in many cases, the probability of the assigned emotion is not overwhelmingly higher than that of the other emotions, as it happened for the data set with hashtags. As mentioned before, although it is interesting to look at the correctly classified examples, we can learn more from the misclassified examples. Those are the ones presented in Table 5. 12.

Table 5. 12 Misclassified texts by BERT with the Portuguese data set.

Text	true emotion	predicted emotion	anger	fear	joy	sadness
Fui segurar minha gata pra aplicar injeção e acabei chorando	sadness	fear	0.024129	**0.9164367**	0.014892	0.044541
Já são quase meio dia e eu ainda não recebi a mensagem da sogra para ir almoçar na casa dela	sadness	fear	0.100308	**0.4748542**	0.036507	0.38833
Mas a maioria das histórias q tô lendo no fim a mulher termina descobrindo da pior forma possivel que o cara era um traidor e ela não sabia	fear	sadness	0.201706	0.0677291	0.009820	**0.720745**
Crueldade Cachorros são resgatados debilitados sem comida e água em MS <url>	sadness	anger	**0.533579**	0.0205806	0.035593	0.410247
Como se não bastasse ontem eu esqueci o carregador hoje eu esqueci o fone Mas pelo menos cheguei cedo e antes da chuva	anger	sadness	0.122601	0.10531	0.032572	**0.739517**
Cara me da raiva ver como a sociedade lida com certos assuntos	anger	fear	0.051742	**0.7901411**	0.030079	0.128038
MEU DEUS EU TERMINEI A CÔMODA	joy	fear	0.061218	**0.5876598**	0.115000	0.236122
Existem pensamentos projetos e sonhos que é melhor só o nosso coração saber	fear	joy	0.001552	0.0040252	**0.985628**	0.008795
Muito ódio muita briga e até tiro Assim são as manifestações próBolsonaro <url>	anger	joy	0.191126	0.0292648	**0.471154**	0.308455
Eu pensando em alguma possibilidade de ser suficiente pra alguém algum dia <url>	sadness	joy	0.007867	0.0393762	**0.80648**	0.146277
Hj a tarde eu achei 3 barra de chocolate q eu esqueci q tinha comprado	joy	sadness	0.068717	0.035781	0.018555	**0.876946**

Some of the misclassified examples presented in Table 5. 12 may present some degree of difficulty to classify even by human judges. In the first example, we could say that “aplicar a injeção” (giving the injection) may invoke the emotion fear. As for the third example, it does not seem obvious that it expresses fear rather than another negative emotion such as anger or sadness. We can also find expressions that are associated with other emotions in some of the texts. Specifically, in the seventh example, the text includes the interjection “MEU DEUS” (“MY GOD”) that usually expresses fear or surprise.

Finally, there are other factors to consider that may influence the results. One of the most important and that has been previously mentioned is the quality and quantity of data. Twitter is a very useful resource to collect written data, however Twitter messages are usually too short, contain abbreviations and spelling mistakes very often and are sometimes presented out of context. This way, it is difficult even to humans to always understand and correctly classify their emotional content. For the Portuguese data set, there were no human annotators due to time and cost constraints.

Chapter 6
Conclusion

This chapter concludes this dissertation by providing a summary of the main conclusions and contributions of this work and by discussing the work that can be done in the future.

6.1. Summary

This dissertation addresses the subject of emotion recognition in text and its main objective was to contribute to the state of the art of emotion recognition in Portuguese. The achievement of this goal required several previous steps.

The first necessary step was to carry out a literature review to deepen knowledge about emotion models and computational techniques used to classify emotions expressed in texts. This allowed us to provide a theoretical framework and present the current state of the art in emotion recognition and the available data resources for this task, as well as the means to obtain these resources.

Motivated by the literature review, we chose two deep learning models that have given state-of-the-art results for English, namely the emotion recognition model of Polignano *et al.* (2019), based on BiLSTM, CNN and self-attention, and the more recent BERT language model, based on Transformer.

Thus, the second step was to replicate one of the emotion recognition experiments of Polignano *et al.* (2019) for English to ensure that we were able to obtain similar results. This experiment was run on the Semeval-2018 data set and consisted of classifying tweets in one of four emotions (anger, fear, joy, and sadness). Our replication differs only by 0.001 points in micro-average F1 from the original results, which we consider to be a successful replication.

The third step was then to implement BERT using the same data set for the same task, for which we tried different values for the hyperparameters in search of the best possible results.

We consider that the data set used in the previous experiments is not ideally suited for the stated goals of this work. This data set was built by automatically gathering and labelling tweets on the basis of emotional cues, that is hashtags composed by emotional words or expressions that were present in the tweets. As our intention is not only the classification of emotions expressed in texts based simply on the presence of certain emotional words and expressions, but also the classification of implicitly expressed emotions, we created a new data set by removing the emotional cues used in data collection from the examples that still contained them, and we repeated the experiments using both the approach of Polignano *et al.* (2019) and BERT in this new data set, thus taking the fourth and fifth steps towards

achieving our ultimate goal. The performance of both models dropped, which was expected as we were making the task more difficult. Polignano's approach outperformed BERT in both data sets with a difference of about 0.02 in the micro-averaged F1 score.

Regarding the Portuguese language, we were faced with the lack of Portuguese data sets for this task that suited the purpose of this dissertation and that allowed the comparison of results with those obtained in English. So, our sixth step was the collection and automatic emotional labelling of texts in Portuguese from Twitter, adopting a similar methodology to the one used in the creation of the English data set. We collected a total of 11219 tweets between January and July 2020.

Finally, it was possible to carry out the experiments with the Portuguese data using both approaches used for English. The results differ from the results obtained for English and, in Portuguese, BERT outperformed Polignano's approach, with results of 0.6124 and 0.5972, respectively. These results were lower than the results obtained for English, which was also expected due to the challenges in adopting the procedures of data labelling.

Table 6. 1 Summary of micro-average F1 results obtained for all data sets and models.

	μF_1		
	Semeval-2018 - original	**Semeval-2018 without emotion-word hashtags**	**Portuguese**
Polignano *et al.*	**0.8348**	**0.7044**	0.5972
BERT	0.8139	0.6892	**0.6124**

Our data set is available at https://hdl.handle.net/21.11129/0000-000E-75BA-D so that the experiments can be replicable.

6.2. Future Work

The quality and quantity of data are important factors that influence the performance of Machine Learning models. Regarding the quality of the data, it is important to undertake manual revision of the Portuguese data set, putting it in line with what was done for the English data set, and setting a gold standard Portuguese data set for this task. Regarding the quantity of data, more tweets will continue to be gathered, which will also enable further studies addressing, for instance, how emotions in tweets vary over time.

We used BERT-Base and BERTimbau-Base models for the English and Portuguese experiments, respectively, both with 12 encoder layers, 12 attention heads and a total of 110 million parameters. However, there are bigger BERT models, BERT-Large and BERTimbau-Large. These models have 24 encoder layers with 16 attention heads and a total of 340 million parameters and require more computational resources than the ones used in this work. Without restrictions regarding computational resources, the use of these larger BERT models would also be a good possibility for further work.

Comparing the performance of computational models in emotion recognition with human performance would also be interesting. For this, the experiment of emotion recognition by humans using the same textual data could be carried out.

Transfer Learning is a Machine Learning technique that has been gaining traction in the past few years. Roughly, it consists in tackling a target task, not by training a system from scratch, but by training a system that has been previously trained on a different, related task, hoping that some

knowledge from the related task can improve performance in the target task. This technique has achieved promising results for a range of NLP tasks and, in fact, the use of BERT in this work is a form of transfer learning where emotion recognition is the target task and language modelling is the "related" task. Further research is needed to assess which tasks better help tackling the task of emotion recognition and, conversely, whether training a model on an emotion recognition task can successfully transfer what is learned to other tasks.

References

Abdul-Mageed, M., & Ungar, L. (2017). EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 718-728. doi:10.18653/v1/P17-1067

Agrawal, A., & An, A. (2012). Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. *WI-IAT '12: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, 01*, pp. 346-353. doi:10.1109/WI-IAT.2012.170

Agrawal, A., An, A., & Papagelis, M. (2018). Learning Emotion-enriched Word Representations. *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 950-961.

Alm, C. O. (2008). Affect Dataset.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 579-586. doi:10.3115/1220575.1220648

Amam, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In V. Matoušek, & P. Mautner, *Text, Speech and Dialogue. TSD 2007.* (Vol. 4629, pp. 196-205). Berlin: Springer. doi:10.1007/978-3-540-74628-7_27

Arnold, M. B. (1960). *Emotion and personality.* Columbia University Press.

Arriaga, P., Franco, A., & Campos, P. (2010). Indução de emoções através de excertos musicais. *Laboratório de Psicologia, 8*(1), pp. 3-20. doi:10.14417/lp.645

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*.

Bard, P. (1928). A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology, 84*(3), pp. 490-516.

Batbaatar, E., Li, M., & Ryu, K. (2019). Semantic-Emotion Neural Network for Emotion Recognition From Text. *IEEE Access, 7*, pp. 111866-111878. doi:10.1109/ACCESS.2019.2934529

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 747–754. doi:10.18653/v1/S17-2126

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, pp. 135-146. doi:10.1162/tacl_a_00051

Brum, H., & Nunes, M. V. (2018). Building a Sentiment Corpus of Tweets in Brazilian Portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 4167 - 4172.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psycholoogy, 39*(1-4), pp. 106-124. doi:10.2307/1422695

Carlson, A., Cumby, C. M., Rosen, J. L., & Roth, D. (1999). SNoW User Guide.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39-48. doi:10.18653/v1/S19-2005

Chen, S., Hsu, C., Kuo, C., Huang, T., & Ku, L. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1597-1601.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (n.d.). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated Feedback Recurrent Neural Networks. *Proceedings of the 32nd International Conference on Machine Learning, PMLR 37*, pp. 2067-2075.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals.* London, UK: John Murray.

Devlin, J., Chang, M.-W., Kenton, L., & Kristina, T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Vol 1*, pp. 4171-4186.

Duarte, L., Macedo, L., & Oliveira, H. G. (2019). Exploring emojis for emotion recognition in portuguese text. *EPIA Conference on Artificial Intelligence*, pp. 719-730. doi:10.1007/978-3-030-30244-3_59

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), pp. 169-200. doi:10.1080/02699939208411068

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), pp. 384-392. doi:10.1037/0003-066X.48.4.384

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14*(2), pp. 179-211. doi:10.4324/9781315784779-11

Gers, F., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation, 12*(10), pp. 2451-2471. doi:10.1162/089976600300015015

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5-6), pp. 602-610. doi:10.1016/j.neunet.2005.06.042

Gupta, U., Chatterjee, A., Srikanth, R., & Agrawal, P. (2017). A Sentiment-and-Semantics-Based Approach for Emotion. *arXiv abs/1707.06996*.

Harris, Z. (1954). Distributional Structure. *Word, 10*(23), pp. 146-162.

Hasan, M., Agu, E., & Rundensteiner, E. A. (2014). Using Hashtags as Labels for Supervised Learning of Emotions in Twitter Messages.

Hasan, M., Rundensteiner, E. A., & Agu, E. (2014). EMOTEX: Detecting Emotions in Twitter Messages.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation, 9*(8), pp. 1735-1780. doi:10.1162/neco.1997.9.8.1735

Huang, Y. H., Lee, S., Ma, M., Chen, Y. H., Yu, Y., & Chen, Y. (2019). EmotionX-IDEA: Emotion BERT–an affectional model.

James, W. (1884). What is an Emotion? *Mind, 9*(34), pp. 188-205.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56-65. doi:10.1145/1348549.1348556

Kleinginna, P. R., & Kleinginna, A. M. (n.d.). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*, pp. 345-379. doi:10.1007/BF00992553

Lang, P. J. (2010). Emotion and Motivation: Toward Consensus Definitions and a Common Research Purpose. *Emotion Review, 2*(3), pp. 229-233. doi:https://doi.org/10.1177/1754073910361984

Lange, C. G. (1885). *Om Sindsbevaegelser: Et Psyko-Fysiologisk Studie.* Copenhagen, Denmark: J. Lund.

Lazarus, R. S. (1966). *Psychological Stress and Coping Process.* McGraw-Hill. doi:10.2307/1420698

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist, 46*(8), pp. 819-834. doi:10.1037/0003-066X.46.8.819

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., & Howard, R. E. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), pp. 541-551. doi:10.1162/neco.1989.1.4.541

Li, Q., Wu, C., Wang, Z., & Zheng, K. (2020). Hierarchical Transformer Network for Utterance-Level Emotion Recognition. *Applied Sciences, 10*(13). doi:10.3390/app10134447

Luo, L., & Wang, Y. (2019). EmotionX-HSU: Adopting pre-trained BERT for emotion.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60. doi:10.3115/v1/P14-5010

Metri, P., Ghorpade, J., & Butalia, A. (2011). Facial Emotion Recognition Using Context Based Multimodal Approach. *International Journal of Interactive Multimedia and Artificial Intelligence, 1*(4), pp. 12-15. doi:10.9781/ijimai.2011.142

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2*, pp. 3111-3119.

Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 Shared Task on Emotion Intensity. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 34-49. doi:10.18653/v1/W17-5205

Mohammad, S. M., & Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence, 31*(2), pp. 301-326. doi:10.1111/coin.12024

Mohammad, S. M., & Kiritchenko, S. (2018). Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in Tweets. *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. doi:10.18653/v1/S18-1001

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. doi:10.3115/v1/D14-1162

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik, & H. Kellerman, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3-33). New York, NY, USA: Academic.

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4), pp. 344-350.

Polignano, M., Basile, P., de Gemmis, M., & Semeraro, G. (2019). A Comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention. *UMAP'19 Adjunct: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 63-68. doi:10.1145/3314183.3324983

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pretraining.

Rodrigues, J., & Branco, A. (2018). Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2403-2409.

Rodrigues, J., Branco, A., Neale, S., & Silva, J. (2016). LX-DSemVectors: Distributional Semantics Models for the Portuguese Language. *Lecture Notes in Artificial Intelligence, 9727*, pp. 259-270. doi:10.1007/978-3-319-41552-9_27

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology, 39*(6), pp. 1161-1178. doi:10.1037/h0077714

Saif, M. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. doi:10.18653/v1/S18-1001

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), pp. 379-399. doi:10.1037/h0046234

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), pp. 310-328. doi:10.1037/0022-3514.66.2.310

Shmueli, B., & Ku, L. (2019). SocialNLP EmotionX 2019 Challenge Overview: Predicting Emotions in Spoken Dialogues and Chats. *arXiv, abs/1909.07734*.

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. *9th Brazilian Conference on Intelligent Systems, BRACIS 2020, 12319*, pp. 20 - 23.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis.* M.I.T. Press. doi:10.2307/1161774

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 task 14: affective text. *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 70-74. doi:10.5555/1621474.1621487

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1556-1560. doi:10.1145/1363686.1364052

Strapparava, C., & Valitutti, A. (2004). WordNet Affect: an Affective Extension of WordNet. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pp. 1083-1086.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 4339-4344.

Werbos, P. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE, 78*(10), pp. 1550-1560. doi:10.1109/5.58337

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Dean, J. (2016). Google's neural machine translation system: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Yang, K., Lee, D., Lee, T. W., & Lim, H. (2019). EmotionX-KU: BERT-Max based contextual emotion classifier.

Zahiri, S. M., & Choi, J. D. (2018). Emotion detection on TV show transcripts with sequence-based convolutional neural. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, p. 44.52.