# Sub-language Sentiment Analysis in WhatsApp Domain with Deep Learning Approaches

Leonardo P. de Morais, Anderson da Silva Soares, Vinicius da C. M. Borges, Nadia Félix Felipe da Silva and Fabíola S. F. Pereira.

***Resumo*—Abordagens de análise de sentimento têm oferecido uma ferramenta útil para sistemas de apoio à decisão em vários campos, incluindo política, gerenciamento de rede, marketing e saúde. Devido ao crescente impacto das redes sociais online nesses campos e ao fato de serem ricas fontes de informação, as atuais técnicas de análise de sentimento nesse cenário evoluíram com sucesso. O WhatsApp é uma plataforma de rede social que permite aos usuários interagir com laços próximos de uma maneira particular para comunicar informações mais significativas, genuínas, tangíveis e pessoais ao destinatário, tais como um sentimento. Portanto, o domínio do WhatsApp pode ser definido como um sub-linguagem da primeira linguagem. No entanto, apenas alguns estudos se concentraram na análise de sentimentos do WhatsApp. Esses trabalhos geralmente empregam técnicas de léxico de sentimento desatualizadas e não avaliam as técnicas mais modernas baseadas em aprendizado profundo. Este estudo tem como objetivo avaliar essas técnicas de análise de sentimentos baseadas em redes neurais profundas e aprendizado por transferência, considerando as características intrínsecas da sub-linguagem no domínio do WhatsApp. As técnicas de BERT1 e ALBERT1 (abordagens de transferência de aprendizado) alcançam o melhor desempenho em precisão e F1 (88% em média para métricas e classificadores) de forma semelhante a outros domínios (Twitter). Embora DCNN e LSTM com static embbedings geralmente alcancem bom desempenho quando são pré-treinados em um corpus maior de outros domínios, essas abordagens atingem o pior desempenho para o domínio do WhatsApp. Além disso, o ELMo fornece uma boa troca entre a precisão e a complexidade do tempo de treinamento, principalmente quando levamos em consideração o pequeno tamanho da nossa base de treinamento do WhatsApp. Assim, pode-se inferir que as características específicas da sub-linguagem do WhatsApp tem impacto no desempenho de alguns classificadores tradicionais de análise de sentimento.***

***Palavras-chave*—Análise de Sentimentos, WhatsApp, Aprendizado de Máquina, Aprendizado Profundo.**

***Abstract*—Sentiment analysis approaches have offered a useful tool for decision support systems in various fields, including politics, network management, marketing, and healthcare. Owing to the increasing impact of online social networks on these fields and the fact that they are rich sources of information, the current sentiment analysis techniques in this scenario have evolved successfully. WhatsApp is a social network platform that enables users interact with close ties in a particular manner to communicate more meaningful, genuine, tangible, and personal information to the recipient, such as a sentiment. Hence, WhatsApp domain can be defined as a sub-language of the first language. However, only few studies have focused on WhatsApp sentiment analysis. These works usually employ outdated sentiment lexicon techniques and do not assess the most modern techniques based on deep learning. This study aims to evaluate this techniques for sentiment analysis based on deep neural networks and transfer learning, considering the intrinsic features of sub-language in WhatsApp domain. BERT1 and ALBERT1 (transfer learning approaches) achieve the best performance in accuracy and F1 (88% on average for both metrics and classifiers) similarly to other domains (Twitter). Although DCNN and LSTM with static embeddings usually achieve good performance when they are pre-trained on a larger corpus of other domains, these approaches reach the worst performance for WhatsApp domain. Furthermore, ELMo provides a good trade-off between the accuracy and training time complexity, mainly when taking into account the small size of our corpus training of WhatsApp. Hence, it can be inferred that the specific characteristics of the WhatsApp sub-language has an impact on the performance of some traditional SA classifiers.***

***Index Terms*—Sentiment Analysis, WhatsApp, Machine Learning, Deep Learning.**

## I. Introduction

Sentiment Analysis (SA) helps in classifying the emotions of textual data, usually, as polarities; that is, positive, negative, or neutral [1]. SA is a powerful tool for decision-making in applications with high availability of texts, such as stock market prediction, financial markets, network optimization, trading strategies, labor market intelligence, e-commerce, e-learning, and social media marketing [2]. Although there are several ways to perform SA, Natural Language Processing (NLP) is the most common owing to the data type. In NLP, according to [3], [4], a sub-language is a variety of a language that forms subsets of the general language, typically exhibiting particular types of lexical, semantic, and other restrictions and performance. Generally, a sub-language is formed by a restricted domain in technical or nontechnical applications. However, there are few studies about the applicability of NLP algorithms in sub-languages, and most of them are in the technical domain [5], [6].

Social network platforms, such as blogs, forums, and microblogs are rich sources of sub-language data for various applications, including SA [7]–[14]. Online social networks such as Twitter and Facebook were initially designed for

Leonardo, Anderson, Vinicius, Nadia are with the Institute of Informatics (INF), Federal University of Goiás (UFG), Goiânia. Brazil, e-mail: leopereiramorais@gmail.com, andersonsoares@ufg.br, vcmborges@ufg.br, nadia.felix@ufg.br

Fabíola is with Federal University of Uberlândia, Uberlândia, Brazil, e-mail: fabiola.pereira@ufu.br

Manuscript received February 08, 2023; revised May 08, 2023; accepted June 14, 2023.

promotional purposes to express the personal, commercial, and political views of users, their groups, or brands. For example, people posted about their vacations, any new product consumed, social events, and consumer reviews. Messenger applications, such as WhatsApp, introduced a more direct relationship among users, with a specific language, i.e., a sub-language of the social network dialect.

WhatsApp is the most popular global mobile messenger application, with approximately 1.6 billion monthly active users [15], [16]. WhatsApp is an instant messaging communication application through which users can share texts, images, and videos with other users and groups with a limited number of people (approximately 280 people). Furthermore, WhatsApp is one of the three most popular social networks worldwide [15], [16] and it can help in extrapolating for other communication platforms such as WeChat and Facebook messenger.

This type of platform communication is advantageous for SA, mainly because it is used to primarily interact with close ties comparison to more public platforms such as Facebook, Twitter, and Instagram, which revolve more around communication with weak tie [17]. When people send a specific message to an individual or group of closest people in a more private manner, they generally want to communicate something meaningful and personal to the recipient; for instance, a sentiment. Thus, the analysis of texts on WhatsApp can help in classifying sentiments that translate the current and most accurate sentiment of a user.

On the other side, this type of data analysis involves difficulties and challenges; for example, because the messages are private, it is not easy to find an available public database to analyze sentiments in this type of social media. Furthermore, in this work, we emphasize that the average sentence size was 23 characters for WhatsApp messages, which is smaller than the average sentence length of tweets (i.e., 53 characters after changing character the limit [18]). Other challenges and characteristics, such as audience, corpus size, social ties (described in detail in Section II) show that the WhatsApp domain has a distinct sub-language.

Various methods and models [19], [20] have helped in evolving the state-of-the-art in Twitter SA, such as competitions, benchmarks, and scientific studies that developed classification systems based on lexicon and recurrent neural networks (Convolutional Neural Networks (CNNs) and Long Short Term term Memory (LSTM) being the most common choice), transformer networks, and fine-tuning models built upon Bidirectional Encoder Representations from Transformers (BERT) [21]–[23]. Recent studies have demonstrated significant performance improvements on several NLP tasks using the representations extracted from the pre-trained models on the large unannotated Twitter corpora through transfer learning. BERT [21] and variations of it (A Lite BERT (ALBERT) [24], Robustly optimized BERT approach (RoBERTa) [25], Structural BERT (StructBERT) [26], and others) are widely used Deep Neural Language Models (DNLMs) with the best performance. However, there is no information about whether these transfer learning models and traditional Deep Neural Networks (DNNs) also yield high accuracy for SA in the context of the WhatsApp sub-language.

DNLMs that achieved state-of-the-art results on downstream NLP tasks have recently been trained for numerous languages and sub-languages, including Portuguese [27], [28]. More recently, this research domain was integrated with transfer learning [29]–[33]. In order to understand the impact of transfer learning on the SA models, we can revisit the classic supervised machine learning paradigm. In SA with supervised learning, isolated learning is considered from a predictive model using a single dataset. This approach requires several training examples that exhibit the best performance for well-defined objective tasks. Transfer learning refers to a set of methods that extend this approach, leveraging data from domains, or additional tasks to train a model with better generalization properties.

Considering the different sub-languages employed in social media platforms, our hypothesis is that the most relevant and recent models along with DNLMs and DNNs for SA exhibit different performances in the WhatsApp sub-language in comparison to their performances on Twitter and/or consumer reviews/blogs/forums. Therefore, we aims to evaluate the performance of WhatsApp SA for Portuguese. Our main contributions can be summarized as follows: (i) A WhatsApp corpus in the Portuguese language was annotated with authorization of users; (ii) a comprehensive performance evaluation of the established sentiment classification methods and features was presented; iii) We show through experiments that there are a domain and context adaptation (a form of transfer learning), in which the knowledge accumulated through the language models (that were previously trained using formal texts) contribute to WhatsApp SA (which are informal texts); and iv) to the best of our knowledge, there are few studies that have focused on DNLMs and DNNs approaches with evaluations in the Portuguese language for SA, especially with WhatsApp.

In order to achieve these goals, the remainder of this paper is organized as follows. Section II compares the sub-languages used in distinct social network platform. Section III presents the main related work in the WhatsApp domain. Section IV details the WhatsApp sentiment corpus. Section V describes the DNLMs used. Sections VI and VII report our experiments and the obtained results. The conclusions and future work are presented in Section VIII.

## II. Sub-languages of Social Network

The differences between the domain of the WhatsApp sub-language and the sub-languages for other social network platforms' domains are described below.

**i) The frequency of misspellings and slangs in WhatsApp messages**. These characteristics occurs more frequently than in other domains because users typically use a more informal and direct language since it is used for relationships with people who are closer and therefore more intimate [17], [34], which can result in shorter sentences (fewer characters). Furthermore, in this sub-language, the development of a specific vocabulary and distinct culture by users, while being limited in terms of the length (e.g., number of characters), may convey rich meaning. In contrast, Facebook and Twitter are more temporal and public because they are designed to depict the messages, groups, brands of a single user, including information about their vacations, parties, new products, and social events; their posts, whether personal or commercial, are meant to foster an image or promote something to a group of people (indirect). Thus, the informality degree is balanced. On the other hands, other media, such as blogs, forums, and consumer reviewers consider more technical subjects that demand more formal and indirect language.

**ii) There are very few public corpora annotated with a sentiment label for WhatsApp**. In addition, the few existing corpora or datasets are significantly small (with small number of messages). In contrast, there are several public datasets for Twitter, forums, blogs, and consumer reviews [35].

**iii) Multiple Topics**. There are diverse topics in WhatsApp because people usually communicate with their personal contacts in a private way; thus, they talk about different subjects which change in time. In this case, the use of domain-specific sentiment lexicons can result in low accuracy. Furthermore, this approach is static and, over time, would require expansions to increase its generalization.

**iv) Class imbalance**. Class of Positive messages are more common than classes of negative and neutral messages in WhatsApp datasets [36]–[38]. This condition is different from other SA domains (e.g., product reviews), which tend to be predominantly negative. Nevertheless, Facebook and Twitter tend to have positive and negative classes more balanced than Whatsapp because they establish communication between people with strong and weak ties; thus, the most used class will depend more on the topic and social ties.

Based on these characteristics, we can come to the conclusion that WhatsApp has distinct aspects. Thus, this social online network is a sub-language. Furthermore, these differences can influence the performance of the state-of-the-art of SA classifiers. Hence, an evaluation performance of main algorithms and methods of SA in a WhatsApp corpus is necessary to assess whether this sub-language impacts on the performance of these classifiers.

## III. Related Work

Sentiment Analysis has been studied on many media, including reviews [39], forum [40], discussions [41], and blogs [42]. In this paper, we are interested in analyzing texts from WhatsApp.

Data from WhatsApp can be analyzed from different perspectives. Various studies are interested in understanding the communication patterns in group chats [43] and user behavior [44] using exclusively general statistics and metadata from conversations, without exploring the message content. Meanwhile, WhatsApp texts have been used in applied analysis, without necessarily involving SA tasks or general NLP approaches. For instance, the work in [45] analyzed the engagement of pregnant women and mothers by observing the WhatsApp messages exchanged between them. It should be noted that only manual and human analyses have been conducted without using any NLP method. The authors in [46] employed Amazon Comprehend[1] to perform automated topic modeling on WhatsApp data to identify the most prominent topics applied to large-scale engagement contexts.

In another work, the authors argued that recognizing sarcastic statements will help in improving the SA of data collected from WhatsApp conversations in [47]. However, no machine learning approach was adopted, only manual and human classification. These scenarios illustrate that while the applicability of WhatsApp data has been increasingly explored, there is a lack of implementation of automatic learning approaches, such as deep learning, in state-of-the-art NLP over texts from WhatsApp.

The R programming and the R studio library are employed in [36] to analyze the sentiments of text messages from people in a WhatsApp group (all organization members). This work aimed to determine the positivity level of the group. The author analyzed eight emotions in the WhatsApp group database and found that positive sentiment had the highest frequency, with trust being the second-highest frequent sentiment. The author employed the *get_nrc_sentiment* method based on the lexicon by which it can extract sentiments and eight emotions according to their frequency.

The work in [48] analyzed the 1) textual features: a) language usage, b) main topics, and c) sentiment of message content between two sets of messages in Brazilian WhatsApp public groups about politics; 2) misinformation as well as 3) the propagation dynamics of both sets of messages. They extracted these types of features from the texts using a version of the Linguistic Inquiry and Word Count (LIWC), which is a psycholinguistic lexicon system based on a Portuguese dictionary that categorizes words into psychologically meaningful groups. They also used a Portuguese version of the SentiStrength method [49] to measure the polarity of each instance of content, which is a combination of supervised learning techniques with a set of rules that impact the strength of the opinion contained in the message. A substantial volume of negative messages was observed in both sets of messages due to the election/politics subject. Moreover, the authors reported that there were more positive messages than neutral ones (also in both groups), which evidences the polarized nature of the data [48].

The authors in [37] focused on classifying sentiments on small datasets, such as sentiment strength-tweet, health care reforms, US airlines, and WhatsApp datasets (public WhatsApp chat). A hybrid deep method was proposed

[1] https://aws.amazon.com/comprehend/

that combined the CNNs and Bidirectional-LSTM (Bi-LSTM) DNNs. CNN-BiLSTM employed four conv1d and two max pool layers along with Bi-LSTM. The proposed method was compared with several machine learning-based methods, such as logistic regression, Naive Bayes (NB), support vector machine, decision tree, K-nearest neighbor, Ada-boost, classifier extra tree, gradient boosting machine, eXtreme Gradient Boosting (XGB) classifier, and random forest. The experimental results demonstrated that the proposed method outperformed the existing methods over WhatsApp datasets, obtaining an accuracy and precision of 80% and 75%, respectively. The hybrid deep method method is computationally slower than the several classical machine learning-based methods that were mentioned before. In addition, it was not compared with the isolated CNN and Bi-LSTM models.

The user sentiments were analyzed in WhatsApp public groups based on the ensemble learning method in [31]. The authors employed five classifiers to analyze the sentiments and used voting classifiers to aggregate the results from the five classifiers and obtain the final result. The five classifiers include the Gaussian NB classifier, multinomial NB classifier, Bernoulli NB classifier, logistic regression classifier, and linear Support Vector Classification (SVC) classifier. These five classifiers use different methods to perform classification.

Authors in [38] proposed the use of WhatsApp data to leverage user profiles by combining an SA classifier based on a CNN from ParallelDots API [2] with emoji information. They intended to utilize the intensive usage of emojis in WhatsApp conversations to improve SA considering six emotion labels (happy, sad, angry, excited, sarcasm, and fear).

Most of the WhatsApp datasets are private and have a significantly small number of messages in comparison to the datasets of other social networks such as Twitter. Most of the related works on WhatsApp SA employed the lexicon approach. However, the lexicon approach has drawbacks; for example, WhatsApp users post messages on various topics, and in this case, the use of domain-specific sentiment lexicons yields low accuracy. Furthermore, this approach is static and it must be expanded over time to increase its generalization. Deep learning-based classification has been examined by the NLP research community, and for big datasets, greater performance than classic machine learning algorithms such as regressors and decision trees has been obtained for various tasks, including tweet SA [50]. However, this research scenario is not presented or documented in WhatsApp texts considering the small size of text datasets and sub-language aspects in WhatsApp.

## IV. WhatsApp Sentiment Corpora

We built our corpora from existing and closed groups considering that WhatsApp SA is not widespread in the literature [36], [51], and a *corpus* with textual WhatsApp data in Portuguese labeled with sentiments is a non-existent resource.

It is worth mentioning that WhatsApp messages are not stored on servers; end-to-end encryption is employed during the transmission of these messages to ensure that WhatsApp and the third parties cannot read them. Therefore, we collected the WhatsApp messages from people who volunteered to participate in the research. This work was supported by a research project submitted, evaluated, accelerated, and approved by the university's ethics committee to legalize volunteer involvement. The volunteers participants of this study authorized the WhatsApp messages collection by signing the consent term. This term defines the rights and duties of volunteers and researchers, such as the clause that establishes that the messages must not be released publicly in order to preserve the volunteer' privacy.

The built corpus had 20,000 messages in the Portuguese language, collected from WhatsApp private groups. These groups comprised people with diverse age and social class, living in different states of Brazil. Most of the people know each other. We devised two groups of 37 women and 32 men, with ages varying between 18 and 65 years.

Six thousand, two hundred and ninety-eight (6,298) messages from the WhatsApp groups were randomly selected[3] from the WhatsApp groups for the *Training Corpus.* Three volunteers, who did not participate in the collection, were chosen for the task of WhatsApp annotation and they manually labeled the sentences in positive, negative, or neutral classes. The same three volunteers annotated all messages, and the messages were labeled by the majority consensus (when two people agreed). Table I lists some examples. Just one annotator has some experience in SA annotation and in the linguistics research fields. All annotators are native Portuguese speakers, and undergraduated. In order to minimize the disagreement between the annotators and the subjectivity of the task, periodical individual training meetings were held to understand and discuss the concept of WhatsApp Sentiment.

We used the Test Corpus to evaluate the results of the sentiment classification models that consider the positive, negative, and neutral classes. The Test Corpus is different from the corpora used for training. Similar to the training corpus, the WhatsApp messages were taken from private/closed groups, adopting the same manual labeling methodology by three people with the tiebreaker criteria. This corpus contained 579 messages. Table II presents Statistics for each class and Fleiss' Kappa for our corpora.

We computed the Fleiss' Kappa score [52] to assess the agreement between the three raters. We interpreted this score using the normalization proposed by Sim et al. [53]. Table II lists the Fleiss's Kappa for each corpus. In the Training Corpus, the Fleiss' Kappa score was 0.91, implying an "almost perfect agreement" [53]. For the Test Corpus, a Fleiss' Kappa score of 0.52 was obtained, and ac-

[2] https://apis.paralleldot.com/text_docs/index.html

[3] We considered a smaller sample than the original dataset due to annotation costs.

Table I: Examples of WhatsApp Messages

| **Messages with Majority Agreement (All annotators said the same)** | Class |
|---|---|
| *Dá pra ver a irmã dele chorando tadinha.* (You can see his sister crying.) | Negative |
| *Palavra abençoada.* (Blessed word.) | Positive |
| **Messages with Low Agreement** | Class |
| *Tenho uma entrevista hj a tarde, mas comecei a estudar.* (I have an interview today, but I started studying) | Positive/Negative |
| *Pensei que você já tinha visto desculpas.* (I thought you had already seen excuses). | Positive/Negative |

cording to [53], we can interpret this value as a "moderate agreement". There is a significant difference of Kappa score between training and test corpus. It is important to point out that the text from WhatsApp is very subjective (due to peculiar aspects of this sub-language) and this increases the complexity of the annotation process. In addition, we had much greater number of messages in the training corpus than in testing and this added more context to the messages and increased agreement between annotators. Hence, this difference between these two corpus does not affect the reliability of the results.

Table II: Statistics for each class and Fleiss'Kappa for individual categories

| **Class** | **#** | **Kappa** |
|---|---|---|
| *Training Corpus* | | |
| Positive | 4,146 | 0.91 |
| Negative | 1,252 | |
| Neutral | 900 | |
| *Test Corpus* | | |
| Positive | 289 | 0.52 |
| Negative | 184 | |
| Neutral | 106 | |

It is worth mentioning that the corpora were collected over a period of ten months. The messages from the first seven months were used for training, and those from the next three months were used for testing. Although they include texts from the same groups of people, the training and testing corpora had individual characteristics. It can be observed from Figures 1 and 2 that the texts in the Training Corpus are larger than those in the Testing Corpus and provide more context for the analysis. This can be responsible for the different annotation kappa values because smaller sentences have no context or a smaller context and cause a more significant disagreement in the annotation. Furthermore, the average sentence size was 23 characters for the Training Corpus, which is smaller than the average sentence length of tweets (i.e., 53 characters after changing character the limit [18]). This confirms what was asserted in section II, WhatsApp sub-language has shorter sentences (few characters). Figures 1 and 2 also show that the most number of sentences have approximately 0 to 50 characters for the Training Corpus and Testing Corpus. Hence, the WhatsApp sentences are mostly short[4].

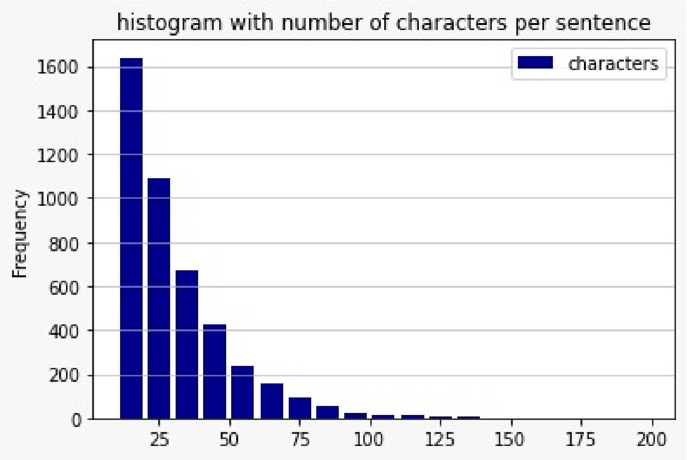


Fig. 1: Histogram: Frequency × Sentences size (characters) – Training Corpus

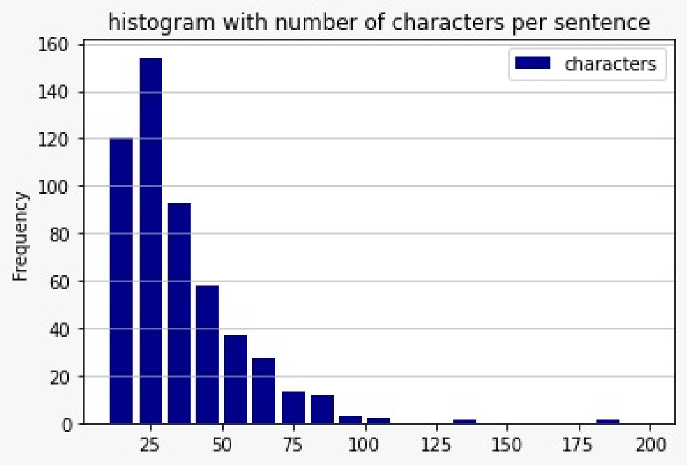


Fig. 2: Histogram: Frequency × Sentences size (characters) — Testing Corpus

In WhatsApp messages, users cannot ensure that responsive postings are placed in a directly adjacent position. In [55], the authors analyzed this in German WhatsApp dialogs and compared them to other languages. They argued that users often communicate by formatting their messages in such a way that it makes a relevant response. In Portuguese, we analyzed this aspect by counting the occurrences of the user sentences that were separated only by a line break. We observed that in 10% of the corpus messages of training, the users did not present their messages in a unique text format (without line breaks).

[4] We take into consideration character count which is directly related to sentence size [54].

When we combined these successive sentences in a single paragraph from a user (Figure 3), we observed that this had a low impact on the size of sentences from the corpus. Therefore, we considered the line break of each sentence as the delimiter of sentences for our SA evaluation. As it was expected, Figure 3 also shows that the sentence sizes increases when successive sentences are joined because the number of sentences between 0 and 50 characters reduces while the amount of those comprising between 51 and 200 characters increases.

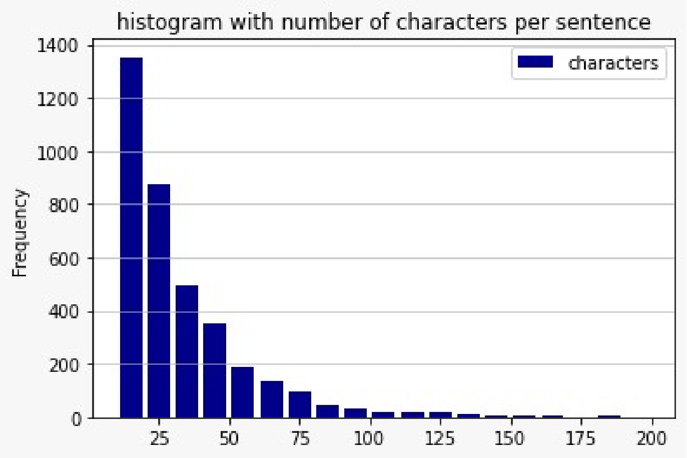


Fig. 3: Grouped Sentences' sizes from Training Corpus

## V. Deep Neural Language Models

In this study, we evaluate the performance of DNLMs for WhatsApp SA against classical word embeddings which uses DNNs. Two word representation methods were evaluated via SA in Portuguese, i.e., contextualized and static word representations. Static word representations were provided as input to the DNNs in a network training stage [56]; in this study, two DNNs were used, namely: Dynamic Convolutional Neural Network (DCNN) and LSTM. These networks were selected for SA in WhatsApp owing to their good performance in the literature for other domains [57], [58]. Each approach is explained in the following sections.

### A. Static Word Representations

Word representations are numerical vectors that can represent words or concepts in a low-dimensional continuous space, reducing the inherent sparsity of traditional vector space representations [59]. These vectors, also known as embeddings, can capture useful syntactic and semantic information, such as the regularities in natural language. They are based on the distributional hypothesis, which establishes that the meaning of a word is given by its context of occurrence [60]. The ability of static word embeddings to capture knowledge has been extended in several tasks, such as machine translation [61], SA [62], word sense disambiguation [63] and language understanding [64]

Although very useful in many applications, the static word embeddings, like those generated by Word2Vec [65], Global Vectors (GloVe) [66], Wang2Vec [67] and FastText [68] have an important limitation: each word is associated with only one vector representation, ignoring the fact that polysemous words can assume multiple meanings. This limitation is called the meaning conflation deficiency, which is a mixture of the possible meanings of a single word [69]. Because they create a single representation for each word, a notable problem with static word embeddings is that all meanings of a polysemous word must share a single vector. Qiu et al. [70] emphasized that these models are context-free and fail to capture higher-level concepts in context, such as polysemous disambiguation, syntactic structures, semantic roles, and anaphora. The embedding for a word is always the same, regardless of its context. Another issue is the out-of-vocabulary problem. In order to tackle this problem, character-level word representations or sub-word representations are used in many NLP tasks, such as CharCNN [71] aand FastText. In this study, we apply the static word embeddings Word2Vec, Wang2Vec, FastText, and Glove. Although we could have chosen other static word embeddings, the ones adopted here have been widely used in practice [27], [72], therefore they are suitable as a proof of concept.

**a) Word2Vec:** [65] is a method used in NLP for generating word embeddings. It has two different training strategies: (i) Continuous Bag-of-Words (CBOW), in which the model is given a sequence of words without the middle one; it attempts to predict this omitted word; (ii) Skip-Gram, in which the model is given a word and attempts to predict its neighboring words. In both cases, the model consists of only a single weight matrix (apart from the word embeddings), which results in fast log-linear training that can capture the semantic information.

**b) GloVe**: The Global Vectors (GloVe) method was proposed by [66]. This method consists of a co-occurrence matrix, M, constructed by observing the context words. Each element $M_{ij}$ in the matrix represents the probability of word $i$ being close to word $j$. In the matrix $M$, the rows (or vectors) are randomly generated and trained by obeying the equation:

$$P(w_i, w_j) = log(M_{ij}) = w_i w_j + b_i + b_j, \qquad (1)$$

where $w_i$ and $w_j$ are word vectors, and $b_i$ and $b_j$ are biases.

**c) Wang2Vec:** is a modification of Word2Vec that considers the lack of word order in the original architecture. Two simple modifications were proposed in Wang2Vec such that the embeddings can better capture the syntactic behavior of words [67]. In the continuous window architecture, the input is the concatenation of the context word embeddings in the order of their occurrence. In structured skip-gram, a different set of parameters is used to predict each context word depending on its position relative to the target word.

**d) FastText:** is a method [68] in which the embeddings are associated with character n-grams and words are represented as the summation of these representations. In this method, a word representation is produced by adding

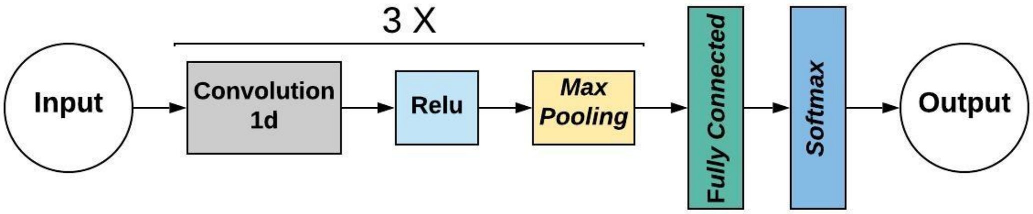


Fig. 4: DCNN network architecture.

character n-gram vectors with the vectors of surrounding words. Therefore, this method attempts to capture the morphological information to create word embeddings.

### B. Contextualized Word Representations

The limitations of static word embeddings led to the creation of context-sensitive word representations. Embeddings from Language Models (ELMo) [29], BERT [30], and Generative Pre-trained Transformer 2 (GPT-2) [73] are examples of DNLMs that are fine-tuned to create models for various downstream NLP tasks. Because GPT-2 is not yet available for the Portuguese language, we performed our experiments solely on ELMo, a multilingual version of BERT, Portuguese BERT, and Portuguese ALBERT. The internal representations of words for these language models are called contextualized word representations because they are a function of the entire input sentence. In this study, the sentence embeddings were built by adding these representations. The success of this approach suggests that these representations capture highly transferable and task-agnostic properties of natural languages [33]. Next, we describe the main characteristics of the two representation models used in this work.

**a) ELMo:** ELMo is a hybrid deep neural model language. ELMo is composed of LSTM and CNN. In other words, ELMo employs a two-layer bidirectional LSTM language model, built over two context independent character CNN layers context-sensitive features from a left-to-right and a right-to-left language model. The contextual representation of each token is the concatenation of the independently trained left-to-right and right-to-left representations.

**b) BERT:** [30] is a language model that improves previous language representation models by replacing left-to-right language model pre-training with two unsupervised pre-training tasks, thereby enabling a deep bidirectional architecture. BERT is one of the key innovations in the recent development of contextualized representation learning. Even though the word embedding layer is trained from large-scale corpora, it is insufficient to train a wide variety of neural architectures that encode contextual representations only from the limited supervised data on end tasks. Unlike ELMo, which intends to provide additional features for a particular architecture that bears the human understanding of the end task, BERT adopts a fine-tuning approach that requires almost no specific architecture for each end task. This is a desirable property because an intelligent agent should minimize the use of prior human knowledge in the model design. Instead, it should learn such knowledge from the data [74].

BERT implements the transfer learning concept, whereby the language representations are pre-trained on large corpora and fine-tuned in a variety of downstream tasks, such as sentiment analysis. The pre-training occurs with two tasks, the Masked Language Model and Next Sentence Prediction via a large cross-domain corpus. Unlike previous biLMs that are limited to a combination of two unidirectional language models (i.e., left-to-right and right-to-left), BERT uses a masked language model to predict words that are randomly masked or replaced. BERT is the first fine-tuning-based representation model that achieved state-of-the-art results for various NLP tasks, demonstrating the enormous potential of the fine-tuning method [75].

## VI. Experimental Setup

Experiments were performed to evaluate whether the DNLMs and DNNs approaches yield different overall precision achievements for WhatsApp sentiment classification in comparison with other social platforms. Details regarding the parameters of each method are given below.

### A. DCNN and LSTM

A DCNN and LSTM were used as DNNs because they achieved good results for Twitter, reviews products and blogs [57], [58] when solving the sentiment classification problem in textual data. The DCNN architecture is sized with three convolution layers; each convolution layer contains a one-dimensional convolution, Rectified Linear Unit (ReLU), and a Max Pooling, followed by a Fully Connected layer, and ending with Softmax, as shown in Figure 4 based on [57], [76]. ReLU is non-linear, which means that the errors from the training process are copied; thus, several layers of neurons are activated by the ReLU function. The Softmax activation function assists in generating nonlinearities, giving the network more abrangency (softmax is used for multivariate classification tasks) in learning the functionalities [57], [76].

The LSTM network is the second DNN evaluated as a classifier of sentiment for WhatsApp texts. The architecture was defined in a traditional format, since the parameter for the number of hidden layers was dimensioned with 100 layers [57], [76]. Table III lists the parameters used in the architecture of DCNN and LSTM.

Table III: Parameters of the DCNN and LSTM.

| **Parameters** | **DCNN** | **LSTM** |
|---|---|---|
| Batch Size | 15 | 15 |
| Training Times | 100 | 15 |
| Hidden Layers | - | 100 |
| Convolution Layers | 3 | - |
| ReLu Layers | 3 | - |
| Max Pooling Layers | 3 | - |
| Fully Connected Layers | 1 | - |
| Softmax Layers | 1 | - |

## *B. Static Embeddings*

Static word embeddings are used for word representations, one of the most popular forms currently used for representing word vectors; they can be used to capture the semantic and syntactic similarities as well as the relationship with other words. This approach is provided as an entry for DNNs in one of the stages in training the networks [56]. Regarding word embeddings, we experimented with four different pre-trained word embeddings [27]: Word2Vec [65], FastText [77], Wang2vec [78] and Glove [79], each containing 300 dimensions, and Word2Vec with the CBOW strategy. Details regarding these models can be found in [80].

## *C. Dynamic Embeddings*

We considered ELMo, Portuguese BERT, BERT- multilingual, and Portuguese ALBERT with two different training sets for dynamic embedding. The parameters used have been described below.

**A) ELMo:** We linearly concatenated all three ELMo layers without learning any task-specific weights to obtain a representation for each word. During our experiments, we considered the ELMo language model that was exclusively trained for the Portuguese language[5]. We trained the model with four epochs: maximum sequence length of 128, train batch size of 32, and learning rate of 2e-5 (which are frequently used in literature). We used the AllenNLP framework [81].

**B) BERT:** The BERT-based model contains an encoder with 12 transformer blocks, 12 self-attention heads, and a hidden size of 768. BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The sequence has one or two segments in which the first token of the sequence is always [CLS], which contains the special classification embedding, and another special token [SEP] is used for separating segments. For text classification tasks, BERT takes the final hidden state $h$ of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of the BERT to predict the probability of label $c$:

$$p(c|h) = softmax(Wh) \tag{2}$$

where W is the task-specific parameter matrix. We used the Hugging Face tool [82] for the experiments, considering four training epochs, with a maximum sequence length of 128, training batch size of 32, learning rate of 2e-5, and the following pretrained versions of BERT were utilized during our experiments:

**a) Portuguese BERT:** Portuguese BERT [28] was the first BERT model pre-trained exclusively for the Portuguese language on a large dataset to be publicly released. It was pre-trained on the Brazilian Web as Corpus dataset (BrWaC) [83], a dataset crawled and filtered from more than 60 million Brazilian Portuguese pages. BrWaC is composed by 3.53 million documents in general themes, 2.68 billion tokens and is now openly available, both for download and for navigation in a NoSketch Engine interface, and you can request access from Neurocognition and Natural Language Processing Research Lab of UFRGS[6]. We refer to Portuguese BERT as BERT1.

**b) BERT-multilingual:** BERT-multilingual was pre-trained on the Wikipedia dumps for the top 100 languages[7] with the largest Wikipedia pages, including Portuguese. Even though it was trained following the standard BERT procedure and without a cross-lingual objective, it was proven to generalize well across languages for a wide variety of tasks [84]. We refer to BERT-multilingual as BERT2.

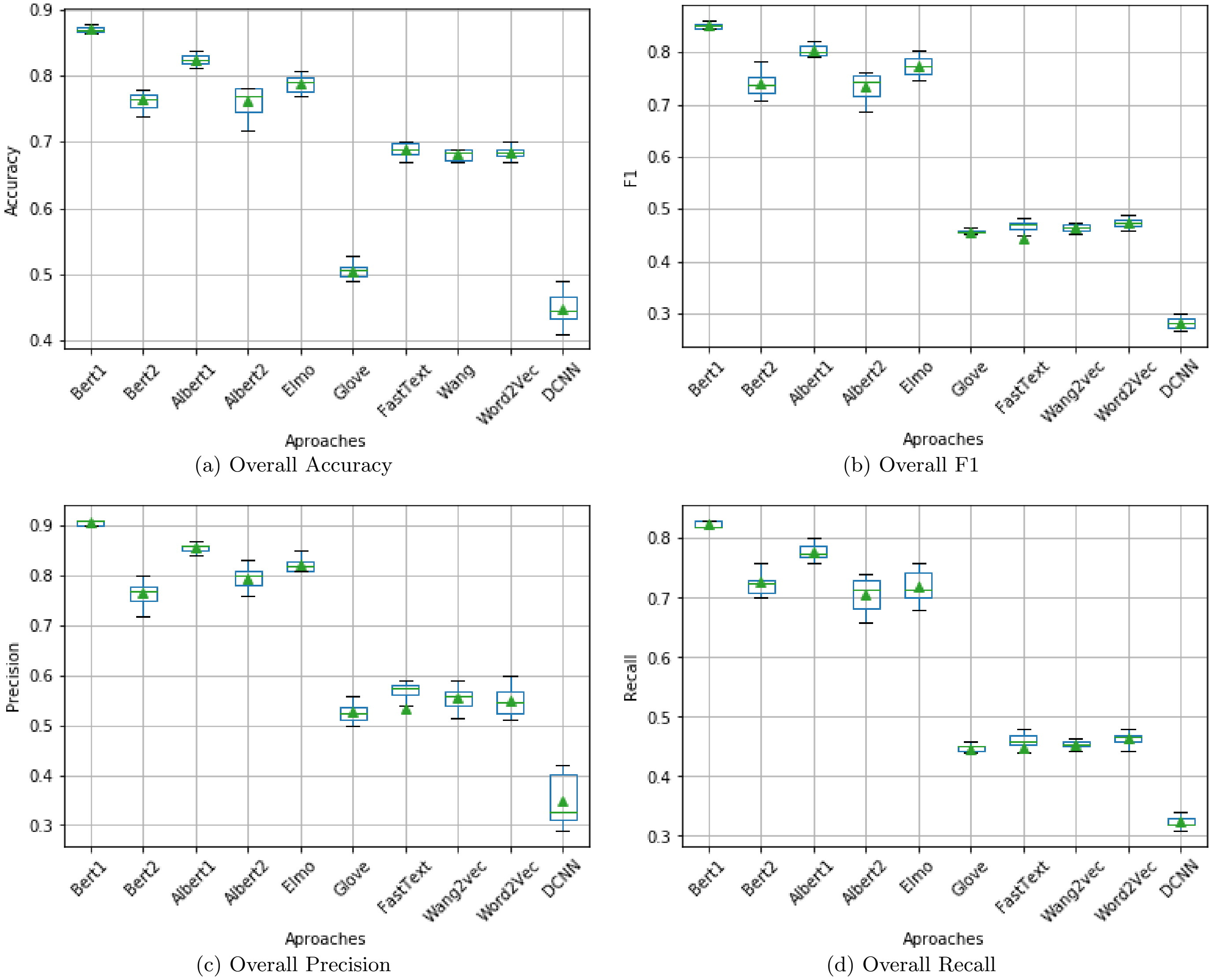
**c) Portuguese ALBERT**: ALBERT [32] incorporates two parameter reduction techniques that eliminate the major obstacles involved with scaling pretrained models. The first one is factorized embedding parameterization; by decomposing the large vocabulary embedding matrix into two small matrices, they separated the size of the hidden layers from that of the vocabulary embedding. This separation facilitates the growth of the hidden layer size without significantly increasing the parameter size of the vocabulary embeddings. The second technique is cross-layer parameter sharing. This technique prevents the parameters from growing with the depth of the network. Both techniques significantly reduce the number of parameters for BERT without impacting the performance, thus improving the parameter efficiency. An ALBERT configuration similar to BERT-large has 18 times fewer parameters and can be trained approximately 1.7. times faster. The parameter reduction techniques also act as a form of regularization that stabilizes the training and helps with generalization. In this study, we pre-trained a model on the BrWaC dataset [83] and another with Wikipedia dump[8]. We refer to ALBERT1, the ALBERT trained on

[5] https://allennlp.org/elmo

[6] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC

[7] https://github.com/google-research/bert/blob/master/multilingual.md

[8] https://dumps.wikimedia.org



(a) Overall Accuracy

(b) Overall F1

(c) Overall Precision

(d) Overall Recall

Fig. 5: Comparison between Deep Neural language Models considered for WhatsApp SA considering F1, Precision, Recall, and Accuracy. BERT1 is related to Portuguese BERT, BERT2 is related to BERT-multilingual, ALBERT1 is equivalent to ALBERT pre-trained with Brazilian Web Corpus [83], ALBERT2 is equivalent to ALBERT pre-trained with Wikipedia dump.

the first, and ALBERT2, the ALBERT trained on the second.

### *D. Evaluation metrics*

The overall accuracy was evaluated to assess the performance of DNNs and DNLMs during the training stage. The overall accuracy, F1 score, precision, precision by class, and recall are the metrics used to assess the performance of DNNs and DNLMs. More details about the calculated evaluation metrics can be found in [85], [86].

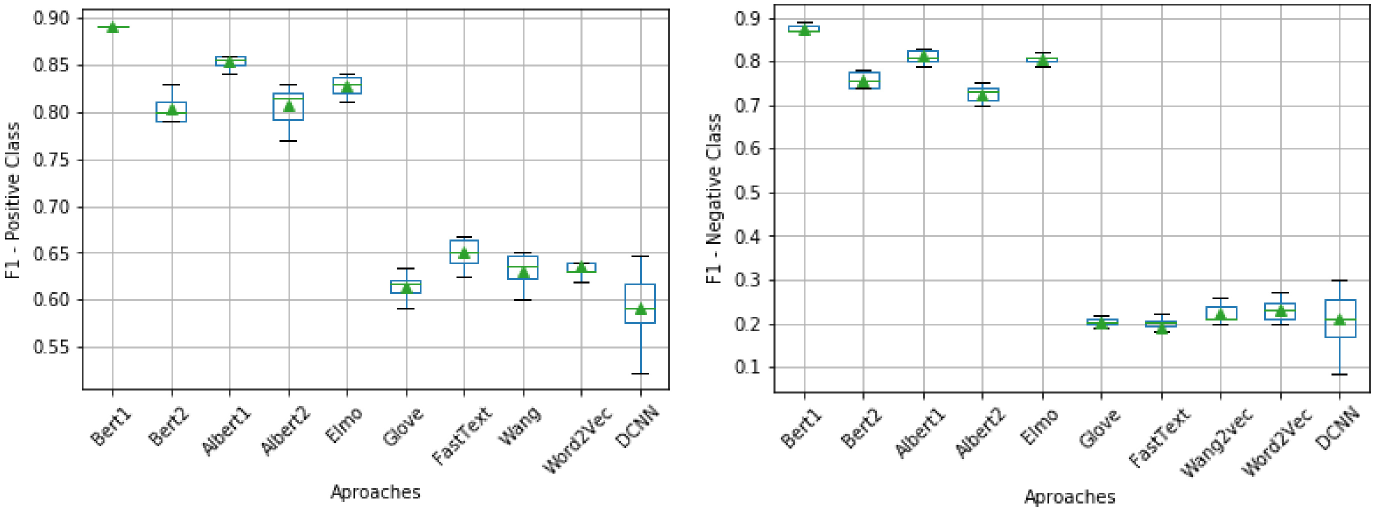
## VII. Results and Discussions

In this section, we evaluate the performance of WhatsApp sentiment classification with the DNLMs and DNNs described in Section V and Section VI. Figures 5 and 6 present a comparative between all considered models. All classifiers were carried out 10 runs, and we present the average and standard deviations. We used Python with the Hugging Face library [9] for BERT and its variations, AlleNLP[10] to implement the ELMo model, and Tensorflow [87] for LSTM and DCNN.

Two ways of word representation were evaluated in WhatsApp Sentiment Analysis for Portuguese: contextualized and static word representations. Even though the DCNN and LSTM with static embeddings obtained a good accuracy in SA on Twitter and other social media [57], [58], according to static embeddings, considering the overall accuracy (see Figure 5 (a)), it can be observed

[9] https://github.com/huggingface/transformers

[10] https://allennlp.org/



(a) F1 Positive

(b) F1 Negative

Fig. 6: Comparison between the Deep Neural language Models considering F1 per class positive and negative. BERT1 is related to Portuguese BERT, BERT2 is related to BERT-multilingual, ALBERT1 is equivalent to ALBERT pre-trained with Brazilian Web Corpus [83], ALBERT2 is equivalent to ALBERT pre-trained with Wikipedia dump.

that FastText, Wang2Vec, and Word2Vec exhibited similar performance; while GloVe demonstrated poor performance due to the bad performance on the positive and negative classes (see Figure 6 (a,b)). We can also conclude that the DCNN, a CNN-based approach, had the worst performance, which was expected because this approach does not use any external knowledge, and the same words will always have the same representation regardless of the context in which they occur.

On the other hand, despite the fact that DNN approaches usually achieve good performance when they are pre-trained on a larger corpus of other domains [57], [58], the achieved results are quite different for WhatsApp domain. This can be explained by the fact that the characteristics of the sub-language domain in the static word embeddings are different from the WhatsApp sub-language (e.g., shorter sentences and misspellings. From Figure 5, we can derive some useful insights. First, Portuguese BERT (BERT1) was the best classifier, with the lowest standard deviation and the highest average, considering all evaluation metrics. ALBERT, pre-trained with Brazilian Web Corpus [83], was the second best classifier; with a simpler network architecture, it guaranteed lesser complexity in training time than BERT1. All dynamic strategies were better than the static ones, which was expected due to the improvement in the employed context. Interestingly, we found that the performance of BERT-Multilingual, which is not a exclusively Portuguese knowledge base, and with the help of the transfer learning from the pre-training stage, was better than all other static strategies trained in Portuguese (Word2Vec, FastText, Wang2Vec, and Glove).

Figure 6 allows us to visualize analyze the performance by class, considering that the neutral class is the minority and the positive is the majority. Portuguese BERT (BERT1) remained the best classifier, with the highest average and the lowest standard deviation.

From Figures 5, 6, and 7, we can conclude that the dynamic embeddings were the best choice to solve the WhatsApp SA problem; even when transfer learning is not pre-trained in a specific language, the best option is dynamic embedding with the BERT-Multilingual. It is important to notice that the ELMo embedding approach achieved high accuracy, precision, recall, and F1 score; that is, it exhibited the third-best overall performance, significantly similar to the transfer learning approaches. ELMo offered a good trade-off between the accuracy and training time complexity [88], and in our context it performed efficiently in spite of the small size of our corpus training, and also in spite of the fact that it was associated with the class imbalance from training corpus and the noise within the data.

BERT and its variations are different from ELMo primarily because they target a different training objectives. The main limitation of the earlier works is the inability to consider both the left and right contexts of the target word, because the language model objective is generated from left to right, by adding successive words to a sentence. Moreover, ELMo model that implements Bi-LSTM, simply concatenated the left-to-right and right-to-left information; this indicates that the representations could not utilize the left and right contexts simultaneously. Besides, ELMo has much fewer parameters than ALBERT which has simpler deep neural model than BERT therefore, ELMo can achieve the lowest training time complexity. BERT essentially uses transformers, whereas ELMo uses LSTMs and CNNs. In addition to the fact that these two approaches work differently, it should also be noted that using transformers enables the parallelization of training, which is an important factor when working with large amounts of data.

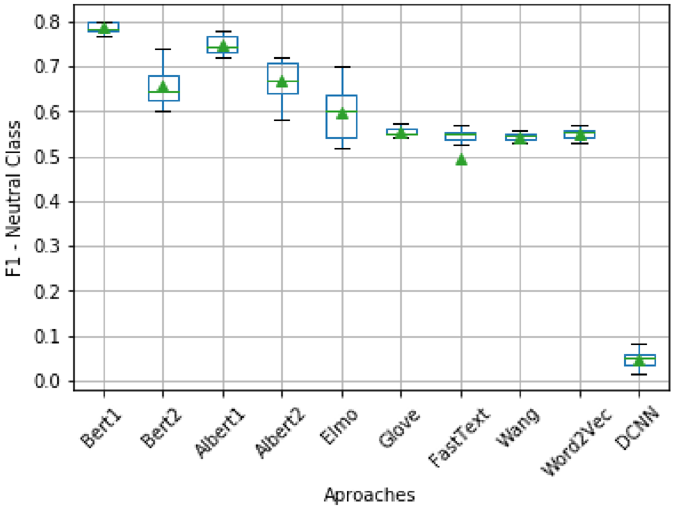


Fig. 7: F1 Neutral.BERT1 is related to Portuguese BERT, BERT2 is related to BERT-multilingual, ALBERT1 is equivalent to ALBERT pre-trained with Brazilian Web Corpus [83], ALBERT2 is equivalent to ALBERT pre-trained with Wikipedia dump.

To the best of our knowledge, no study has compared all methods included in this study. However, it is a consensus among the community [89] that BERT, a pretrained deep deep bidirectional Language Model (biLM) proposed by Google, and its variations (ALBERT, for example), achieved the state-of-the-art performance for most NLP tasks, including SA.

Considering the Stanford Sentiment Treebank task of Glue benchmark[11], which consists of movie reviews, it should be noted that the variations of ALBERT were ranked first and third in our experiment, whereas a variation of BERT achieved the second rank on the leaderboard. In the WhatsApp scenario, BERT exhibited the best performance, and the accuracy of ELMo increased, similar to the variations of ALBERT and BERT. DCNN and static word embeddings in LSTM achieved lower accuracy than that usually obtained in the analysis of Twitter and consumer reviews.

### A. *Statistical Evaluation*

In Table IV we present the P-values of T-test application comparing each technique against others in the Test Corpus, and according to the data shown, BERT2 does not statistically differ from ALBERT2; i.e., they are equivalent. GloVe does not statistically differ from FastText, and FastText is equivalent to Wang2Vec and Word2Vec. It is worth noting that ELMo achieved a $p - value$ of 0.012 in relation to ALBERT2, close to the threshold ($> 0.05$), demonstrating that ELMo and ALBERT2 are almost the same statistically.

## VIII. Conclusions and Future Work

SA approaches are powerful tools for decision-making and recommender systems in various domains [90]–[94]. WhatsApp is a popular social network that has peculiar characteristics (misspelling, slangs, few and smaller size corpus, close strong social ties, shorter sentences), which defines it as a distinct sub-language. DNLMs and DNNs approaches are currently employed for Twitter, blogs and consumer/products reviews. However, to the best of our knowledge, no study has evaluated the most recent and relevant DNLMs and DNNs approaches using textual data from WhatsApp.

In this study, we built our corpora from closed groups with WhatsApp sentences from the Portuguese language. The sentences were labeled with three sentiment classes: positive, negative, or neutral. We computed the Fleiss's Kappa score to evaluate the agreement between the raters. The Fleiss's Kappa score showed almost perfect and moderate agreement for each training and test corpus, respectively. In addition, we found that the size of WhatsApp sentences was 23 characters, on average (shorter than that of Twitter[12]).

We also performed a comparative study on static and dynamic transfer learning-based deep learning models in WhatsApp SA. We implemented some transfer learning approaches (BERT1, BERT2, ALBERT 1, ALBERT2), hybrid deep neural networks (ELMo) and static word embeddings (DCNN and LSTM with FastText, Wang, Word2Vec and Glove) to evaluate the performance considering the main metrics for sentiment classification, such as the overall accuracy, precision, recall, and F1 score. The experiments demonstrated that transfer learning based on deep learning architectures obtained the highest accuracy and precision, which is similar to the results of SA on Twitter and consumer reviews; whereas static word embeddings in LSTM and DCNN yielded the lowest precision (much lower than those usually obtained for Twitter and consumer reviews).

In future work, we intend to adopt and develop a deep learning approach that can classify sentiments in embedded systems and smartphones. Therefore, sentiment could be useful knowledge for developing decision support and recommender systems, mostly for e-commerce, tourism, financial market, and network management on mobile networks, such as the correlation between sentiment and network usage, products, and places. Furthermore, a comparison between other platforms such as Telegram would bring a greater understanding of the sentiments represented in these environments.

Another question to be raised from this study is how representative the WhatsApp corpus needs to be to be generalized to that language variety [95]. This question is impacted by the fact that the corpus was collected over a period of ten months, from existing and closed groups, and the messages from people who volunteered to participate in the research.

[11] https://gluebenchmark.com/leaderboard

[12] https://smk.co/article/the-average-tweet-length-is-28-characters-long-and-o

Table IV: P-values of T-test application comparing each technique against others in the Test Corpus. If $p - value > 0.05$, the techniques do not statistically differ from each other, while a $p - value \leq 0.05$ indicates that they statistically differ among themselves. We considered the overall F1.

| | **BERT1** | **BERT2** | **ALBERT1** | **ALBERT2** | **Elmo** | **Glove** | **FastText** | **Wang2Vec** | **Word2Vec** | **DCNN** |
|---|---|---|---|---|---|---|---|---|---|---|
| **BERT1** | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **BERT2** | 0 | - | 0 | **0.471** | 0.007 | 0 | 0 | 0 | 0 | 0 |
| **ALBERT1** | 0 | 0 | - | 0 | 0.006 | 0 | 0 | 0 | 0 | 0 |
| **ALBERT2** | 0 | **0.471** | 0 | - | 0.012 | 0 | 0 | 0 | 0 | 0 |
| **ELMo** | 0 | 0.007 | 0.006 | 0.012 | - | 0 | 0 | 0 | 0 | 0 |
| **Glove** | 0 | 0 | 0 | 0 | 0 | - | **0.648** | 0.024 | 0.001 | 0 |
| **FastText** | 0 | 0 | 0 | 0 | 0 | **0.648** | - | **0.422** | **0.250** | 0 |
| **Wang2Vec** | 0 | 0 | 0 | 0 | 0 | 0.024 | **0.422** | - | 0.017 | 0 |
| **Word2Vec** | 0 | 0 | 0 | 0 | 0 | 0.001 | **0.250** | 0.017 | - | 0 |
| **DCNN** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - |

## Referências

[1] B. Liu, “Sentiment analysis: A multifaceted problem,” *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 76–80, Aug. 2010, ISSN: 1094-7167. DOI: 10.1109/MIS.2010.75.

[2] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (Studies in Natural Language Processing). Cambridge University Press, 2020, ISBN: 9781108486378.

[3] R. Grishman and R. Kittredge, *Analyzing language in restricted domains: sublanguage description and processing*. Psychology Press, 2014.

[4] R. I. Kittredge, *The oxford handbook of computational linguistics, chapitre sublanguages and controlled languages*, 2003.

[5] T. Lippincott, D. Ó. Séaghdha, and A. Korhonen, “Exploring subdomain variation in biomedical language,” *BMC bioinformatics*, vol. 12, no. 1, p. 212, 2011.

[6] C. Mihaila, R. T. Batista-Navarro, and S. Ananiadou, “Analysing entity type variation across biomedical subdomains,” in *Third workshop on building and evaluating resources for biomedical text mining*, 2012, pp. 1–7.

[7] N. F. F. da Silva, E. R. Hruschka, and E. R. H. Jr., “Tweet sentiment analysis with classifier ensembles,” *Decis. Support Syst.*, vol. 66, pp. 170–179, 2014. DOI: 10.1016/j.dss.2014.07.003. [Online]. Available: https://doi.org/10.1016/j.dss.2014.07.003.

[8] L. F. S. Coletta, N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, “Combining classification and clustering for tweet sentiment analysis,” in *2014 Brazilian Conference on Intelligent Systems*, 2014, pp. 210–215.

[9] N. F. F. D. Silva, L. F. S. Coletta, and E. R. Hruschka, “A survey and comparative study of tweet sentiment analysis via semi-supervised learning,” *ACM Comput. Surv.*, vol. 49, no. 1, Jun. 2016, ISSN: 0360-0300.

[10] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” *Information Sciences*, vol. 311, no. 0, pp. 18–38, 2015, ISSN: 0020-0255.

[11] J. M. Chenlo and D. E. Losada, “An empirical study of sentence features for subjectivity and polarity classification,” *Information Sciences*, vol. 280, no. 0, pp. 275–288, 2014, ISSN: 0020-0255.

[12] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011, ISSN: 0020-0255.

[13] C. Catal and M. Nangir, “A sentiment classification model based on multiple classifiers,” *Applied Soft Computing*, vol. 50, pp. 135–141, 2017, ISSN: 1568-4946.

[14] R. M., V. R. Hulipalled, K. Venugopal, and L. Patnaik, “Consumer insight mining: Aspect based twitter opinion mining of mobile phone reviews,” *Applied Soft Computing*, vol. 68, pp. 765–773, 2018, ISSN: 1568-4946.

[15] J. Clement, *Whatsapp - statistics and facts*, https://www.statista.com/topics/2018/whatsapp/, Statista.

[16] S. KEMP, *The state of digital in april 2019: All the numbers you need to know*, https://wearesocial.com/uk/blog/2019/04/the-state-of-digital-in-april-2019-all-the-numbers-you-need-to-know, SWe Are Social.

[17] E. Karapanos, P. Teixeira, and R. Gouveia, “Need fulfillment and experiences on social media: A case on facebook and whatsapp,” *Computers in Human Behavior*, vol. 55, pp. 888–897, 2016, ISSN: 0747-5632.

[18] A. B. Boot, E. T. K. Sang, K. Dijkstra, and R. A. Zwaan, “How character limit affects language usage in tweets,” *Palgrave Communications*, vol. 5, no. 1, p. 76, 2019. DOI: 10.1057/s41599-019-0280-3. [Online]. Available: https://www.nature.com/articles/s41599-019-0280-3.

[19] M. C. Dıaz-Galiano, M. G. Vega, E. Casasola, *et al.*, “Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus,” in *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th,*

*2019*, M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, *et al.*, Eds., ser. CEUR Workshop Proceedings, vol. 2421, CEUR-WS.org, 2019, pp. 550–560. [Online]. Available: http://ceur-ws.org/Vol-2421/TASS%5C_overview.pdf.

[20] V. Basile, C. Bosco, E. Fersini, *et al.*, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. [Online]. Available: https://www.aclweb.org/anthology/S19-2007.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[22] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., Curran Associates, Inc., 2017, pp. 3856–3866. [Online]. Available: http://papers.nips.cc/paper/6975-dynamic-routing-between-capsules.pdf.

[23] D. Cer, Y. Yang, S.-y. Kong, *et al.*, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. DOI: 10.18653/v1/D18-2029. [Online]. Available: https://www.aclweb.org/anthology/D18-2029.

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *Albert: A lite bert for self-supervised learning of language representations*, 2019. arXiv: 1909.11942 [cs.CL].

[25] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692.

[26] W. Wang, B. Bi, M. Yan, *et al.*, "Structbert: Incorporating language structures into pre-training for deep language understanding," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJgQ4lSFPH.

[27] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio, *Portuguese word embeddings: Evaluating on word analogies and natural language tasks*, 2017. arXiv: 1708.06025 [cs.CL].

[28] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese named entity recognition using bert-crf," *arXiv preprint arXiv:1909.10649*, 2019. [Online]. Available: http://arxiv.org/abs/1909.10649.

[29] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, *Deep contextualized word representations*, 2018. arXiv: 1802.05365 [cs.CL].

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018. arXiv: 1810.04805 [cs.CL].

[31] L. Zhang, "Social sentiment analysis using classifiers and ensemble learning," *Journal of Physics: Conference Series*, vol. 1237, p. 022193, Jun. 2019.

[32] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS.

[33] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[34] S. F. Waterloo, S. E. Baumgartner, J. Peter, and P. M. Valkenburg, "Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp," *new media & society*, vol. 20, no. 5, pp. 1813–1831, 2018.

[35] B. Agarwal, R. Nayak, N. Mittal, and S. Patnaik, *Deep Learning-Based Approaches for Sentiment Analysis* (Algorithms for Intelligent Systems). Springer Singapore, 2020, ISBN: 9789811512162.

[36] S. Joshi, "Sentiment analysis on whatsapp group chat using r," in Jan. 2019, pp. 47–55, ISBN: 978-981-13-6346-7.

[37] D. Jain, A. Garg, and M. Saraswat, "Sentiment analysis using few short learning," in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 2019, pp. 102–107.

[38] S. Dahiya, A. Mohta, and A. Jain, "Text classification based behavioural analysis of whatsapp chats," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 717–724. DOI: 10.1109/ICCES48766.2020.9137911.

[39] R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, *et al.*, "Sentiment analysis of customers' reviews using a hybrid evolutionary svm-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022.

[40] M. Li, M. Ge, H. Zhao, and Z. An, "Modeling and analysis of learners' emotions and behaviors based on online forum texts," *Computational intelligence and neuroscience*, vol. 2022, 2022.

[41] H. Yin, X. Song, S. Yang, and J. Li, "Sentiment analysis and topic modeling for covid-19 vaccine discussions," *World Wide Web*, vol. 25, no. 3, pp. 1067–1083, 2022.

[42] X.-z. Yang, S. Wu, T. Ren, and N. Li, "Research on intelligent sentiment analysis and theme evolution of

personal technology blog," in *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 2022, pp. 81–87.

[43] M. Seufert, T. Hoßfeld, A. Schwind, V. Burger, and P. Tran-Gia, "Group-based communication in whatsapp," in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, 2016, pp. 536–541. DOI: 10.1109/IFIPNetworking.2016.7497256.

[44] A. Rosenfeld, S. Sina, D. Sarne, O. Avidov, and S. Kraus, "WhatsApp usage patterns and prediction of demographic characteristics without access to message content," *Demographic Research*, vol. 39, no. 22, pp. 647–670, 2018. DOI: 10.4054/DemRes.2018.39.22.

[45] J. Kaur, A. S. Wani, and P. Singh, "Engagement of pregnant women and mothers over whatsapp: Challenges and opportunities involved," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '19, Austin, TX, USA: Association for Computing Machinery, 2019, pp. 236–240, ISBN: 9781450366922. DOI: 10.1145/3311957.3359481. [Online]. Available: https://doi.org/10.1145/3311957.3359481.

[46] D. Lambton-Howard, R. Anderson, K. Montague, *et al.*, "Whatfutures: Designing large-scale engagements on whatsapp," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14, ISBN: 9781450359702. DOI: 10.1145/3290605.3300389. [Online]. Available: https://doi.org/10.1145/3290605.3300389.

[47] Afiyati, E. Winarko, and A. Cherid, "Recognizing the sarcastic statement on whatsapp group with indonesian language text," in *2017 International Conference on Broadband Communication, Wireless Sensors and Powering (BCWSP)*, 2017, pp. 1–6. DOI: 10.1109/BCWSP.2017.8272579.

[48] G. Resende, P. Melo, J. C. S. Reis, M. Vasconcelos, J. Almeida, and F. Benevenuto, "Analyzing textual (mis)information shared in whatsapp groups," in *Proceedings of the ACM Conference on Web Science*, ser. WebSci'19, Boston, USA, 2019.

[49] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[50] M. I. Prabha and G. Umarani Srikanth, "Survey of sentiment analysis using deep learning techniques," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2019, pp. 1–9. DOI: 10.1109/ICIICT1.2019.8741438.

[51] N. Rupavathy, M. J. C. M. Belinda, G. Nivedhitha, and P. M. Abhinaya, "Whatsapp sentiment analysis," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no. 11-12, pp. 3462–3465, 2018.

[52] J. Fleiss *et al.*, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[53] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy*, vol. 85, no. 3, pp. 257–268, Mar. 2005.

[54] J. Mikk, "Sentence length for revealing the cognitive load reversal effect in text comprehension," *Educational Studies*, vol. 34, no. 2, pp. 119–127, 2008. DOI: 10.1080/03055690701811164. eprint: https://doi.org/10.1080/03055690701811164. [Online]. Available: https://doi.org/10.1080/03055690701811164.

[55] K. König, "Sequential patterns in sms and whatsapp dialogues: Practices for coordinating actions and managing topics," *Discourse and Communication*, vol. 13, no. 6, pp. 612–629, 2019. DOI: 10.1177/1750481319868853. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1750481319868853.

[56] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, Aug. 2018, ISSN: 1556-603X. DOI: 10.1109/MCI.2018.2840738.

[57] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *Meeting Association Computational Linguistics*, vol. 1, no. 52, pp. 655–665, 2014.

[58] A. F. Agarap and P. Grafilon, *Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn)*, 2018. arXiv: 1805.03687 [cs.CL].

[59] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, ISSN: 0001-0782. DOI: 10.1145/361219.361220.

[60] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research*, vol. 49, pp. 1–47, 2014.

[61] T. Mikolov, Q. V. Le, and I. Sutskever, *Exploiting similarities among languages for machine translation*, 2013. arXiv: 1309.4168 [cs.CL].

[62] R. Socher, A. Perelygin, J. Wu, *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://www.aclweb.org/anthology/D13-1170.

[63] X. Chen, Z. Liu, and M. Sun, "A unified model for word sense representation and disambiguation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Lin-

guistics, Oct. 2014, pp. 1025–1035. DOI: 10.3115/v1/D14-1110. [Online]. Available: https://www.aclweb.org/anthology/D14-1110.

[64] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.," in *Interspeech*, 2013, pp. 3771–3775.

[65] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

[67] W. Ling, C. Dyer, A. W. Black, and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1299–1304.

[68] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[69] J. Camacho-Collados and M. T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning," *Journal of Artificial Intelligence Research*, vol. 63, pp. 743–788, 2018.

[70] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, *Pre-trained models for natural language processing: A survey*, 2020. arXiv: 2003.08271 [cs.CL].

[71] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, *Character-aware neural language models*, 2015. arXiv: 1508.06615 [cs.CL].

[72] C. P. Soto, G. M. Nunes, J. G. R. Gomes, and N. Nedjah, "Application-specific word embeddings for hate and offensive language detection," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27 111–27 136, 2022.

[73] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[74] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, *What does bert look at? an analysis of bert's attention*, 2019. arXiv: 1906.04341 [cs.CL].

[75] C. Raffel, N. Shazeer, A. Roberts, *et al.*, *Exploring the limits of transfer learning with a unified text-to-text transformer*, 2019. arXiv: 1910.10683 [cs.LG].

[76] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *HLT-NAACL*, 2016.

[77] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[78] W. Ling, C. Dyer, A. W. Black, and I. Trancoso, "Two/too simple adaptations of Word2Vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1299–1304.

[79] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *In EMNLP*, 2014.

[80] NILC, *Repositório de word embeddings do nilc*, http://nilc.icmc.usp.br/embeddings, NILC - Núcleo Interinstitucional de Linguística Computacional.

[81] M. Gardner, J. Grus, M. Neumann, *et al.*, *Allennlp: A deep semantic natural language processing platform*, 2018. arXiv: 1803.07640 [cs.CL].

[82] T. Wolf, L. Debut, V. Sanh, *et al.*, *Huggingface's transformers: State-of-the-art natural language processing*, 2019. arXiv: 1910.03771 [cs.CL].

[83] J. A. Wagner Filho, R. Wilkens, M. Idiart, and A. Villavicencio, "The brwac corpus: A new open resource for brazilian portuguese," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[84] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, *Cross-lingual ability of multilingual bert: An empirical study*, 2020.

[85] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall, 2010.

[86] K. Faceli, A. C. L. anda João Gama, and A. C. P. L. F. de Carvalho, *Inteligência Artificial Uma abordagem de Aprendizado de Máquina*, 1a edição. LTC, 2015, ISBN 978-85-216-1880-5.

[87] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[88] L. H. Li, P. H. Chen, C. Hsieh, and K. Chang, "Efficient contextual representation learning without softmax layer," *CoRR*, vol. abs/1902.11269, 2019. arXiv: 1902.11269. [Online]. Available: http://arxiv.org/abs/1902.11269.

[89] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. [Online]. Available: https://www.aclweb.org/anthology/W18-5446.

[90] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word

of mouth," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, Nov. 2009, ISSN: 1532-2882.

[91] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media.," in *HLT-NAACL*, The Association for Computational Linguistics, 2012, pp. 656–666, ISBN: 978-1-937284-20-6.

[92] M. Cheong and V. C. Lee, "A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter," *Information Systems Frontiers*, vol. 13, no. 1, pp. 45–59, Mar. 2011, ISSN: 1387-3326.

[93] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10, Atlanta, Georgia, USA: ACM, 2010, pp. 1195–1198, ISBN: 978-1-60558-929-9.

[94] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 406–418, Feb. 2011, ISSN: 1532-2882.

[95] D. Biber, "Representativeness in corpus design," in *Current Issues in Computational Linguistics: In Honour of Don Walker*, A. Zampolli, N. Calzolari, and M. Palmer, Eds. Dordrecht: Springer Netherlands, 1994, pp. 377–407, ISBN: 978-0-585-35958-8. DOI: 10.1007/978-0-585-35958-8_20. [Online]. Available: https://doi.org/10.1007/978-0-585-35958-8_20.