On Text Preprocessing for Early Detection of Depression on Social Media

José Solenir L. Figuerêdo[1], **Rodrigo Tripodi Calumby**[1]

[1] University of Feira de Santana – Feira de Santana – BA – Brazil

jslfigueredo@ecomp.uefs.br, rtcalumby@uefs.br

Abstract. *Depression is a serious challenge to public health. Many of those who suffer from this disease use social media for information or relief. The text data produced by these users can be used to support research in this field. However, this raw information is not always suitable for use directly in machine learning. Hence, a comparative analysis was performed between different preprocessing techniques to verify the impact on the effectiveness of early depression detection on social media. The results show that the preprocessing contributes to an increase in the prediction effectiveness. Moreover, the mapping of emoticons to real emotion words was decisive to improve not only model's effectiveness, but also to keep the balance between different evaluation measures.*

1. Introduction

Mental illnesses are becoming one of the most prevalent health problems worldwide. Among these diseases, depression stands out due to the numerous problems that it can trigger on the person. According to recent estimates by the World Health Organization (WHO), more than 300 million people suffer from this disease worldwide [Organization 2017]. The WHO also indicated an increase of 18.8% in the number of cases of depressed people, considering the years 2005 and 2015. In relative terms, in 2015 people with depression represented about 4.4% of the world population, more commonly in women (5.1%) compared to men (3.6%). In addition, in the same study, it was found that depression affects people of different ages, but especially those between 55 and 74 years, with the elderly being the most vulnerable.

Depression, unlike other diseases, beyond directly affecting the population, can also negatively affect other general health conditions such as cardiovascular disease, cancer, diabetes mellitus and obesity [Choudhury et al. 2013b]. Moreover, depression is also known to have negative influences on individuals' family and personal relationships, work and school life [Choudhury et al. 2013b]. Of an even greater concern is the fact that depression is one of the main reasons that lead to suicide. In 1990, a pioneer study by Goodwin and Jamison suggests that depression is the leading cause of suicide [Goodwin and Jamison 1990], considering that between 15% and 20% of all patients with depression take their lives. In another study, conducted by Richards and O'Hara [Richards and O'Hara 2014], it was found that approximately two thirds of people who die by suicide are dealing with depression at the time of death. The WHO data also showed this relationship between depression and suicide. According to the WHO, approximately 788,000 people died from suicide in 2015, with far more people trying unsuccessfully [Organization 2017]. Among other factors, these facts demonstrate the

severity of this disease, and the urgency to find ways to treat it, or at least reduce its symptoms or impacts.

For several reasons, many of the people who suffer from depression do not receive adequate treatment. One of these reasons is the fact that most patients are unaware of their condition and, therefore, do not seek clinical intervention until the symptoms become severe. Ideally however, the detection of depression in early stages is critical, given it would allow a more adequate and effective treatment. Besides, it is important to note that late diagnosis is not the only problem when it comes to enabling adequate treatment. In this context, the strong social stigma associated with clinical depression leads many patients to avoid seeking professional help to assess their situation. Hence, for many reasons, many people end up resorting to less formal alternatives, such as social media and the Internet, to obtain information about their condition and also to discuss about their mental state [Yates et al. 2017].

The availability of social media platforms, such as Facebook[1], Twitter[2] or Reddit[3] made it possible for people to share their personal experiences, ideas or thoughts in a free and comprehensive way. This kind of technology allows the production a large amounts of data which, therefore, generates a myriad of opportunities to solve problems in different application fields [Aggarwal 2011]. From this, many studies have been conducted on subjects related to sentiment analysis, personalized recommendation systems, public opinion monitoring, among others [Schoen et al. 2013]. Most of these studies rely on machine learning techniques as well as other purely statistical methods. In the context of this paper, some studies have been conducted using social media to investigate people's mental state, especially depression [Choudhury et al. 2013c, Wang et al. 2013, Tsugawa et al. 2015, Coppersmith et al. 2015, Benamara et al. 2018].

Several studies indicate that mental disorders also interferes on the use of language by affected people. Thus, the fact that social media are an abundant source of textual data makes them an interesting source for the investigation of such disorder and its possible identification. The relationship between the use of language and clinical disorders has been studied for decades [Pennebaker et al. 2003, Rude et al. 2004]. However, it is important to notice that these data in their raw form may include information that is not relevant to support these tasks.

Therefore, an adequate preparation of this data is necessary. Hence, understanding the importance of this step is of great relevance, given its direct impact on the construction machine learning models. It is expected that a more appropriate treatment of the data will contribute to the construction of more appropriate models and, consequently, allow better effectiveness on the intended task.

In summary, this work investigates the role of textual data preprocessing in the task of early detection of depression on social media. To this end, benchmark experiments were conducted with multiple preprocessing techniques to assess their impact on the construction of depression prediction models. Consequently, it contributes to the task by providing evidences of the relevance and effectiveness impact of different text pre-

[1]https://www.facebook.com/ - As of March 18, 2020
[2]https://twitter.com/ - As of March 18, 2020
[3]https://www.reddit.com/ - As of March 18, 2020

processing techniques. Moreover, it may allow better choices for the construction of applications or future research works on the field.

The remainder of this paper is organized as follows. Section 2 presents the related work and Section 3 describes the proposed experimental process. The results and discussions are presented in Section 4. Finally, Section 5 brings the conclusions and future work.

2. Related Works

Despite the advances on its treatment, depression is one of the fastest growing diseases in the world [Organization 2017]. Among other things, this fact motivates the development of studies that can contribute to reduce the impact of the problem. Many researchers have conducted studies aimed at detecting depression and other mental disorders, using data from different sources [Cavazos-Rehg et al. 2016, Losada and Crestani 2016, Yang and Srinivasan 2016, Santana et al. 2018]. One of these sources and which has enabled the development of many studies are the social media platforms, considered a promising instrument for public health [Choudhury et al. 2013a] investigations. Online domains, such as social media, have created a new ecosystem for innovative research with a rich source of data and social metadata to capture users' behavioral trends. In these applications, Natural Language Processing techniques have been used in combination with machine learning methods, e.g., for the discovery of predictive models. In terms of data sources, many social media platforms have been used to carry out these studies, with emphasis on Twitter and Reddit.

Most of the work in the literature is based on supervised learning [Nakamura et al. 2014, Vedula and Parthasarathy 2017]. These studies show that people with depression tend to have common patterns of behavior. In general, these studies indicate that people who suffer from depression tend, for example: to talk more about relationships and life (e.g., friends, home, dating and death); become more concerned with themselves (they use the first person pronoun very often); use more emoticons (e.g., ":(", ":-(", ":-c"), words of negative emotions (e.g., "anger" and "anxiety") and denial terms (e.g., "no", "none" and "never"); and constantly remember the past and worry about the future. Considering such findings, more studies can be developed focused on the feature extraction process that are later used with classic machine learning models [Nadeem 2016, Almeida et al. 2017, Trotzek et al. 2020].

In [Nadeem 2016], using a crowdsourcing method to build a list of Twitter users who express being diagnosed with depression, a model was proposed to define which of these tweets could indicate depression. For this, the bag-of-words model (frequency of occurrence of words to characterize the content of a tweet) was used for feature representation. To train the prediction models, logistic regression, Naïve Bayes and SVM methods were applied. Using a Corpus of 2.5 million tweets, the experimental results reached about 81% classification accuracy.

In the context of the CLEF eRisk Pilot Task 2017[4] on early risk detection on the Internet [Losada et al. 2017], the work in [Almeida et al. 2017] relied on Information Retrieval (IR) and supervised learning (SL) techniques. The proposed IR method retrieved

[4]http://erisk.irlab.org/2017/index.html - As of February 14, 2020.

similar documents from a test document used as a query. The intuition is that using the full content of a user's post as a query should allow a search engine to retrieve semantically similar posts [Almeida et al. 2017]. In the SL approach, a set of feature extraction strategies (e.g., n-grams, word dictionaries, Part of Speech (PoS) selection) were used and submitted to several classifiers. The authors found that based on an ensemble approach, merging the output candidates from all SL-based systems (considering three classifiers and all features), with the output candidates from the IR-based systems outperformed the results obtained by each approach separately. In same context, the approach proposed by [Malam et al. 2017] used different types of features (statistics and linguistics) and the Random Forest method to predict depression. The best results were obtained when using all of the features simultaneously.

In [Trotzek et al. 2020], a convolutional neural network based on different word embeddings was evaluated and compared to a simple logistic regression based on linguistic metadata at the user level (e.g., average number of the term "I", possessive pronouns, and personal pronouns in posts, frequency of use of the expression "my depression" in posts, frequency of use of words describing medicines for "treatment" of depression, among others). An ensemble of both approaches was performed and the authors indicated they have reached the state-of-the-art in early detection of depression, using the same data from the eRisk 2017.

Over the past few years, many works related to depression prediction have been developed. However, to the best our knowledge, none of these works specifically assessed the effect of preprocessing techniques on the learning step, in general, leaving it a secondary role. Hence, unlike previous studies, this work aims at evaluating the preprocessing step in the construction of predictive models for depression detection.

3. Experimental Pipeline

The experimental process carried out in this work is illustrated in Figure 1. There are five stages: The dataset collection, preprocessing, feature extraction (text embeddings), CNN training, and, finally, the effectiveness evaluation. These modules are described in detail in the following.

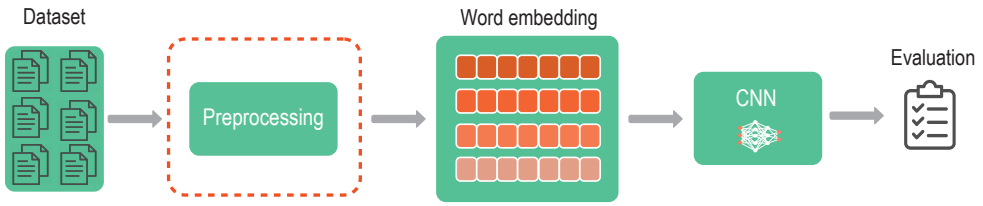

Figure 1. Experimental Benchmark Workflow.

3.1. Dataset

The experiments were conducted with the dataset initially provided by [Losada and Crestani 2016] and published as part of the CLEF eRisk 2017 Task [Losada et al. 2017]. The dataset was built by crawling messages from Reddit users. Reddit has a large community of members and many of them have a long history of submissions (spanning several years). It also has topic divisions (named

"subreddits"), e.g., regarding different medical conditions, such as anorexia and depression [Losada and Crestani 2016]. Each user has a list of posts and each post includes the following fields: title, date, information and text. The information field simply contains the string "reddit post" and was discarded.

The data from 887 users were collected, with 135 users considered as depressive and a control group of 752 labeled as non-depressive. The dataset was randomly divided into training and test sets. The training set consists of 486 users (83 positive, 403 negative). In turn, the test set contains 401 users, with 52 positive and 349 negative. The retrieved posts are organized in chronological order. Thus, it is possible to verify both the difference in the use of language between users with and without depression, and also the evolution of the language used over time.

3.2. Preprocessing techniques

As highlighted in Figure 1, the preprocessing step is the focus of this work. For all evaluated scenarios, a basic preprocessing, which is common in the literature, was applied. Hence, for all cases, this standard preprocessing comprised the removal of numbers, links to websites and mentions to other users. In addition, all words were changed to lowercase, due to the practical requirements of the pre-trained embeddings used for text feature representation in this work (see Section 3.3). In the benchmark proposed here, the specific preprocessing techniques considered were:

- Stop Words Removal: Removal of many words considered irrelevant for the textual analysis (e.g., "a","an", "of","the", etc.);
- Lemmatization: Changing of the inflected forms of a word to a common root (e.g., "We are the champions" becomes "We are the champion");
- Lemmatization with PoS tagging: Reduces the inflected forms of each word to a common base or root, but taking into account its context by Pos tagging (e.g., "I am the happiest person in the world" becomes "I be the happy person in the world");

These techniques were assessed according to multiple use cases regarding their combinations as presented in Table 1. In the table, for "Case 1" no other additional preprocessing was carried out, only the basic preprocessing already mentioned was maintained.

All the cases from Table 1 were also assessed considering an additional dimension related to the importance of emotions represented by emoticons. In the first approach all emoticons were discarded. In turn, the second approach was performed in a way that the emoticons included in the posts were preserved through a mapping to representative terms. For instance, in this mapping, the ":(" symbol was replaced by the term "sad". These variations were assessed with the aim of analyzing whether emotions, represented in this context by emoticons, impacted the identification of users with depression, as suggested by the literature.

This entire preprocessing step was performed using the NLTK[5] library and the Keras preprocessing module[6]. The emoticons mapping was performed according to a

[5]https://www.nltk.org - As of February 14, 2020

[6]https://keras.io/preprocessing/text - As of February 14, 2020

Table 1. Preprocessing cases evaluated.

Category	Stop Words Removal	Lemmatization	PoS tagging
Case 1	-	-	-
Case 2	✓	-	-
Case 3	-	✓	-
Case 4	-	✓	✓
Case 5	✓	✓	-
Case 6	✓	✓	✓

predefined mapping[7].

3.3. Word Embeddings

The experiments carried out in this work relied on pre-trained word embedding models trained based on large datasets, not necessarily associated with depression to any other mental illness. This choice of these pre-trained models was due to the unavailability of large amounts of data related to depression and the recurrent success of this transfer learning procedure in multiple applications.

The specific embedding models used were fastText [Mikolov et al. 2018] and GloVe [Pennington et al. 2014]. The fastText 300-dimensional embedding was trained with data from Wikipedia 2017 and UMBC webbase corpus and statmt.org news dataset (here named "FastText WN"), and also with data obtained via a common crawl (here named "FastText Crawl"). The GloVe model was trained with data from Wikipedia 2014 and Gigaword 5, respectively named "GloVe Crawl" and "Glove WN".

3.4. Convolutional Neural Network Classifier

The CNN classifier architecture used in the experiments is illustrated in Figure 2. This architecture was proposed in [Trotzek et al. 2020], basically consisting of a simple convolution layer (100 filters of height 2 and width corresponding to the dimension of the output text of embedding). In this network, Concatenated Rectified Linear Units (CReLU) are used as an activation function. Then, the 1-max polling method is used to obtain a scalar from each filter, generating a 100-dimensional vector. The output represented by this vector is then propagated through three dense layers that also have CReLU as the activation function. In the output of the first layer, dropout is applied, as a way of reducing overfitting. Finally, in the last layer, the softmax function is applied to obtain the class probabilities (depressive,non-depressive).

3.5. Effectiveness Assessment

The effectiveness assessment was carried out using classical machine learning measures, such as Precision, Recall and F_1. In addition, we also used the evaluation measure used in eRisk 2017, called *Early Risk Detection Error* ($ERDE_o$) [Losada and Crestani 2016], which attributes penalties for late decision making. Basically, decision delay is measured by counting the number (k) of different textual items seen before making a decision. Considering a binary decision (d) made by an early risk detection system at k, $ERDE_o$ is defined as:

[7]https://en.wikipedia.org/wiki/List_of_emoticons - As of February 14, 2020

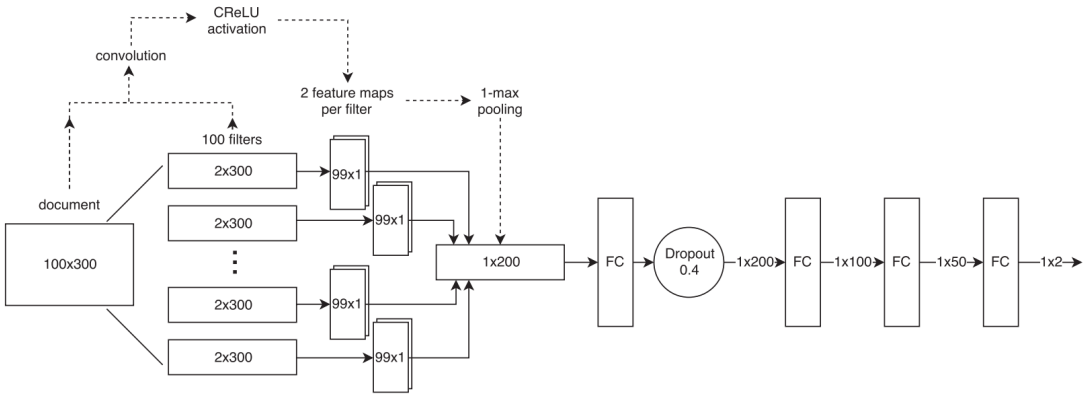

Figure 2. Baseline CNN architecture used in the benchmark. Source [Trotzek et al. 2020]

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{for false positives (FP)} \\ c_{fn} & \text{for false negatives (FN)} \\ lc_o \cdot c_{tp} & \text{for true positives (TP)} \\ 0 & \text{for true negatives (TN)} \end{cases}$$

where, following [Losada et al. 2017], $c_{fn} = c_{tp} = 1$ and c_{fp} is set according to the proportion of positive cases in the test set (0.1296). As indicated in [Losada et al. 2017], c_{fn} and c_{tp} were set at 1 since delayed detection could have serious consequences (i.e., delayed detection would be equivalent to not detecting the problem).

The parameter o controls the point where the cost of a late decision begin to grow faster. The general error is calculated by taking the average of the values of $ERDE_o$. In this study, we used the $ERDE_5$ and $ERDE_{50}$ to evaluate the models. In turn, the function $lc_o(k)$ encodes a cost associated to the delay in detecting true positives, and is calculated according to Eq. 1.

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{1}$$

3.6. Experimental Setup

The experiments carried out in this work strictly followed the procedures established in eRisk 2017. The objective is to detect initial evidences of depression by analyzing users' posts. The idea was to simulate systems that monitor posts on social media to early identify potentially depressive users. To simulate this, the data was divided into ten blocks, each one containing 10% of each user's messages in chronological order. Before the data was submitted to the training or test stages, preprocessing was carried out, following the settings described in the subsection 3.2.

Training is performed with the entire training set only once for each evaluated preprocessing scenario. On the other hand, for the test phase, following the CLEF eRisk 2017 Task, the ten blocks are processed iteratively and incrementally. As mentioned, each block contains 10% of each user's posts, therefore at each iteration i the prediction

is conducted based on the concatenation of all the post blocks till the i^{th} position, thus simulating an incremental chronological cut-off in a user's feed.

Standard processing receives each user post individually in the form of the first 100 words per post, with each word having a vector representation with a dimension of 300, attributed to by embedding model. After this step, zeros are filled for posts with less than 100 words, resulting in a 100x300 matrix as the CNN classifier input. The CNN performs the classification for each post individually, for each user. Thus, it is necessary to aggregate these results and make a single decision for each specific user. For this, the 98^{th} percentile of the set of per post probabilities (of depression) is calculated, considering all the user's posts till that cut-off point. The 98^{th} percentile is used instead of another strategy, such as the mean, to give more weight to the posts with greater probability. This probability is taken as the candidate decision for that user.

For decision definition, considering the early detection task, a probability threshold (τ) was used. This threshold determines whether the model has enough confidence to make the subject's prediction as positive (depressive) or should wait for more data, that is, more posts. Comprehensive experiments were conducted with thresholds in the range of [0.5-1] with step 0.05. We selected the results that achieved the best effectiveness for each text embedding model. Notice that the decision that a user is non-depressive is taken only after analyzing the last block of posts. Considering the predictions provided by the classifier, the evaluation measures were computed according to the ground-truth available with the dataset.

4. Results and discussion

The experiments were carried out using multiple decision-making thresholds. However, for the sake of space and simplicity, only the configurations that allowed the best effectiveness for each of the cases described in Section 3.2. The results are separate into two sections, for no preprocessing (Section 4.1) and for preprocessing and combinations (Section 4.2).

4.1. No Preprocessing - Baseline Results

Table 2 presents the results achieved for the approach where the database was not submitted to specific preprocessing steps, except for the transformation of words to lowercase, due to the inherent characteristics of the embedding models used. Considering the multiple evaluation measures, it is clear that the results without using preprocessing are not promising in terms of effectiveness. This corroborates what is found in the specialized literature, which indicates the importance of the data preparation stage, including text preprocessing. These baseline results are important to assess the impact of different preprocessing techniques and their combinations.

4.2. Preprocessing

Tables 3, 4, 5 show the best results achieved for each preprocessing case (described in Section 3.2), considering, respectively, F_1 a *Precision* (*P*) and $ERDE_{50}$ as the selection criteria. The results are presented both for the experiments with the mapping of emoticons (on the left) and without (on the right). In the tables, the "Basic" preprocessing corresponds to Case 1.

Table 2. Effectiveness results achieved without text preprocessing.

Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R
FastText Crawl	0.6	13.2717	9.3414	0.5667	0.5000	0.6538
FastText WN	0.8	12.9135	8.8304	0.5872	0.5614	0.6154
GloVe Crawl	0.5	13.1309	9.7109	0.5586	0.5254	0.5962
Glove WN	0.5	13,0503	10.1119	0.5263	0.5814	0.4808

In the results presented in Table 3, a particular characteristic emerges. The results suggest that when applying a more robust preprocessing, e.g., *Case 6*, the mapping of emoticons in general allows better results in terms of effectiveness. On the other hand, when using basic standard preprocessing, such as *Case 1*, the mapping the emoticons does not significantly influence the results.

Table 3. Best results based on F_1 as the selection criteria.

Preprocessing	Category	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R
Basic	Case 1	FastText Crawl	0.6	12.8767	8.5415	0.6549	0.6066	0.7115	FastText Crawl	0.7	12.5323	9.4425	**0.6667**	0.7273	0.6154
Stop	Case 2	GloVe Crawl	0.65	12.7405	8.9057	0.6408	0.6471	0.6346	Glove WN	0.75	12.5464	8.9033	**0.6526**	0.7209	0.5962
Lemma	Case 3	FastText WN	0.55	12.5578	9.6927	0.6667	0.7021	0.6346	GloVe Crawl	0.55	12.5443	9.5678	**0.6809**	0.7619	0.6154
Lemma + Pos	Case 4	FastText Crawl	0.65	12.6082	9.0734	**0.6400**	0.6667	0.6154	GloVe Crawl	0.5	12.8707	8.4864	0.6218	0.5522	0.7115
Stop + Lemma	Case 5	FastText Crawl	0.5	12.7643	8.9386	**0.6667**	0.6429	0.6923	Glove WN	0.55	12.4407	9.3656	0.6526	0.7209	0.5962
Stop+ Lemma + Pos	Case 6	FastText WN	0.6	12.6374	9.4152	**0.6735**	0.7174	0.6346	FastText WN	0.55	12.7808	8.4709	0.6333	0.5588	0.7308

Considering the results in Table 4, which considers *Precision* as a selection criterion, the systems that used the *emoticons* mapping presented results similar to those described in Table 3, that is, allowed greater effectiveness with the application of the most complete preprocessing. This shows that applying a more complete preprocessing in combination with the mapping of emoticons made it possible to more effectively identify those people with depression, when compared to similar cases that did not apply the mapping.

Table 4. Best results based *Precision* (*P*) as the selection criteria.

Preprocessing	Category	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	P	R
Básico	Case 1	FastText WN	0.8	12.6460	8.9763	0.6170	0.6905	0.5577	FastText Crawl	0.7	12.5323	9.4425	0.6667	**0.7273**	0.6154
Stop	Case 2	Glove WN	0.55	12.5788	9.1833	0.6170	0.6905	0.5577	Glove WN	0.75	12.5464	8.9033	0.6526	**0.7209**	0.5962
Lemma	Case 3	FastText Crawl	0.5	12.3998	9.5183	0.6292	0.7568	0.5385	Glove WN	0.55	12.5443	9.5678	0.6809	**0.7619**	0.6154
Lemma + Pos	Case 4	FastText Crawl	0.65	12.7359	9.8833	0.6237	**0.7073**	0.5577	Glove WN	0.5	12.5857	9.9614	0.6186	0.6667	0.5769
Stop + Lemma	Case 5	GloVe Crawl	0.75	12.5874	9.5046	0.6207	**0.7714**	0.5192	GloVe Crawl	0.55	12.4407	9.3656	0.6526	0.7209	0.5962
Stop + Lemma + Pos	Case 6	FastText WN	0.6	12.6374	9.4152	0.6735	**0.7174**	0.6346	GloVe Crawl	0.5	12.5405	8.8485	0.6275	0.6400	0.6154

Regarding the best results in terms of $ERDE_{50}$, Table 5 shows that the preprocessing without the mapping of emoticons allowed the best results. However, there are other important aspects to consider. By analyzing the *Precision* of these systems, one can notice that the systems that applied the mapping, in general, obtained much better results in most cases. This situation suggests a demand for application-oriented trade-off adjustment given systems that did not use the mapping have a lower early detection error, but on the other hand make more mistakes in the detection of depressive individuals. In addition, the results suggest that using the mapping allowed to maintain a better balance between the two measures (P and $ERDE_{50}$).

These findings suggest that using the mapping is possibly more appropriate, because in addition to more effectively identifying people with depression, it still maintains competitive results in terms of $ERDE_{50}$. Nevertheless, depending on the objective of the system, it is important to assess whether it is desired to detect more risk cases, to detect

risk cases earlier by looking at less data, or to find a model that keeps a balance between these two measures.

Table 5. Best results based on $ERDE_{50}$ as the selection criteria.

Preprocessing	Category	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	*P*	*R*	Model	τ	$ERDE_5$	$ERDE_{50}$	F_1	*P*	*R*
Basic	Case 1	FastText Crawl	0.6	12.8767	**8.5415**	0.6549	0.6066	0.7115	GloVe Crawl	0.7	12.5464	8.9378	0.6237	**0.7073**	0.5577
Stop	Case 2	FastText WN	0.7	12.8812	8.8503	0.6055	0.5789	0.6346	Glove WN	0.7	12.7842	**8.6911**	0.6226	**0.6111**	0.6346
Lemma	Case 3	Glove WN	0.55	12.9002	9.1537	0.6226	**0.6111**	0.6346	FastText WN	0.55	12.6812	**8.5776**	0.6549	0.6066	0.7115
Lemma + Pos	Case 4	FastText WN	0.65	12.6082	9.0734	0.6400	**0.6667**	0.6154	FastText WN	0.5	12.8707	**8.4864**	0.6218	0.5522	0.7115
Stop + Lemma	Case 5	FastText Crawl	0.5	12.7643	8.9386	0.6667	**0.6429**	0.6923	GloVe Crawl	0.55	12.9436	**8.9198**	0.6154	0.5538	0.6923
Stop + Lemma + Pos	Case 6	Glove WN	0.55	12.4261	9.0308	0.6535	**0.6735**	0.6346	GloVe Crawl	0.55	12.7808	**8.4709**	0.6333	0.5588	0.7308

Still regarding the findings from Table 5, when it comes to depression, it is important to consider how critical it is to detect true positives. Hence, although late detection is a problem, detecting positive cases is crucial, as they require intervention, in order to reduce emergence of serious risks, including suicide attempts. Therefore, it highlights the importance of the results obtained with the mapping of emotions, given in addition to enabling the detection of true cases, on a highly unbalanced dataset, it allowed to maintain a certain balance between different measures when compared to not using the mapping.

5. Conclusion

Depression is one of the most prevalent public health problems worldwide due to the numerous consequences that may arise from it, such as suicide attempts. In this sense, finding solutions to assist in the identification, prevention or treatment is an essential task. Due to multiple factors, such as the strong social stigma, many people who suffer from depression rely on less formal environments, such as social media, to talk about their condition and a find some relief. Thus, the textual data generated by these users can be exploited in the development of tools related to depression identification. However, using such data without adequate preparation may lead to unsatisfactory effectiveness. Hence, in this work, we performed a comparative analysis of different data preprocessing techniques to verify their contribution to prediction effectiveness.

The experimental results suggests not applying the preprocessing achieves inferior results. This highlights the importance of the data preparation stage for machine learning model construction. The results achieved showed that using the mapping of *emoticons* can generate better representations and that allow effectiveness improvement, especially when associated with more robust preprocessing techniques, such as *Stop Words Removal*, *Lemmatization* and *Pos Tagging*. On the other hand, it was observed that, without the mapping, a simple preprocessing is sufficient to achieve good results. In addition, it was found that the application of the mapping along with more complete preprocessing techniques allowed a better balance between the different measures compared, especially for *Precision* and $ERDE_{50}$.

As future work, we intend to evaluate these preprocessing approaches in other databases, especially with a larger amounts of data. It is also intended to investigate other preprocessing techniques, as a way to enable the construction of even better predictive models. Finally, with the increase of identification reliability of machine learning solutions, recommendations for intervention could be issued, for instance, through advertisements, informational links and online advisement.

References

Aggarwal, C. C. (2011). *An Introduction to Social Network Data Analytics*, pages 1–15. Springer US, Boston, MA.

Almeida, H., Briand, A., and Meurs, M. (2017). Detecting early risk of depression from social media user-generated content. In *Working Notes of CLEF 2017, Dublin, Ireland, September 11-14, 2017.*

Benamara, F., Moriceau, V., Mothe, J., Ramiandrisoa, F., and He, Z. (2018). Automatic detection of depressive users in social media. In *CORIA 2018, 15th French Information Retrieval Conference, Rennes, France, May 16-18, 2018. Proceedings.*

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., and Bierut, L. J. (2016). A content analysis of depression-related tweets. *Computers in Human Behavior*, 54:351–357.

Choudhury, M. D., Counts, S., and Horvitz, E. (2013a). Predicting postpartum changes in emotion and behavior via social media. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, pages 3267–3276.

Choudhury, M. D., Counts, S., and Horvitz, E. (2013b). Social media as a measurement tool of depression in populations. In *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*, pages 47–56.

Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013c). Predicting depression via social media. In *Proceedings of ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Goodwin, F. K. and Jamison, K. R. (1990). *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford University Press, New York.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 28–39.

Losada, D. E., Crestani, F., and Parapar, J. (2017). CLEF 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *Working Notes of CLEF 2017, Dublin, Ireland, September 11-14, 2017.*

Malam, I. A., Arziki, M., Bellazrak, M. N., Benamara, F., Kaidi, A. E., Es-Saghir, B., He, Z., Housni, M., Moriceau, V., Mothe, J., and Ramiandrisoa, F. (2017). IRIT at e-risk. In *Working Notes of CLEF 2017, Dublin, Ireland, September 11-14, 2017.*

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Nadeem, M. (2016). Identifying depression on twitter. *CoRR*, abs/1607.07384.

Nakamura, T., Kubo, K., Usuda, Y., and Aramaki, E. (2014). Defining patients with depressive disorder by using textual information. In *2014 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 24-26, 2014*.

Organization, W. H. (2017). Depression and other common mental disorders: Global health estimates.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Richards, C. S. and O'Hara, M. W., editors (2014). *The Oxford Handbook of Depression and Comorbidity*, volume 1. Oxford University Press.

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Santana, R. C., de Lima, T. H. N., Pinto, S. A. P., Zárate, L. E., and Nobre, C. N. (2018). Otimização automática de classificadores para auxiliar no diagnóstico da depress ão. In *Anais do XVIII Simpósio Brasileiro de Computação Aplicada à Saúde*, Porto Alegre, RS, Brasil. SBC.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., and Gloor, P. A. (2013). The power of prediction with social media. *Internet Research*, 23(5):528–543.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2020). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. Knowl. Data Eng.*, 32(3):588–601.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. (2015). Recognizing depression from twitter activity. In *Proceedings of the CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 3187–3196.

Vedula, N. and Parthasarathy, S. (2017). Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health, London, United Kingdom, July 2-5, 2017*, pages 127–136.

Wang, X., Zhang, C., and Sun, L. (2013). An improved model for depression detection in micro-blog social network. In *13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7-10, 2013*, pages 80–87.

Yang, C. and Srinivasan, P. (2016). Life satisfaction and the pursuit of happiness on twitter. *PloS one*, 11(3).

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2968–2978.