 DOI 10.1109/REW53955.2021.00060

Applying Transfer Learning to Sentiment Analysis in Social Media

Ariadna de Arriba
Research intern
Universitat Politècnica de Catalunya
Barcelona, Spain
ariadna.de.arriba@upc.edu

Marc Oriol
GESSI Research Group
Universitat Politècnica de Catalunya
Barcelona, Spain
marc.oriol@upc.edu

Xavier Franch
GESSI Research Group
Universitat Politècnica de Catalunya
Barcelona, Spain
franch@essi.upc.edu

***Abstract* — Context: Sentiment analysis is an NLP technique that can be used to automatically obtain the sentiment of a crowd of end-users regarding a software application. However, applying sentiment analysis is a difficult task, especially considering the need of obtaining enough good quality data for training a Machine Learning (ML) model. To address this challenge, transfer learning can help us save time and get better performance results with a limited amount of data. Objective: In this paper, we aim at identifying to which degree transfer learning improves the results of sentiment analysis of messages shared by end-users in social media. Method: We propose a tool-supported framework able to monitor and analyze the sentiment of tweets with different ML models and settings. Using the proposed framework, we apply transfer learning and conduct a set of experiments with multiple datasets. Results: The performance of different ML models with transfer learning from different datasets are obtained and discussed, showing how different factors affect the results, and discussing how they have to be considered when applying transfer learning.**

Keywords — sentiment analysis, transfer learning, natural language processing, machine learning, social media analysis

I. INTRODUCTION

With the advent of social media, crowds of end-users can provide vast amounts of feedback through app reviews, forums and social networks. This feedback is a relevant source of information for requirement engineers to understand end-user needs and their satisfaction regarding existing features, missing functionalities or the overall quality of experience of a software application. To facilitate the analysis of this feedback, several Natural Language Processing (NLP) techniques exist. One of these techniques is Sentiment Analysis, which provides the automatic classification of the sentiment that a message evokes [1]. Sentiment analysis can be classified based on the approach used to conduct the analysis: machine learning (ML), lexicon-based, hybrid (i.e., ML & lexicon-based) and graph-based [2]. Currently, the most commonly used technique for automatic sentiment analysis is ML. ML-based sentiment analysis fits in the scope of supervised ML, which requires manual labelling of the sentiment of the messages to train the ML models. Labelling is costly and there might be insufficient training data to provide reliable results. To overcome this issue, techniques like transfer learning have been proposed. Transfer learning is an ML technique used when a model trained for a specific task is reused for another related task [3]. Among its advantages, it has been proven to be very useful for saving time and improving the performance of an ML model [4]. However, transfer learning can be counterproductive and cause some unexpected issues. For instance, a common problem in ML is that providing a big quantity of data can cause a ML model to learn too many irrelevant details that may, ultimately, have a negative impact on the results, which is a common issue known as overfitting [5]. This is because the model may focus on features that may not be relevant for the aim, and only produces noise on the data provided. Another common problem that can arise is negative data transfer, which means that the ML model decreases its performance when applying transfer learning, instead of getting better results [6]. Negative data transfer may be caused by several issues. For example, having the origin and target datasets from very different contexts, or using comparatively unbalanced datasets.

Apart from these issues, we should emphasize that transfer learning and ML processes, in general, are costly in time and resources. These considerations altogether drive us to the following research question:

RQ: To which degree applying transfer learning improves the results of sentiment analysis?

To answer this research question, we have developed a framework able to monitor and analyze the sentiment of tweets using several ML models in a multilingual setting. We have used this framework to apply transfer learning using several corpora from different contexts, sources, and number of samples written in English. As a use case, given the current advent and popularity of several Covid-19 related applications and software services, we used the trained ML models to assess the sentiment of tweets written in Spanish related to Covid-19.

Our contribution to the CrowdRE community is the following:

C1: A versatile and publicly available framework capable of monitoring social media messages and performing sentiment analysis through different ML models. The CrowdRE community may benefit from this framework to support the requirements elicitation process by identifying the sentiment of end-users regarding specific software features or characteristics.

C2: An initial experiment used in the context of Covid-19 as a use case, that provides insights on how transfer learning can be applied for sentiment analysis. The results demonstrate

that transfer learning can be applied for beyond-polarity sentiment analysis, which ultimately saves researchers time and resources to build the required ML models.

The rest of the paper is structured as follows: Section II presents the background and related work to sentiment analysis and transfer learning; Section III shows the architecture of the framework developed and Section IV describes the steps followed to carry out this experiment. At the end of this paper, Section V reflects the results obtained and we end the paper with a discussion and threats to validity, which belong to Section VI and VII, respectively. A conclusions section is added to close the document.

II. Background And Related Work

A. Sentiment analysis

Sentiment analysis determines the emotional tendency of a piece of text applying NLP and ML techniques. We can distinguish two very clear types of sentiment analysis based on the scope of their results: most of the existing approaches provide just basic sentiment analysis in the positive-negative spectrum (known as polarity classification) [7]. More advanced sentiment analysis techniques provide the results in terms of emotions (e.g., happiness, fear, sadness, anger), which is also known as beyond-polarity sentiment analysis [8][9][10].

The polarity sentiment analysis consists of classifying a text according to its polarity. The simplest polarity that exists is positive/negative, but often, the neutral sentiment is also added. To classify a text into a polarity, rule-based methods are frequently used, which involves a basic routine of NLP using an extensive list of words usually classified into positive or negative. The algorithm analyzes the content of the text and searches into the list looking for terms that match those in the text [11]. Then, it calculates the frequency of positive and negative words using some of the existing techniques to quantify words such as Term Frequency - Inverse Document Frequency (TF-IDF) [12].

Emotion classification is a more complex but also more accurate procedure. There are different methods depending on the number of emotions you want to get from the classifier. The most common emotions, identified as basic emotions according to Paul Ekman [13], are: sadness, fear, happiness (or joy), anger, surprise and disgust. The neutral feeling is typically also added to texts that do not show any of the aforementioned emotions when applying sentiment analysis.

B. Transfer learning

Transfer learning consists in storing knowledge obtained in one specific task to resolve another related task. In ML, transfer learning has been evolving and increasing the number of research studies in the last few years. It has been gaining ground in opinion mining and sentiment analysis, being the main focus in several research studies in that field.

In deep learning, knowledge is transferred from one model to another. According to Tan et al. [14], there are four categories to classify deep transfer learning: instances-based deep transfer learning, mapping-based deep transfer learning, network-based deep transfer learning, and adversarial-based deep transfer learning.

For our purpose, we will apply network-based deep transfer learning which refers to the reuse of the partial network that has been pre-trained in the source domain, including its network structure and connection parameters, and transfer it to be a part of deep neural network which is used in the target domain [15]. In summary, we will train the source dataset with a group of emotions and then we will use the pre-trained network to test the target dataset and see how accurate and good the model is with different corpora focused on different topics.

III. Proposed Framework

To conduct the empirical study to evaluate the performance of transfer learning, we propose an architecture that enables the classification of emotions from social media in real time, using transfer learning in a multilingual setting.

Fig. 1 shows the architecture of our proposed framework. The central component of the architecture is the *Orchestrator*, which is responsible for maintaining the flow of information between all subsystems of the framework: The *Twitter Monitor*, the *Tweets Preprocessing* (a REST API to preprocess tweets) and the *Sentiment Analysis* REST API[1]. This last API is connected to *Microsoft* API, for tasks of translating, and to several ML tools, including the developed ML models and external tools such as ParallelDots API.

The behavior is as follows: The *Twitter Monitor* receives the tweets in real-time by calling the Twitter API[2] and applying the desired filters (e.g., collecting tweets from a specific language, applying some keywords, or filtering out retweets and replies in order to get original tweets only). Then, the *Twitter Monitor* sends the tweets to the *Apache Kafka* server, which implements the observer pattern, and forwards the collected tweets to the *Orchestrator. The Orchestrator* calls the *Tweets Preprocessing* API to apply the preprocessing stage in every tweet received. In the current stage of implementation, the *Tweets Preprocessing* component applies some common techniques in NLP, such as checking that a word exists in the specified dictionary (which can be Spanish or any other language), removing unnecessary URLs and mentions, lemmatizing the sentence or replacing emojis with the emotion that they express; all of this in order to obtain a "cleaner" text to be more understandable for the machine.

Once all tweets have been obtained and preprocessed, there are two possible paths to follow. The first one is that the *Orchestrator* generates a CSV file designed to train our ML models, whereas the other one is to apply sentiment analysis using the trained model. In both cases, the system checks if the tweets are in the specified language and, if not, it calls the translator endpoint of *Sentiment Analysis* API, a generic RESTful API, to translate the text. The *Sentiment Analysis* API provides a common interface for applying sentiment analysis and translating short texts, and connects to third-party services to conduct the translation. At the current stage of implementation, this API connects to the Microsoft translation API, but other translation tools could be added.

[1] gessi-sw.essi.upc.edu:8080/sentiment-analysis-api/swagger-ui.html

[2] https://developer.twitter.com/en/docs/twitter-api/early-access

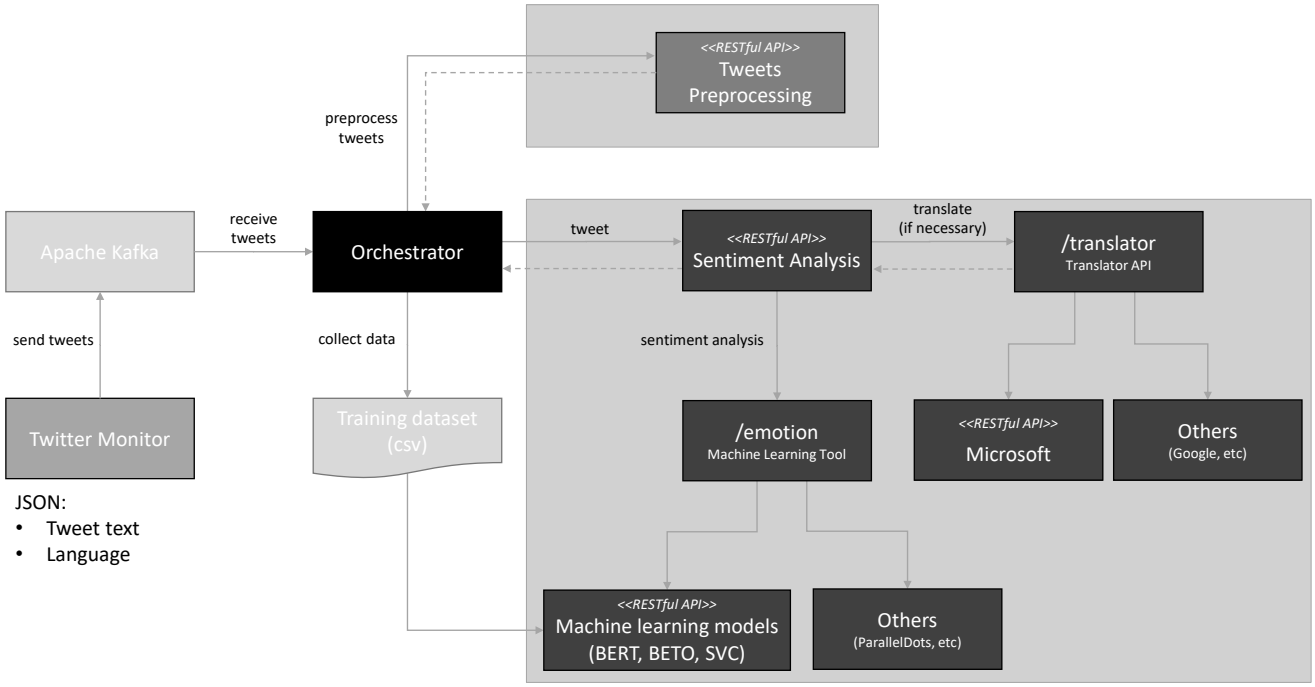

Fig. 1. System architecture

The last stage, in the second option, is to call the *Sentiment Analysis* API to apply sentiment analysis to tweets and obtain the probability of each emotion. In this case, and similarly to the translator endpoint, this API calls another API to classify the text into emotions. We can either call an existing public Sentiment Analysis API or the APIs we have created with our ML models developed. Currently we have implemented the following ML models: BERT Multilingual, BETO and SVM. The BETO model is part of our previous work [14] and has been integrated into the proposed architecture, whereas the BERT Multilingual and SVM have been implemented as a result of this work. The source code of the framework is publicly available in Github[3].

IV. Protocol Of The Experiment

In this section, we explain the protocol used to prepare the experiment and the decisions made to carry it out. Taking advantage of the system designed and explained in the previous section, the procedure we follow to apply the transfer learning tasks is the following:

- **Getting datasets.** Our own dataset has been obtained by monitoring tweets using the proposed framework explained in the previous section and the others were obtained in different external sources as we will explain below.
- **Training models.** We have trained each ML model (i.e., BERT Multilingual, BETO and SVM) with different datasets and fine-tuning some hyperparameters. We have split an 80% of each dataset to train the model and a 20% to evaluate it.
- **Testing models.** We have tested the models in two stages. First, we have validated the model trained with the 20% left of the same origin dataset used for training. Then, we have evaluated it with the complete target dataset on each corpus.

We describe below the datasets chosen for the study and the modifications done to adjust them to the experiment. Then, we brief the ML models and techniques used to conduct the analysis, and finally, we sum up this part describing the measures applied to evaluate the experiment performance.

A. Selection of datasets

We distinguished two relevant groups of datasets to carry out our experiment. We picked out three datasets from external sources tagged with different emotions - diverse emotions in each dataset - and placed them into the origin datasets collection. In the target datasets group, we took our own corpus and labelled it with five distinct emotions.

1) Origin datasets

We selected three datasets in English, due to the fact that most Spanish corpus found are poor or ineffective, and we translated them to Spanish using the *Translator* API to train our ML models. All of the datasets contain short texts extracted from Twitter except one that contains a collection of documents. Below we describe each dataset to understand better our experiment and decisions made. Table I summarizes the number of entries and data distribution for each origin dataset.

The "Emotion Dataset for NLP" [16] is a collection of documents with an emotion flag which contains 20000 entries labelled with one of the following six emotions: 'sad', 'angry',

[3] https://github.com/twittersentimentanalysis

'joy', 'surprise', 'fear', 'love'. To adapt it to our target dataset, we first linked similar emotions, in this case, we only replaced 'joy' with 'happy', and removed emotions that we did not consider in our target, i.e., 'fear' and 'love'.

The second selected corpus, "SMILE Twitter Emotion Dataset" is a collection of tweets related to the British Museum and created for the purpose of classifying emotions expressed on Twitter towards arts and cultural experiences in museums [17]. The original dataset contains 3085 tweets tagged with 5 emotions: 'anger', 'disgust', 'happiness', 'surprise' and 'sadness', plus 'not-relevant'. For our goal, we have slightly changed some names and removed the 'disgust' emotion since it is not included in our target. After these changes and after applying a preprocessing stage and removing the empty tweets, the dataset obtained contains 1238 entries.

Finally, the "Twitter Reviews Dataset" consists of 10017 Twitter user reviews with their emotion labelled [18]. The starting collection is classified into 6 emotions ('happy', 'sad', 'surprise', 'fear', 'disgust', 'angry') but again, for our purpose, we have removed 'fear' and 'disgust' tags.

TABLE I. DATA DISTRIBUTION FOR ORIGIN DATASETS

Emotion	Number of samples		
	Emotion Dataset for NLP	***SMILE Twitter Emotion Dataset***	***Twitter Reviews Dataset***
angry	2709	55	1316
happy	6761	1116	3725
sad	5797	32	2658
surprise	719	35	315

2) Target datasets

Given the advent of multiple Covid-19 related apps in the current pandemic, we decided to use as target dataset a collection of tweets related to Covid-19 as use case, which we name hereafter Covid-19 Twitter Monitor Dataset [19]. To build this dataset, we collected an initial amount of 3346 short texts extracted from Twitter via Twitter API by applying several keywords to obtain Covid-19 related tweets[4] and a language filter to get those written in Spanish. Afterwards, we labelled manually with the following emotions: 'angry', 'happy', 'sad', 'surprise' and 'not-relevant'.

The task of labelling was an exhaustive process done in pairs in which two of the researchers tagged independently each tweet and then, we measured the inter-rater agreement to see the rate of concordance between us. We did three pilot iterations in which we discussed the differences and tried to come to an agreement to label the tweets. In the third iteration, we got an interrater agreement of 74.2% using Cohen's Kappa coefficient, which indicates that the agreement reached is substantially enough [20]. We held a weekly meeting to revise an amount of approximately 200 tweets and see which ones we agreed on and came to an agreement on the ones we differ. After these iterations, we decided to keep only the tweets we initially agreed upon. In the initial process, we had the intention of including 'disgust' emotion but, in the following iterations, during the tagging process, we realized that only one tweet was classified in that way, so we decided to exclude that tag. As we have seen in the previous section, for the origin dataset we focused on only four emotions, thus we had to remove entries tagged as 'not-relevant' and delete null tweets obtained from the preprocessing stage just as the tweets do not match in the emotion tag. We finally obtained a collection of 1359 entries following the distribution reflected in Fig. 2.

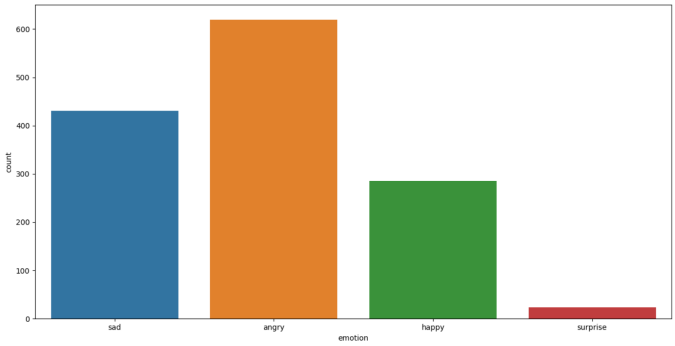

Fig. 2. Data distribution for target dataset

B. Selection of Machine Learning Models

We compared the different training datasets using the following ML models: BERT Multilingual, BETO and SVM.

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based ML technique developed by Google. BERT is a bidirectional transformer [21] pre-trained using a combination of masked language modelling objective and next sentence prediction (NSP) on a large corpus comprising the Toronto Book Corpus and Wikipedia [22]. The model acts in a way in the training process in which it has to predict the words previously masked (a 15% of each sequence) based on the context and uses NSP to predict if a pair of sentences are connected to each other [23].

We use two different variants of the BERT model. The first one chosen for the training process was BERT Multilingual, which had been trained in 104 languages, including Spanish, with the largest Wikipedia [23]. This variant was pre-trained on a large corpus of raw texts only in multiple languages. The second selected variant was BETO, a Spanish BERT model trained with a corpus in a similar size as the BERT model but exclusively in Spanish.

SVM (Support Vector Machines) are a set of supervised learning models with associated learning algorithms focused on different learning tasks such as classification as in this case. In order to solve classification problems, we used SVC [24] which is based on LIBSVM [25], a library for SVM tasks available for many programming languages. We applied the typical use of SVM to deal with multi-class classification which involved first training the corpus to obtain the model and then evaluating it predicting test data from the pre-trained model [26].

[4] The list of keywords related to Covid-19 is provided by Twitter developers, and can be obtained at: https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules

V. EXPERIMENT AND RESULTS

To evaluate the performance of every ML model chosen, the measures used are accuracy, F_1 Score, precision and recall. We trained every dataset by fine-tuning some parameters such as 'learning rate' and 'epsilon' of the optimizer, or the number of 'epochs' in the training process to obtain the best model results. Then we tested it with the other datasets in all the ML models developed.

From all models developed, SVC is the one that gives the worst results as opposed to BERT and BETO (see Table II). Whereas BERT and BETO are quite similar in terms of performance in "Emotion dataset for NLP" and "SMILE Twitter Emotion Dataset", there are some differences for the remaining corpora. BERT is better for "Twitter Reviews Dataset" and BETO for our target dataset. That may be this way because although we translated all datasets, all of them except ours were originally in English.

Regarding the application of transfer learning, the accuracy obtained when it was trained with the origin dataset "Emotion Dataset for NLP" and tested with "Twitter Reviews Dataset", is considerably high (0.62) as we can observe in Table III. "SMILE Twitter Emotion Dataset" presented a higher accuracy (0.73). However, the difference with the F_1 Score makes us suspicious that unbalanced data was causing this higher accuracy. In the confusion matrices [18], we can confirm this hypothesis. For our corpus, "Covid19 Twitter Monitor Dataset", the accuracy is higher without applying transfer learning. The application of this technique causes the dataset to reduce its accuracy from 0.76 to 0.43, more than 30%. We are facing a clear case of negative transfer learning, which probably mainly fails due to the difference of context and the original language of the target dataset.

In Table III, we present the metrics obtained when testing our ML model with the "Emotion Dataset for NLP". The first row corresponds to the results of testing the ML model using 20% of the origin dataset (80% of the origin dataset was used for training). The other rows present the results of testing the ML model using 100% of their respective datasets. As we can observe, the accuracy is high in most cases. However, seeing the confusions matrices we can stipulate that in the corpora that performs better is "Twitter Reviews Dataset", as we mentioned above.

TABLE II. PERFORMANCE RESULTS FOR DIFFERENT MACHINE LEARNING MODELS AND CORPORA

Model	Dataset	*Precision*	*Recall*	*F_1 Score*	*Accuracy*
BETO	*SMILE Twitter Emotion Dataset*	0.55939	0.24256	0.24563	0.60484
	Covid19 Twitter Monitor Dataset	0.81107	0.57147	0.56550	0.76471
BERT	*Emotion Dataset for NLP*	0.85320	0.83086	0.84095	0.86036
	Twitter Reviews Dataset	0.85461	0.77967	0.81016	0.83593

TABLE III. PERFORMANCE RESULTS AFTER APPLYING TRANSFER LEARNING ON "EMOTION DATASET FOR NLP" WITH BERT

Dataset	*Precision*	*Recall*	*F_1 Score*	*Accuracy*
Emotion Dataset for NLP	0.86260	0.82091	0.83933	0.86193
Covid19 Twitter Monitor Dataset	0.41932	0.38313	0.32168	0.42531
Twitter Reviews Dataset	0.50217	0.47467	0.47915	0.62104
SMILE Twitter Emotion Dataset	0.35906	0.38171	0.31327	0.73667

VI. DISCUSSION

Once we have finished transfer learning experiments with different datasets and models, we can extract some points we have observed during the process.

The first one is that obtaining a good dataset is a slow and difficult process. In our case, as we explained in section IV before, every week we labelled a collection of 200 tweets approximately for finally obtaining a corpus with a considerable number of entries which took us several weeks in the end. Another fact to consider is that we are not able to choose humans' opinion and reaction about a concrete topic, so what we got is an unbalanced collection due to people being mostly angry and sad in front of the situation of Covid-19, which is totally expected but unfavorable for this study.

The next one is that not all datasets are good enough in terms of the quality of the data (i.e., the content of the messages), balanced distribution of samples, quantity of samples and context similarity, which is a consequence of the first point. This is a well-known factor in the NLP domain and mentioned often as one the main challenges to overcome in the area [27]. This factor affects transfer learning results directly. If we have a balanced dataset with a big amount of data and the model is good enough, we will obtain optimal results in the training and validation process. However, if the other datasets have a topic completely different from the trained dataset, even if it may have good training results, the scores obtained when applying transfer learning can be far from satisfactory. This is

why it is important to have a big dataset to make the training process more accurate and to give models more vocabulary to distinguish better between two emotions.

Another point, and very close to the previous one, is that if the training dataset is poor at the beginning, training and validation results will have the same bad behavior. As a consequence of bad results, testing a bad model with other corpora will result in negative transfer learning (i.e., worsening the accuracy of the trained model).

From the results obtained, we can conclude that transfer learning improves the results of sentiment analysis provided that the input dataset is similar to the output dataset. Even though after applying transfer learning the accuracy is not higher in all cases, it is always very similar except in our target dataset. The facts that produce this worse result are: (1) the three origin datasets follow almost the same distribution, (2) all corpora were originally in English, and (3) the context is not as specific and limited as our use case.

VII. Threats to Validity

In this section we discuss the threats to validity of our evaluation and the actions we have taken to mitigate them.

- *Internal validity:* The internal validity of the evaluation concerns our ability to draw conclusions from the conducted experiments and the outcomes observed. To mitigate such a threat, we applied several ML techniques into a dataset, we extracted tweets related to Covid-19 picked randomly every week and labelled them following a rigorous process in pairs.
- *Construct validity:* Construct validity corresponds to how well are our performance metrics to test our models. Regarding this point, we measured our ML models with common metrics to test, not only how good is the model but how good are our datasets. Furthermore, we provide the confusion matrix and training and validation loss to check the best model for our target corpus.
- *Conclusion validity:* Conclusion validity sets how reasonable are the conclusions research reached. The main threat to deal with is the possibility of obtaining an incorrect conclusion, due to missing a relationship between the data and the conclusions or, on the contrary, conclude a relationship that does not exist. To mitigate these threats, we were cautious with the results and measures extracted. We observed in detail the performance metrics and confusion matrices and derived a conclusion taking into account the type of dataset and features such as the kind of data that contains, or the number of samples of each class.
- *External validity:* External validity refers to the generalizability of our conclusions. In this regard, we have used several publicly available origin datasets to mitigate any possible risk related to any possible bias caused by these datasets. However, the experiment used a single target dataset specific to messages related to Covid-19 written in Spanish. Although the chosen target dataset could be for another language and topic, further experiments are needed to assess this aspect.

VIII. Conclusions

After carrying out the experiment, we corroborated the statement that transfer learning is costly in time and resources. The importance of a good quality dataset can lead the experiment to failure or success. Therefore, the process of obtaining the corpus should be exhaustive and rigorous.

As we have seen in the results extracted, the "Emotion Dataset for NLP" has been successful above the rest of the corpora. The evidence confirms that it is very important to have a dataset with a significant number of samples, equality in number between the different classes, and containing text distinctively different on each emotion labeled to help the machine avoiding confusion.

As a future work and improvement, we plan to collect more data from Twitter, maybe from different contexts instead of only focusing on Covid-19 to widen the number of messages of every single class. Apart from this, it could be a good point to analyze every text to assure that it is a relevant message to distinguish the emotion and discard the ones that are not. Regarding transfer learning, we could apply it in datasets with other contexts and other languages.

Acknowledgements

This work has been partially supported by the Spanish project DOGO4ML (contract PID2020-117191RB-I00).

References

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: https://doi.org/10.1016/j.asej.2014.04.011.

[2] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," *ACM Comput. Surv.*, vol. 49, pp. 1–41, 2016, doi: 10.1145/2938640.

[3] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[4] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, May 2016, doi: 10.1186/s40537-016-0043-6.

[5] X. Ying, "An Overview of Overfitting and its Solutions," *J. Phys. Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[6] W. Zhang, L. Deng, and D. Wu, "Overcoming Negative Transfer: A Survey," *CoRR*, vol. abs/2009.00909, 2020, [Online]. Available: https://arxiv.org/abs/2009.00909

[7] E. Fersini, "Chapter 6 - Sentiment Analysis in Social Networks: A Machine Learning Perspective," in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Boston: Morgan Kaufmann, 2017, pp. 91–111. doi: 10.1016/B978-0-12-804412-4.00006-1.

[8] Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Ho, and J. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," 2016, pp. 1361–1364. doi: 10.1109/FTC.2016.7821783.

[9] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 439–448. doi: 10.1109/ICDM.2016.0055.

[10] B. Gaind, V. Syal, and S. Padgalwar, "Emotion Detection and Analysis on Social Media," *CoRR*, vol. abs/1901.08458, 2019, [Online]. Available: http://arxiv.org/abs/1901.08458

[11] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, USA, 2005, pp. 347–354. doi: 10.3115/1220575.1220619.

[12] Y. Li and B. Shen, "Research on sentiment analysis of microblogging based on LSA and TF-IDF," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 2584–2588. doi: 10.1109/CompComm.2017.8323002.

[13] T. Dalgleish and M. Power, *Handbook of Cognition and Emotion*. John Wiley & Sons, 2000.

[14] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, *A Survey on Deep Transfer Learning*, vol. abs/1808.01974. 2018. [Online]. Available: https://arxiv.org/abs/1808.01974

[15] A. de Arriba Serra, M. Oriol Hilari, and X. Franch, "'Applying Sentiment Analysis on Spanish Tweets Using BETO' To appear in proceedings of the Iberian Languages Evaluation Forum (IberLEF), 2021.," presented at the IberLEF.

[16] "Emotions dataset for NLP," Feb. 03, 2021. https://kaggle.com/praveengovi/emotions-dataset-for-nlp (accessed Feb. 03, 2021).

[17] "SMILE Twitter Emotion Dataset." https://kaggle.com/ashkhagan/smile-twitter-emotion-dataset

[18] "Twitter Reviews for Emotion Analysis," Feb. 10, 2021. https://kaggle.com/shainy/twitter-reviews-for-emotion-analysis (accessed Feb. 10, 2021).

[19] A. de Arriba Serra, M. Oriol, and X. Franch, *Applying transfer learning to sentiment analysis in social media - Supporting material*. Zenodo, 2021. doi: 10.5281/zenodo.5047285.

[20] K. E. Emam, "Benchmarking Kappa: Interrater Agreement in Software Process Assessments," *Empir. Softw. Eng.*, vol. 4, no. 2, pp. 113–133, Jun. 1999, doi: 10.1023/A:1009820201126.

[21] A. Vaswani *et al.*, "Attention Is All You Need," *ArXiv170603762 Cs*, Dec. 2017, Accessed: Feb. 12, 2021. [Online]. Available: http://arxiv.org/abs/1706.03762

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv181004805 Cs*, May 2019, Accessed: Feb. 12, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805

[23] R. Horev, "BERT Explained: State of the art language model for NLP," *Medium*, Nov. 17, 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (accessed Feb. 12, 2021).

[24] "SVC Documentation," Feb. 19, 2021. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC (accessed Feb. 19, 2021).

[25] C. Chih-Chung and L. Chih-Jen, "LIBSVM: A Library for Support Vector Machines," *2001*, p. 39, Jan. 2021.

[26] "LIBSVM: A Library for Support Vector Machines," Feb. 19, 2021. https://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed Feb. 19, 2021).

[27] F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, "Natural language processing for requirements engineering: The best is yet to come," *IEEE Softw.*, vol. 35, no. 5, pp. 115–119, Sep. 2018, doi: 10.1109/MS.2018.3571242.