Merging Datasets for Emotion Analysis

An Approach using BETO on Spanish Tweets

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
moriol@essi.upc.edu

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

ABSTRACT

Context. Applying sentiment analysis is in general a laborious task. Furthermore, if we add the task of getting a good quality dataset with balanced distribution and enough samples, the job becomes more complicated.

Objective. We want to find out whether merging compatible datasets improves emotion analysis based on machine learning (ML) techniques, compared to the original, individual datasets.

Method. We obtained two datasets with Covid-19-related tweets written in Spanish, and then built from them two new datasets combining the original ones with different consolidation of balance. We analyzed the results according to precision, recall, F1-score and accuracy.

Results. The results obtained show that merging two datasets can improve the performance of ML models, particularly the F1-score, when the merging process follows a strategy that optimizes the balance of the resulting dataset.

Conclusions. Merging two datasets can improve the performance of ML models for emotion analysis, whilst saving resources for labeling training data. This might be especially useful for several software engineering activities that leverage on ML-based emotion analysis techniques.

CCS CONCEPTS

• Supervised learning by classification • Natural language processing • Social networks

KEYWORDS

Sentiment Analysis, Emotion Classification, Machine Learning, Merging Datasets, Social Media, Twitter, BETO

1 Introduction

Sentiment analysis is a growing field of research [1] that can be used to support a number of software engineering (SE) activities such as requirements elicitation [2], code review [3] and app review analysis [4]. Social media, online forums and software development repositories have become major sources of information to retrieve and analyze sentiments from developers, end-users, and other stakeholders. A number of techniques have been proposed in the literature to extract sentiment from text, based on machine learning (ML), lexicon-based, graph-based or hybrid [5]. Most of the research conducted so far has mainly focused on the English language corpora, and the comparatively low research in other languages has led to different approximations to support sentiment analysis in specific languages to close such gaps [6][7]. Training these models from scratch is expensive. Furthermore, there might be insufficient data to properly train the models to obtain satisfactory results.

Leveraging the boom of social networks usage and sentiment analysis, we monitor Spanish tweets tagged with emotion with the intention to create a system capable of classifying emotions in a piece of text. As mentioned above, a popular approach for identifying emotions in a piece of text is the use of ML [8]. Despite ML having the potential to obtain more accurate results than the other methods (lexicon-based, graph-based), the results may not always meet stakeholders' expectations. Many times, missing data can result in a poor ML model. Besides, data collection can be a slow and difficult process and specifically in the emotion analysis field, because emotions are subjective and researchers tag the text based on their criteria, so this task should be done carefully and involving many people to finally reach an agreement.

In our previous work, we developed an ML model trained with Spanish corpora to obtain the emotions of twitter messages in Spanish [9]. To improve the accuracy of the results, we proposed subsequently the application of transfer learning to train the ML models with other datasets [10]. In this work, we propose a method to accelerate the process of obtaining more quantity of data by merging datasets coming from similar contexts. This objective motivates the research question addressed in this paper:

> RQ1: Do merged datasets improve ML-based emotion analysis compared to single datasets?

 DOI 10.1109/ASEW52652.2021.00051

To answer this research question, we leveraged Covid-19 and the pandemic situation to define an experiment using tweets written in Spanish as data. Therefore, the corpus obtained from external sources should be in the same context or a similar one.

The paper is structured as follows. We first contextualize the research topic by analysing the background and related work (Section 2). Next, we present the proposed architecture (Section 3) and the protocol of the experiment (Section 4). Finally, we show the results obtained (Section 5) and some discussion (Section 6). We end the study with the summary of threats to validity (Section 7) and the conclusions extracted (Section 8).

2 Background and Related Work

2.1 Data Merging

Data merging is the process of merging two or more datasets into one [11]. There are two common approaches or techniques to proceed: merging new cases or adding new variables. In the first case, both datasets contain the same labels but instances differ. Therefore, the merging technique consists, basically, in extending one dataset with samples of the other [12]. The second approach is based on creating a new dataset by merging labels of both datasets as some, or even all, features may be distinct. In this case, the instances are the same in both datasets [13].

Porting these concepts to our study context, we base our experiment upon two datasets with different instances and each one labelled with a unique emotion (datasets are described in Section 4.1). Both datasets share some tags but differ in others. Therefore, we decided to apply a combination of both approaches mentioned above, obtaining instances from both corpora and, simultaneously, adding the labels in which they differ. Previous studies show that it is possible to merge several datasets to train an ML model for sentiment analysis [14].

2.2 Sentiment analysis

Sentiment analysis is a common technique used to classify a text into the sentiments that it expresses. To carry out the approach, many natural language processing (NLP) methods and techniques are applied. There are two main types of problems related to sentiment analysis: polarity-based sentiment analysis and emotion classification. The polarity-based approach consists in classifying a piece of text into positive or negative, or neutral if text was written in an objective way and does not express any emotion [15]. Emotion classification, instead, is based on classifying a text into emotions as happiness or anger [16]. There are as many variants as emotions and combinations but most commonly used are the ones proposed by Paul Ekman [17]. For our purpose we have selected the following emotions: ‘anger’, ‘happy’, ‘sad’, ‘surprise’ and ‘not-relevant’. As the dataset extracted from external sources contains ‘fear’ emotion, we have included it in the merged datasets.

Although ML-based techniques can be used in both modalities of sentiment analysis, it is more commonly applied in emotion classification, which justifies our choice of basing our framework in ML. More precisely, classification issues, in this case emotion classification, are subject of research in the domain of supervised learning [18], one of the existing ML variants. To deal with the classification matter, we use the BETO model [19], a bidirectional transformer pre-trained on a large corpus as BERT [20], but exclusively in Spanish corpora. This transformer-based ML technique uses a combination of masked language modeling objective and next sentence prediction to predict data and solve classification problems [21].

3 Proposed Framework

To carry out the experimental study, we reuse an architecture that we already developed for conducting a related study in the context of crowd-based requirements engineering [10]. In this paper, we applied transfer learning in social media using several ML models and concluded that transfer learning may improve the results of sentiment analysis under certain conditions.

Figure 1 presents the architecture. It revolves around an *Orchestrator* software component. The Orchestrator has the mission of synchronizing the information flow among the other software components placed in the architecture. Concretely, these components are: the *Twitter Monitor* that gathers the tweets of interest; the *Tweets Preprocessing*, a REST API that applies typical preprocessing steps over the tweets; the *Sentiment* Analysis component, another REST API that is connected to the Microsoft API for translating messages from one language to another, and to the ML models and tools (including the BETO model and the ParallelDots API, among others). In the current study, we only used a part of this architecture: the *Twitter Monitor* to collect the tweets for our custom-made dataset and the *Tweets Preprocessing* API to preprocess the tweets. Finally, we used the ML API (only BETO for this case) to train our ML model.

More details of the architecture and the developed framework are described in our previous work [10].

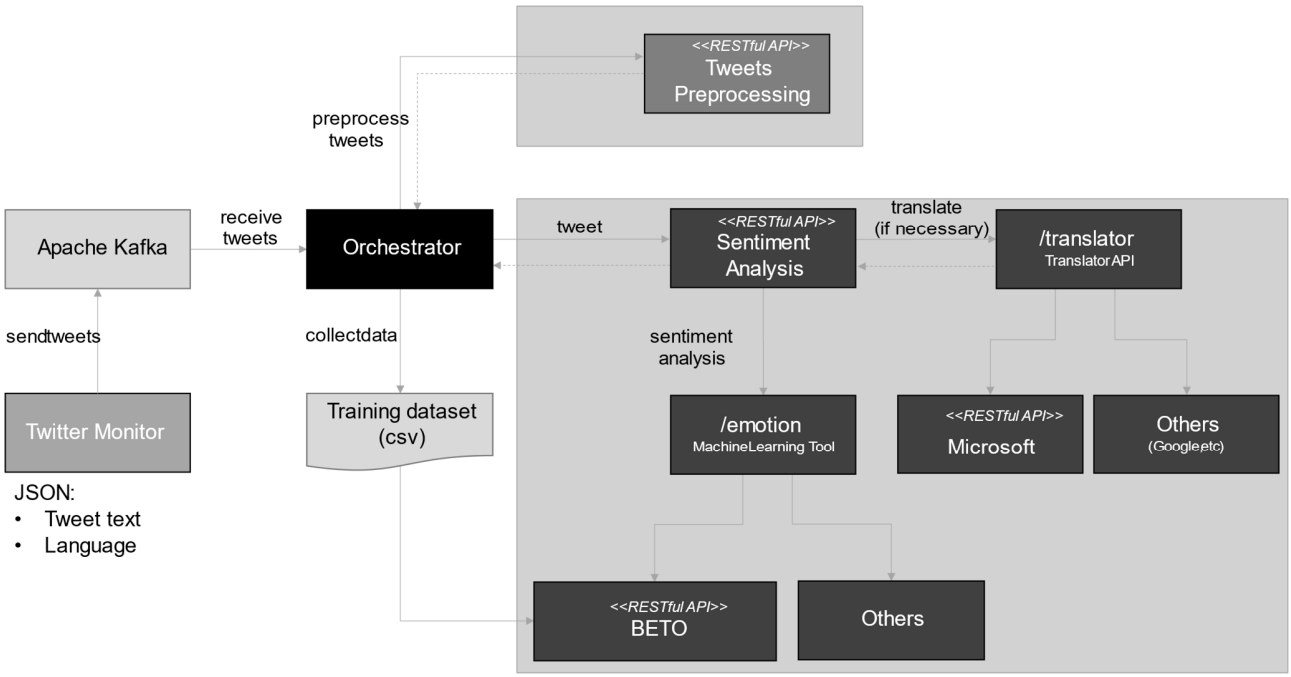

Figure 1: System architecture

4 Proposed Framework

We applied three steps to carry out the experiment as follows:

1. Obtain two datasets in Spanish.
2. Merge datasets into two other datasets according to different balancing strategies.
3. Apply preprocessing and training to the datasets.

4.1 Dataset Obtention

In this stage, we have obtained two datasets as required by our experiment, both of them in related areas to make their merging feasible. We created the first one (*Custom-made dataset*) using the architecture above, and we reused a second one (*Reused dataset*) publicly available. Figure 2 shows the distribution of each dataset (also the two ones presented in Section 4.2) explained below.

Custom-made dataset. We have created a first dataset obtaining Covid-19-related tweets from the Twitter API, applying a two-fold filter: 1) using a list of keywords provided by Twitter developers[1], 2) restricting to tweets written in Spanish only. To obtain this corpus, after a preliminary piloting phase, every week we gathered about 200 tweets and two authors tagged the tweets with the emotion they believed that fits better from the following ones: 'angry', 'happy', 'sad', 'surprise' and 'not-relevant'. After getting a total amount of 3.346 tweets, we only kept those ones in which the two assigned emotion tags matched. Applying this strict rule, we ended up with 2.165 tweets.

Reused dataset. We looked for, and found, a dataset in the Covid-19 context tagged with similar emotions as the ones we had. This corpus[2] is a collection of 3.085 tweets (after applying the preprocessing, once translated, and removing the empty tweets) written in English collected during the lockdown period in India. The labels for the emotions in the collection are: 'fear', 'sad', 'anger' and 'joy'. Therefore, before merging both datasets, we had to adapt this corpus to our tags and language. First, we translated all tweets to Spanish using *Translator API* since we wanted to train a monolingual model and, then, we replaced some tags that differed: we replaced 'joy' by 'happy' and 'anger' by 'angry'.

These datasets have been made publicly available in an online repository [22].

4.2 Dataset Merging

Taking into account from the previous step that (1) the two datasets contain the same metadata and (2) the emotions of the datasets were adapted to match each other, the process of merging the datasets was straightforward. To this aim, we followed two merging strategies that lead to two different merged datasets:

[1] https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules

[2] https://www.kaggle.com/surajkum1198/twitterdata

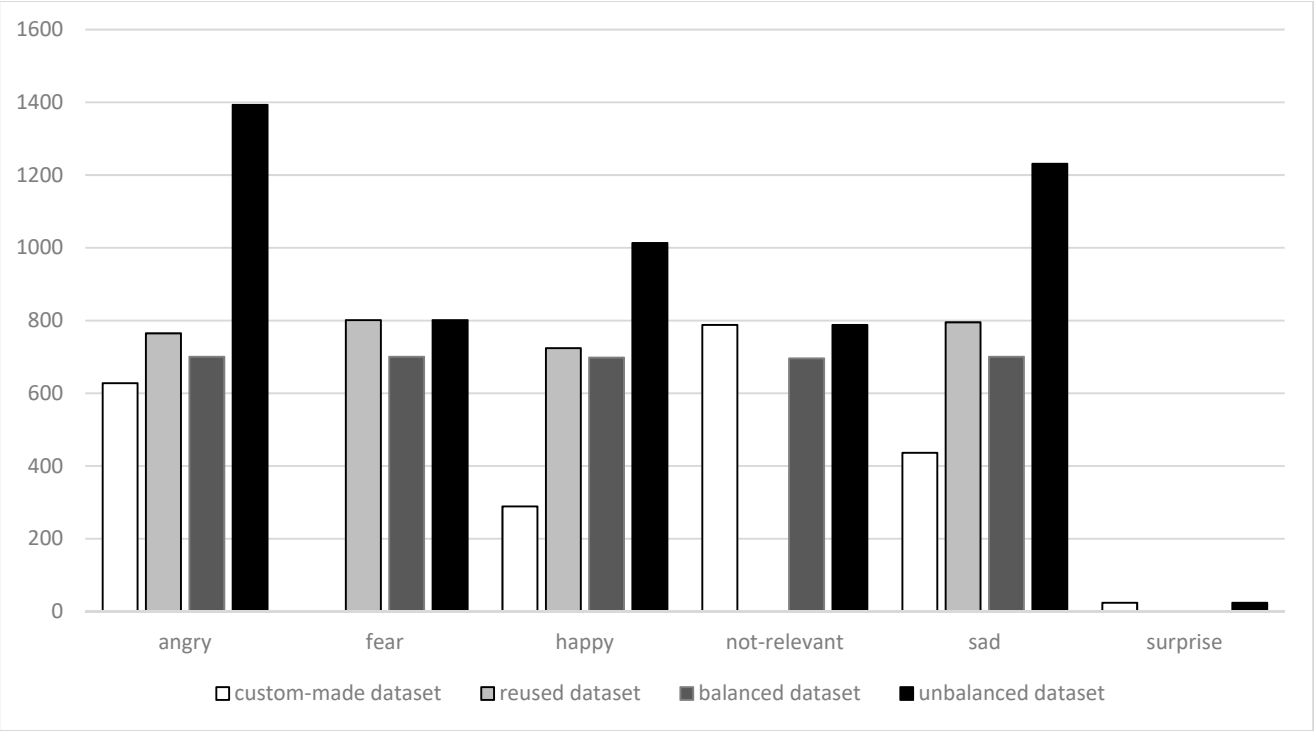

Figure 2: Datasets' distribution

Unbalanced dataset. For this corpus, we simply joined both datasets and we obtained an amount of 5.250 tweets distributed as shown in Figure 2.

Balanced dataset. We removed the 'surprise' tagged entries due to the low number of tweets with this emotion. For the other tags, we got 700 samples of each emotion as the emotion with fewer samples has 700 entries approximately. To get this number of tweets, we obtained 700 random entries of 'not-relevant' from the custom-made dataset and 700 'fear' samples from the reused corpus (we did not mix entries of these emotions from both datasets since they are only contained in one of these corpora). For the 'angry' and 'sad' emotions, we got 350 random samples of each corpus and for the 'happy' tag we obtained 289 samples (all) from our corpus and 411 random samples from the reused dataset. As a result, the balanced dataset obtained contains 3.500 entries.

4.3 Preprocessing and Training

Once we have obtained the datasets, the next step was applying preprocessing to clean tweets. Tweets from the reused dataset were originally clean but we applied again the preprocessing to unify it with our tweets and remove some unnecessary words that may remain after translating (e.g., some stopwords or prepositions). In this stage, focused on tweets collected for our dataset, we also removed unnecessary words or expressions for our purpose, such as user mentions or numbers, and we replaced others that, for the contrary, could be helpful (e.g., we replaced emojis by the emotion that expresses). Details on the tools and libraries used to conduct these preprocessing steps within our framework are described in [9].

With the preprocessed tweets, we could proceed to train the model. For this particular case, we used BETO to train the model, a Spanish version of BERT. We have used 80% of each dataset for training while the 20% remaining was used for validation.

5 Experimental Results

In this section, we compare the results obtained from the BETO training model with each dataset. In Table 1, we show the most used metrics (weighted average) in the ML validation stage for each dataset separately and for both merged datasets: the balanced and the unbalanced ones. For each dataset the method applied was the same, we used an 80% for training and a 20% for validation.

Taking a look at Table 1, we can see the higher accuracy is assigned to our custom-made dataset. Although accuracy is a good metric where the data is balanced, for the unbalancing issues we should observe the F1-Score. Focusing only on this score, the best result corresponds to the balanced merged dataset while the worst results are produced in our custom-made dataset. Apart from analysing the scores, we should examine the confusion matrices in which we clearly observe that best results are from the merged datasets.

For the unbalanced case, we can see that almost all emotions are predicted correctly in 2 of every 3 cases. The 'fear' and 'angry' emotion are confused in several occasions. That is probably because 'fear' does not appear in our custom-made dataset so, in most cases, text that expresses fear also expresses anger and we have classified that way. We could also see that 'surprise' emotion is giving the worst prediction due to the few numbers of samples and is causing lower F1-Score than in the balanced case.

On the other hand, the balanced dataset is giving us better predictions in almost all emotions since we removed the 'surprise' tag. However, once again 'angry' was classified as 'fear' in many cases for the same reason as in the previous case.

Table 1: Experimental results

	Precision	Recall	F1-Score	Accuracy
Custom-made dataset	0.74932	0.51007	0.52039	0.67667
Reused dataset	0.55603	0.56448	0.55365	0.56240
Unbalanced merged dataset	0.69033	0.52522	0.52130	0.62571
Balanced merged dataset	0.59565	0.59519	0.59238	0.59514

6 Discussion

To start with the metrics obtained, we should emphasize that accuracy is not a good measure to compare models' performance if datasets suffer unbalancing problems, as they do not consider how data is distributed. In this regard, F1-score is a more suitable metric to compare results from unbalanced datasets.

Secondly, merging a dataset can be significantly helpful not only to increase the number of samples but also to improve the balance and quality of the training dataset. For instance, in our case, the F1-score for the balanced case is quite high due to the fact that the weakest emotion in our custom-made dataset ('happy') is the strongest one in the reused dataset. The other way around, the 'angry' emotion does not give a good performance in the reused dataset but it does in the custom-made.

Thirdly, and related to the previous point, the next step should be obtaining from each corpus the samples of the emotion that gives better results. In this case, the expected results would be a higher accuracy than the obtained in this experiment. However, we should verify that, with new data (not related to either corpora), the behaviour is as good as in the validation stage.

Finally, merging datasets from two different sources may pose some risks. If the context in which two datasets were built are too different, or if a particular dataset has some bias, the combination of these datasets may cause what is known as negative transfer learning, which means a decrease of performance (in terms of accuracy) through unexpected and undesirable results [23]. It is therefore crucial to understand the context in which the datasets where obtained, their characteristics, and analyse the adequacy of merging them.

7 Threats to Validity

In this section, we discuss the validity threats for this experimental study.

Internal validity. For the custom-made dataset, we monitored all tweets from Twitter selecting random users. For the tagging process, the task of labelling and classifying the text into emotions was done by researchers. We are not psychologists and emotions are subjective, therefore the labelling process can cause inconsistencies and/or confusions for the ML algorithms in some occasions during the training process. To mitigate this risk, we performed piloting and kept only those tweets for which researchers had an agreement from the beginning.

Construct validity. Construct validity refers to the degree in which a test measures what it is supposed to measure. To allay this threat, we used an 80% of the dataset for training and a 20% for validation using 5-fold cross-validation in our ML model development. To evaluate their performance, we used the most common metrics, apart from drawing the confusion matrices that give us more information about the results obtained.

Conclusion validity. To obtain the dataset, we randomly picked tweets every week from different Twitter users for several months. However, picking up text from random people does not assure that the samples collected are representative enough to the whole society. For further experiments, we should extract information about users (e.g., nationality) and select a more significant sample.

External validity. For this study, we have used Covid-19 and Spanish language as a use case. Even though we developed a general framework to be used in any context and for different languages, it may require further experiments to ensure the results in different circumstances.

8 Conclusions

The task of obtaining enough data to finally extract a model accurate enough to be able to predict and classify properly into emotions a piece of text requires time and resources. This experimental study was done to corroborate that, to accelerate the process, we are able to merge data. We used few data and a concrete context area, but it has come in handy as a basis for future experiments.

Regarding our research question formulated in the introduction of this paper, a merged dataset could be considerately helpful. The fact is not so much improving performance regarding an only dataset but saving us time and resources by merging two or more corpora. However, an important point to take into account is that the datasets should be similar in context, kind of emotions and ideally in the language. Additionally, to ensure a good performance the original datasets should be good in terms of balancing of number of samples and, especially, in terms of data quality (i.e. text content).

As future work, we aim to integrate the results of this work into the general concept of data-driven requirements engineering [24], adding emotions to the reports and opinions that users provide about software applications. Emotions can also be helpful when combining this explicit feedback given by users, with implicit feedback gathered from the monitoring of applications at runtime [25]: emotions can help interpreting the actual user behavior in front of particular navigational paths or contexts of use.

ACKNOWLEDGMENTS

This paper has been funded by the Spanish Ministerio de Ciencia e Innovación under project / funding scheme PID2020-117191RB-I00 / AEI/10.13039/501100011033 (DOGO4ML project).

REFERENCES

[1] Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila, 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 27, 16-32. DOI: https://doi.org/10.1016/j.cosrev.2017.10.002.

[2] Bin Lin, Fiorella Zampetti, Gabriele Bavota, Massimilano Di Penta, Michele Lanza, Rocco Oliveto. 2018. Sentiment Analysis for Software Engineering: How Far Can We Go?. In *Proceedings of the 40th International Conference on Software Engineering (ICSE'18)*. ACM Press, New York, NY, 94-104. DOI: https://doi.org/10.1145/3180155.3180195.

[3] Ting Zhang, Bowen Xu, Ferdian Thung, Stefanus Agus Haryono, David Lo, Lingxiao Jiang. 2020. Sentiment Analysis for Software Engineering: How Far Can Pre-trained Transformer Models Go? In *Proceedings of the IEEE International Conference on Software Maintenance and Evolution (ICSME'20)*. IEEE CS, 70-80. DOI: https://doi.org/10.1109/ICSME46990.2020.00017.

[4] Michael Goul, Olivera Marjanovic, Susan Baxley, Karen Vizecky. 2012. Managing the Enterprise Business Intelligence App Store: Sentiment Analysis Supported Requirements Engineering In *45th Hawaii International Conference on System Sciences (HICSS'12)*. 4168-4177. DOI: https://doi.org/10.1109/HICSS.2012.421.

[5] Anastasia Giachanou, Fabio Crestani. 2016 "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Computing Surveys 49(2): 28, pp. 1–41. DOI: https://doi.org/10.1145/2938640.

[6] Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, Marta Villegas. 2021. Are Multilingual Models the Best Choice for Moderately Under-Resourced Languages? A Comprehensive Assessment for Catalan. arXiv:2107.07903v1.

[7] Adil Majeed, Hasan Mujtaba, Mirza Omer Beg. 2020. Emotion Detection in Roman Urdu Text using Machine Learning. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering Workshops* (NLP-SEA@ASE'20), pp. 125-130.. DOI: https://doi.org/10.1145/3417113.3423375.

[8] Peng Yang, Yunfang Chen. 2017. A survey on sentiment analysis by using machine learning methods. In *Proceedings of the* 2nd *Information Technology, Networking, Electronic and Automation Control Conference (ITNEC'17)*, pp. 117-121. DOI: https://doi.org/10.1109/ITNEC.2017.8284920.

[9] Ariadna de Arriba Serra, Marc Oriol, Xavier Franch. 2021. Applying Sentiment Analysis on Spanish Tweets Using BETO. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF)*, pp. 86-93. Available at: http://ceur-ws.org/Vol-2943/emoeval_paper9.pdf

[10] Ariadna de Arriba, Marc Oriol, Xavier Franch. 2021. Applying Transfer Learning to Sentiment Analysis in Social Media. In: *Proceedings of the 5th International Workshop on Crowd-Based Requirements Engineering (CrowdRE'21)*, held at the *29th IEEE International Requirements Engineering Conference (RE'21),* pp. 342-348. DOI: https://doi.org/10.1109/REW53955.2021.00060.

[11] Mattias Engdahl. What is Data Merging? Available at https://www.displayr.com/what-is-data-merging, last accessed August 2021.

[12] Paula Fortuna, Ilaria Bonavita, Sérgio Nunes. 2018. Merging datasets for hate speech classification in Italian. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* co-located with the *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018).*

[13] Kavitha Srinivas, Abraham Gale, Julian Dolby. 2018. Merging datasets through deep learning. arXiv:1809.01604

[14] Jéssica S. Santos, Aline Paes, Flavia Bernardini. 2019. Combining Labeled Datasets for Sentiment Analysis from Different Domains Based on Dataset Similarity to Predict Electors Sentiment. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19)*, pp. 455-460. DOI: https://doi.org/10.1109/BRACIS.2019.00086.

[15] Theresa Wilson, Janyce Wiebe, Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT'05), pp. 347-354. DOI: https://dl.acm.org/doi/10.3115/1220575.1220619.

[16] Cecilia Ovesdotter Alm, Dan Roth, Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT'05), pp. 579–586. DOI: https://dl.acm.org/doi/10.3115/1220575.1220648.

[17] Nancy L. Stein, Linda J. Levine. 1999. The Early Emergence of Emotional Understanding and Appraisal: Implications for Theories of Development. Chapter 9 in Handbook of Cognition and Emotion, Wiley Online library. DOI: https://doi.org/10.1002/0470013494.ch19.

[18] Sotiris B. Kotsiantis. 2007. Supervised Machine Learning: A Review of Classification Techniques. Book chapter in *Emerging Artificial Intelligence Applications in Computer Engineering*, IOS Press.

[19] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Proceedings of Practical ML for Developing Countries Workshop (PML4DC @ICLR'20)*.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2.

[21] Ashish 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] Ariadna de Arriba Serra, Marc Oriol, Xavier Franch. 2021. "Merging Datasets for Emotion Analysis. An Approach using BETO on Spanish Tweets - Supporting material". Zenodo. DOI: https://doi.org/10.5281/zenodo.5191344.

[23] Britannica, T. Editors of Encyclopaedia (Invalid Date). Mind. Encyclopedia Britannica. https://www.britannica.com/topic/mind

[24] Xavier Franch. 2021. Data-Driven Requirements Engineering: A Guided Tour. In: Ali R., Kaindl H., Maciaszek L.A. (eds) Evaluation of Novel Approaches to Software Engineering. *Communications in Computer and Information Science*, vol 1375. Springer.

[25] Marc Oriol *et al*. 2018. FAME: Supporting Continuous Requirements Elicitation by Combining User Feedback and Monitoring. *IEEE 26th International Requirements Engineering Conference (RE)*, pp. 217-227.