Honors Project

Subject: Multilingual Emotion detection in Texts

School of Electrical Engineering and Computer Science
University of Ottawa

Professor: Paola Flocchini

Supervised by: Diana Ikpen

Group members :
Marie Fabienne Sawadogo: 300101795
Gabriel St Pierre: 300146514

1. Introduction

Natural language processing (NLP) is a field of research in Artificial Intelligence that focuses on the connection between computer science and natural language. Emotion detection is one of the most important objectives of this field in which models attempt to detect emotion from various different texts for various applications such as detecting false statements in surveys, analyzing customer sentiment , predicting user behavior and so on. However, most of the research on emotion detection focuses on a single language, English, mainly due to the lack of richness of data available in other languages. Therefore, there is a growing need to develop multilingual emotion detection models to cover multiple languages and cultures.Our model attempts to bridge the gap between English NLP models and the French language, by providing an emotion detection model that is multilingual.

2. Problem being studied

A difficult issue in the realm of natural language processing is the recognition of human emotions in different languages. Although English has been the subject of substantial research, other languages, particularly those with limited resources, have not received the same attention. It is a difficult effort to recognise and decipher the nuanced signs and clues that people employ to describe their emotions. It is challenging to create a model that can correctly identify emotions in multilingual environments since these cues can change between languages and cultures.This is particularly difficult for languages with lower resources, as there is often limited labeled data available for training machine learning models.
This is a problem because analyzing emotions is important in many industries, such as marketing, psychology, and healthcare. Despite these challenges, multilingual emotion detection is an important area of research with numerous potential applications, including cross-cultural marketing, social media analysis, and customer support. Addressing these challenges and improving the accuracy of multilingual sentiment analysis models is currently and will continue to be an important focus for researchers in the field of natural language processing.
In order to overcome those difficulties, we are developing the model to recognize human emotions across different languages, in our case English and French. The model addresses this problem by utilizing a state-of-the-art natural language processing technique called XML-RoBERTa, which is specifically designed for multilingual language processing. This approach allows the model to effectively learn from both English and French labeled data, and can also be expanded to other languages as well. The model is trained on a large, diverse dataset of multilingual text, which enables it to recognize emotional cues and expressions across both languages and cultures.

3. Relevant literature

Emotion detection has been an active area of research in AI for more than two decades, with early work in the late 1990s and early 2000s focused on using machine learning techniques to

classify emotions in text. However, it was not until the invention of social media and the explosion of user-generated content that emotion detection gained widespread attention and became a popular application of natural language processing techniques. Since then, emotion detection has continued to be an active area of research, with new approaches and models being developed to improve accuracy and address challenges such as multilingual emotion detection and detecting sarcasm and irony. In recent years, emotion detection has mostly focused on the study of a single language.We are witnessing the emergence of several documents on Identifying and Categorizing Offensive Language in Social Media and some related shared works as offensive Languages Detection and Analysis [1], Identifying and Categorizing Offensive Language in Social Media [2] and Offensive Language Detection in Social Media [3]. Currently, there is a new trend towards studying multilingualism, and various related tasks have recently garnered a lot of attention from researchers. For example, the Semeval 2019 shared – Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [4], Comparative Evaluation for Tweet Classification [5]. The XLM-EMO model by Bianchi, Nozza and Hovy uses a similar approach to the one we chose, that is building off the XLM-Roberta model however it was trained on tweets, rather than regular texts [9]. As well as being trained on tweets, some languages, such as French in the dataset used were translated, which could result in inconsistencies between the data fed to the model and real French expected from regular tweets. Despite this, satisfactory results were produced by the XLM-EMO model across multiple languages. In addition to the XLM-EMO paper, there is also the TWEETEVAL model and paper created by Barbieri et al [10]. Although the model was only trained using English tweets, it provides a base regarding the ability for the Roberta model to be able to recognize different emotions. The paper demonstrates that the base roberta model that hasn't been trained on the twitter data was successful in predicting most categories of emotion better than the trained model. This gave us confidence that a Roberta based model would be valuable and able to detect emotions in our case as we extended it to include the French language. There are more and more researchers and resources that focus on emotion detection like Giatsoglou et al. [7] who trained a hybrid system based on dictionary-based document vectors, word embeddings, and word polarity to classify the sentiment of tweets. Some searchers also use BERT like XLM-RoBERTa for Multi-language Sentiment Analysis [6]. Our model attempts to use the recent XED dataset to create a more generalized emotion detection model compared to more specific models focussing on areas such as social media responses.

4. Data and Approaches

4.1. Data Description

In order to build our model, we use the datasets in two languages (English and French).The data on which the model was trained is the original training data from the XLM-RoBERTa model on which our model is based. This includes over 2.5TB of data that's found on the internet and acquired by crawling bots. It spans 100 different languages allowing for vast yet extensive

training of the original model. The XED Multilingual dataset for sentiment and emotion detection created at the university of Helsinki [8] was sourced as our French and English text inputs to further train the model on our specific type of text, that being texts containing emotions. The dataset was extended to a further 30 languages containing thousands of data points each including French. Results using the dataset done by the same group indicate that the dataset is at least on par with similar data sources.The dataset contains a large number of data points, with thousands of examples for each language. The dataset can be broken down by category.

Category / Language	**English**	**French**
Anger	2078	1946
Anticipation	1612	1726
Disgust	518	996
Fear	827	1127
Joy	872	1256
Sadness	770	1200
Surprise	646	1198
Trust	511	1259

[Table 1: Distribution of original dataset]

Category / Language	**English**	**French**
Anger	2078	1947
Fear	1345	1277
Joy	2484	2361
Sadness	770	674
Surprise	646	541
Neutral	511	507

[Table 2: Distribution after preprocessing]

4.2. Description of Method designed

The design of the system revolves around the XLM-RoBERTa model, a machine learning model that has been trained on 100 languages and over 2.5TB of data [13]. The first step to using the model was to find a suitable data source that contains sentences and messages in both English and in French. The data source chosen was XED multilingual dataset. This source was chosen as it contained thousands of messages per language as well as multiple languages expanding beyond our original two in the event the model needs to be expanded to other languages. Although the dataset was chosen, it was still required to be cleaned using regular expressions, meaning remove any excess symbols, punctuation and numbers that may cause noise in the model. These can be things like apostrophes, commas, hashtags. We've also made sure that all messages are in lowercase for increased uniformity. Furthermore, as there was not enough data for sentences to be grouped as multiple emotions, we chose the first emotion listed for each sentence. The final distribution of the cleaned data can be seen in Table 2. This allows our model to learn more consistently off the data and produce better results. With the new dataset, now cleaned, it is split up into two categories, training and validation using the sklearn train_test_split method with a split of 80 - 20 for train - validation while using the random_state variable as 2018 . The training category is the data on which the model will attempt to figure out the rules and patterns in the messages and link them to the appropriate resulting emotions. The validation category is used to make sure the model doesn't overfit, meaning that it has learned too much from the training data and recognizes it but cannot make a generalization that will work on non training data. From here, we fed the training data into our model before setting the hyperparameters.

4.3 Setting Hyperparameters

The parameters set for our system are obtained using the Optuna toolchain. The python library in specific will create multiple test runs training our model within a range of hyperparameters that we set. In the case of our model, we set the range for the learning rate of the model as between 1e-6 and 1e-4 based on the similar learning rate of 1e-3 used to train the XLM-EMO model [9]. In addition to the learning rate, we also tune the batch size, or how many different pieces of text the model will analyze and train off of at once. Research from Kandel and Casteilli indicate a direct correlation between batch size and learning rate, A larger learning rate generally works better with a larger batch size [12]. For the reason above, since our learning rate is relatively low, we chose between a batch size of 16 or 8, with findings that 16 generally works better with our dataset. In addition to the reason previously stated, due to the overall size of our dataset being smaller than similar models such as XLM-EMO we were required to keep the size low to avoid overfitting the data which would not let our model generalize the results it obtained. Additionally, to avoid over or under fitting, meaning the model learns the rules for only the specific texts it's trained on, or doesn't learn enough, we were required to monitor the length of time it trained on the data. Not enough time training and the model doesn't learn and too much it will not be able to generalize. Thus, we chose the number of epochs, or iterations over the training data using the Optuna library. By using the library along with ensuring the training loss converges with the validation loss we found that 8 epochs was ideal. In addition, a learning

rate of 1e-5 was determined through optuna to give us the best results. With our new found hyperparameters, we were able to train the model using only these parameters for best results. In general our model works as follows: The pre-processed data enters the model cycle, where the Optuna library is used to select the hyperparameters that best fit the model. The initial learning rate is defined to train the model, and at each trial, the convergence of the training and validation loss is monitored. If the model is not converging or is overfitting the data, a new iteration is performed with a different learning rate chosen from the range of 1e-6 to 1e-4. This process continues until the optimal hyperparameters are found, allowing the model to achieve the best possible performance. By using this method,we were able to avoid under or overfitting the data, and ensure that the model can generalize well to new data.

4.4 Evaluation techniques

A large component of the evaluation of our model revolves around internal model metrics. We believe the more common metrics in the field of machine learning also apply to our model. We believe the recall score, precision, f1-score and overall accuracy of our model help tell an important story about the model's performance. The precision of our model indicates where we should look for the most improvements in making sure we increase the number of true positives, or correct emotion detecting compared to false negative, or statements that are categorized accidentally into that emotion. The recall is similar in the sense that it's the ratio of true positives to the number of true positives and false negatives, that being sentences that should be of this emotion, but got categorized as another. The f1 score provides a look into both metrics, with a higher score indicating a higher precision and a higher recall. Although these metrics give us a good idea of how our model fairs against our test set of data, it's additionally important to evaluate using different techniques. For those reasons, it's important to evaluate our model using various techniques to gain a more comprehensive understanding of its performance. One technique is to use cross-validation, which involves splitting the data into multiple subsets and training the model on each subset while testing on the others. This allows us to assess the model's generalization ability and identify any overfitting issues.Furthermore, it's important to consider the context in which the model will be used and how its performance will impact its intended users. For instance, if the model is being used in a critical application such as healthcare, a small error rate can have serious consequences. Therefore, it's important to not only evaluate the model's performance but also its impact on the concerned users. Overall, evaluating a model's performance is a crucial part of the machine learning process and requires a combination of technical and contextual assessments..

5 Results

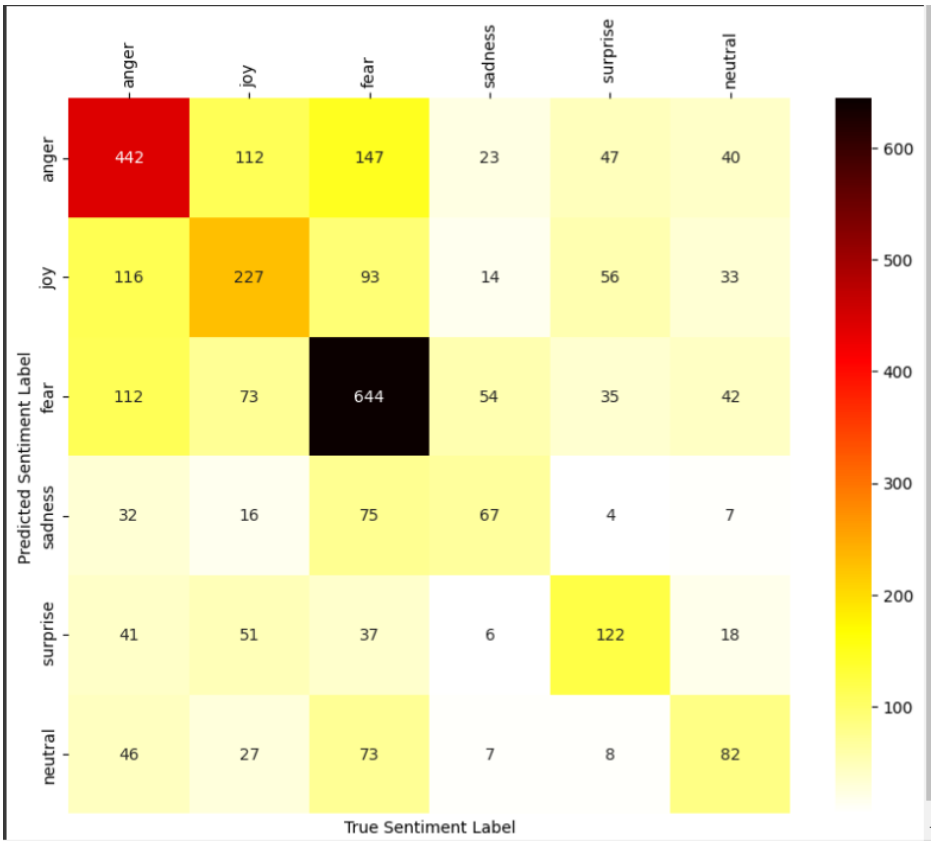
The results obtained by running our model on our test data after training was underwhelming. We conducted tests across both languages as well as a mixed dataset. The mixed dataset results in table 3 indicates we have an overall accuracy of 0.52 or around fifty-two percent. It also states the support, meaning the number of elements from that set that were

used for the test. The accuracy indicates that obtaining the correct emotion will happen about half the time, with the worst changes being neutral, and the best category joy.

	Precision	**recall**	**f1-score**	**support**
Anger	0.56	0.55	0.55	811
Fear	0.45	0.42	0.42	539
Joy	0.60	0.67	0.63	960
Sadness	0.39	0.33	0.36	201
Surprise	0.45	0.44	0.45	275
Neutral	0.37	0.34	0.35	243
Accuracy			0.52	3029

[Table 3: Classification report for mixed test]

To better illustrate the areas our model struggles, we've created a confusion matrix [Figure 1]. Looking at the matrix, the darker the color, the more guesses we have for that pair. We expect the darkest areas to be along the diagonal, as seen in the figure, as this implies we've obtained the correct emotion. This can be useful to determine which emotions are being chosen when a sentence fails, for example we see a large number of neutral sentences actually belong to the fear category, but are being miscategorized. Additionally, we see a lot of misses around the top left corner, however this does not have as big of an effect as they have a larger pool of data. Despite the results, our model was on par with the lower end of the XLM-EMO model, in languages with a limited dataset.



[Figure 1: Confusion Matrix]

Based on relevant work in the multilingual NLP field, we believe the XLM-Roberta model is certainly an appropriate and powerful model upon which to train, however our model was only able to compete with the low end of other research papers. There are many reasons that we observed on why our model ended up with a low accuracy overall when detecting emotions across English and French. The main reasons we've found revolve around the dataset upon which our model was trained. Although the XED dataset is extremely well thought out and greatly appreciated, the classification of sentences as displaying multiple emotions lead to a worse trained model. This is due to the lack of data points that we're multi emotional, thus they could not be integrated into our model's training correctly and we decided to simply take on emotion. We observe a lower accuracy due to the fact that we often ended up predicting one of the original emotions that was removed during data cleaning. For example, the model may predict a text that was originally denoted as [Surprise, Fear] as fear which has been removed during data cleaning. This can be resolved, either by detecting the most likely emotion from the sentence, which could be a project itself, or by increasing the dataset, specifically in data points with multiple emotions attached. Furthermore, we believe an increase in the dataset size would have greatly increased our models accuracy, as although there are a great number of data points, when broken down by category the number is much less impressive. In addition, provided we had access to more powerful equipment to train our model, mainly a graphics card with a larger amount of Vram than the 15GB we used, we believe the accuracy could have been improved by searching upon a wider range of hyperparameters.

6. Relevance of the results

Although our results were not as good as expected based on similar literature, we believe they highlight important details about multilingual NLP moving forward. A major part of training these models rely upon the datasets, which is an area that could use major improvements. This is further supported by the dataset limitations found in the XLM-EMO research paper, in which they were unable to use certain datasets due to lack of data or translated data. We believe that data science research to extract and make this data can be extremely beneficial to machine learning researchers. Furthemore, the results demonstrate the importance of the cleaning and preprocessing area of machine learning creation.Preprocessing and cleaning of data is a crucial step in the machine learning pipeline, and can have a significant impact on the performance of models. In fact, many studies have shown that cleaning and preprocessing can often be more important than the choice of model or algorithm. Therefore, it is important to invest time and resources into this area to ensure that the data is of high quality and suitable for training NLP models. As previously mentioned, the choice to go with a single category when multiple were available made the models learning much more difficult. It highlights the ability for small errors at the beginning of the process to have much larger effects downstream in the model training phase.

7. Future work.

Regarding future work, there is a lot of recent interest within the last couple years towards machine learning and artificial intelligence. Multilingual emotion detection models will become increasingly important, as the advancements made in English will have demand to be ported and to work in more languages. Given the dataset used in this model contains a large amount of average sentences you would expect to read in a book or hear in a movie, it would be interesting to explore a larger set of data containing more slang words and symbols. Data that one would expect to read through a text message or on a social platform such as Twitter. This would open the model up to a much larger use case, with the increasing reliance of the population on the internet and social platforms. We believe that although this would be a large undertaking, the work done on the TweetEval model indicates that the base model can be used on that type of data [10]. In addition to the results from that research, the Roberta model has also been trained on around 38GB worth of reddit content with at least 3 upvotes [11] It would also be interesting to take the dataset we originally used and to explore expanding the language set. The dataset not only includes English and French sentences, but also contains data relating to a multitude of other languages. Opening this up to other languages would widen the appeal of a similar model, however it would probably require knowledge at least to some extent of those languages to make sure the model is working as expected as well as dealing with the problems we encountered with English and French.

8. Conclusion

In this paper, we explore the use of the XED dataset to build upon the XLM-Roberta-base model as a new source for emotion detection. The results achieved from the model indicate are not sufficient to be used in a professional setting, but do indicate the need for larger datasets and greater preprocessing to be done upon the data before training. We believe with a change in dataset, the model could be improved substantially and could be extended to a wider range of languages in order to serve a larger population.

References

[1] "Detecting Offensive Language on Social Networks"
https://arxiv.org/pdf/2203.02123.pdf

[2] "Identifying and Categorizing Offensive Language in Social Media" by Nikhil Oswal
https://arxiv.org/abs/2104.04871

[3] "Offensive Language Detection in Social Media based on Text Classification"
https://ieeexplore.ieee.org/document/9720804

[4] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63
https://aclanthology.org/S19-2007/

[5]Unified Benchmark and Comparative Evaluation for Tweet Classification; by Francesco Barbieri Jose; Camacho-Collados; Leonardo Neves Luis; Espinosa-Anke
https://arxiv.org/pdf/2010.12421.pdf

[6] YNU@Dravidian-CodeMix-FIRE2020: XLM-RoBERTa for Multi-language Sentiment Analysis by Xiaozhi Ou and Hongling Li
https://ceur-ws.org/Vol-2826/T4-13.pdf

[7] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, K. C. Chatzisavvas, Sentiment analysis leveraging emotions and word embeddings, Expert Systems with Applications 69 (2017) 214–224
https://www.sciencedirect.com/science/article/abs/pii/S095741741630584X

[8] Öhman, E., Pàmies, M., Kajava, K. and Tiedemann, J., 2020. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020).
https://aclanthology.org/2020.coling-main.575.pdf

[9]
F. Bianchi, D. Nozza, and D. Hovy, 'XLM-EMO: Multilingual Emotion Prediction in Social Media Text', in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 195–203.

[10]
F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, 'TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification', *arXiv [cs.CL]*. 2020.

[11]
Y. Liu *et al.*, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv [cs.CL]*. 2019.

[12] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," ICT Express, vol. 6, no. 4. Elsevier BV, pp. 312–315, Dec. 2020. doi: 10.1016/j.icte.2020.04.010.

[13] "XLM-Roberta - Hugging Face," *XLM-RoBERTa*. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/xlm-roberta. [Accessed: 23-Apr-2023].