DEPARTMENT OF INFORMATICS
M.Sc. IN DATA SCIENCE

M.Sc. Thesis

Emotion Classification on Greek Tweets

Marina Thalassinou-Lislevand

F3351904

Supervisor: John Pavlopoulos

Athens, November 2020

Abstract

In this thesis, we experimented with zero-shot and transfer learning from the high-resource English language to the low-resource Greek language on the emotion classification task using the multilingual transformer XLM-RoBERTa (XLM-R). In order to tackle emotion analysis, we first fine-tuned XLM-R on the specific task on a big corpus of English tweets. Subsequently, the model was fine-tuned on Greek tweets by using (i) an artificial dataset which was created by retrieving tweets, and (ii) an annotated dataset constructed from scratch in collaboration with the company PaloServices, especially for this project. Furthermore, we compared our model with machine learning models, and we evaluated them on a greek ground truth dataset, which also was constructed with PaloServices for the needs of our task. Finally, we presented the results of all models including the zero-shot learning method, for each category/emotion, mainly on the emotion classification task and secondly on the sentiment analysis task.

Περίληψη

Στη παρούσα διπλωματική πειραματιστήκαμε με τη μηδενική εκμάθηση (zero-shot learning) και τη μεταφορά μάθησης (transfer learning) από την πλούσια (σε πόρους) αγγλική γλώσσα στην ανεπαρκή (σε πόρους) ελληνική γλώσσα στο πρόβλημα της ταξινόμησης κειμένων βάσει συναισθήματος, μέσω του διαγλωσσικού μετασχηματιστή (transformer) XLM-RoBERTa (XLM-R). Προκειμένου να διαχειριστούμε το πρόβλημα της ταξινόμησης κειμένων βάσει συναισθήματος, πρώτα προσαρμόσαμε το XLM-R στο συγκεκριμένο πρόβλημα χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων από αγγλικά tweets. Στη συνέχεια, το μοντέλο προσαρμόστηκε στα ελληνικά tweets χρησιμοποιώντας (i) ένα τεχνητό σύνολο δεδομένων που δημιουργήθηκε από την ανάκτηση tweets, και (ii) ένα ταξινομημένο σύνολο δεδομένων κατασκευασμένο από το μηδέν σε συνεργασία με την εταιρεία PaloServices, ειδικά για αυτό το πρόβλημα. Επιπλέον, συγκρίναμε το μοντέλο μας με μοντέλα μηχανικής μάθησης (machine learning models) και τα αξιολογήσαμε με ένα σύνολο δεδομένων, το οποίο επίσης κατασκευάσαμε με την PaloServices, για τις ανάγκες της εργασίας μας. Τέλος, παρουσιάζουμε τα αποτελέσματα όλων των μοντέλων, συμπεριλαμβανομένης της μεθόδου μηδενικής εκμάθησης (zero-shot learning), για κάθε κατηγορία/συναίσθημα, πρωτίστως για το πρόβλημα της ταξινόμησης κειμένων βάσει εννιά συναισθημάτων και έπειτα βάσει των τριών: θετικό, αρνητικό και ουδέτερο συναίσθημα.

Acknowledgments

I would like to express my heartfelt thanks to my supervisor John Pavlopoulos for his guidance and support, the time he invested, and especially for giving me the opportunity to work on this interesting field. In addition, I would like to thank the company PaloServices and specially Pavlos Polydoras and Konstantinos Korovesis for their help and their valuable advice. Furthermore, I would like to thank my friend Alexandros for the fruitful discussions, and his support during the last year. Finally, a big thanks goes to my family and Giannos for always being there and believing in me.

Contents

1.Introduction

Emotion Detection from text is a recent field of research that is closely related to Sentiment Analysis. Sentiment Analysis aims to detect positive, neutral, or negative feelings from text, whereas Emotion Analysis is a branch of Sentiment Analysis and aims to detect and recognize types of feelings from texts such as *joy, sadness*, etc. In this thesis, we aimed to detect the eight basic emotions from Plutchik's Wheel[1] i.e *joy, trust, fear, surprise, sadness, anticipation, anger, disgust* plus the category *'none'* corresponding to lack of any emotion. *Figure 1.* illustrates Plutchik's Wheel of emotions.

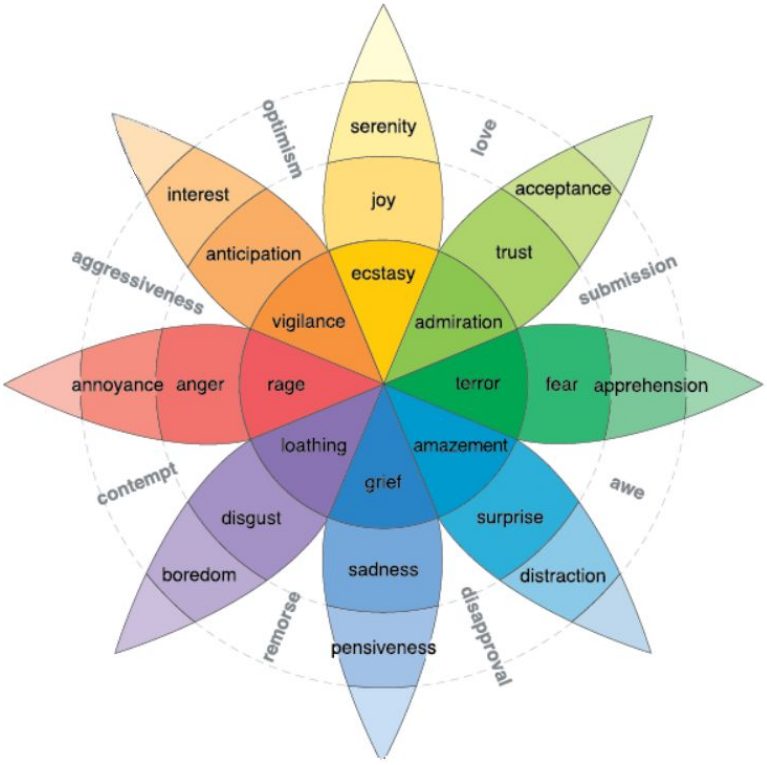

Figure 1: Plutchik's Wheel of emotions.

The advantage of understanding emotions in political science, marketing, human-computer interaction, psychology and many more, made the field of emotion detection in natural language processing (NLP) attractive for the scientific community. In marketing, emotion detection can be used to analyze consumers' reactions to services and products to decide which part of the product or the service should be changed to improve the relationship with customers. Also, emotion detection can be used in recommender systems to provide recommendations or interactions based on the emotional state of the user.

In this thesis, we focused on emotion detection on greek tweets that were annotated in collaboration with the company PaloServices[2]. Paloservices operates a subscription platform

[1] https://en.wikipedia.org/wiki/Robert_Plutchik
[2] http://www.paloservices.com/

for the online reputation of companies, brands, products, organizations, and people. They aggregate news, posts, discussions from the Web and Social Media and provide analytics, including sentiment analysis through machine and deep learning models. For our task, we used zero-shot and transfer learning from a high-resource language (English) to a low-resource language (Greek). Initially, we fine-tuned the multilingual XML-R on the emotion analysis problem by using an English dataset provided by the International Workshop on Semantic Evaluation (SemEval) for the *task*:*"SemEval-2018 Task 1: Affect in Tweets"*. Subsequently, we evaluated our model with a greek ground truth dataset, which we constructed from scratch with the help of PaloServices for the needs of this project. This method is also known as zero-shot cross-lingual classification. Besides the greek ground truth dataset, we constructed two more greek datasets, an artificial dataset by retrieving tweets and a dataset annotated with specific instructions by two professional annotators of PaloServices. Our model was compared to baselines such as Logistic Regression, K-Nearest Neighbors, RandomForest, and all of them were evaluated on the ground truth dataset. We presented the results of the state-of-the-art XLM-R on the low-resource Greek language, mainly for the emotion classification task and secondly for the sentiment analysis task. For the emotion analysis the results of the emotions *disgust* and *none*, and for the sentiment analysis task the results of the categories *negative* and *neutral* are quite impressive especially when applying zero-shot learning.

1.1. Outline

The rest of the thesis is organized as follows:

- **Chapter 2** presents related work.
- **Chapter 3** discusses the methods that we used for the task.
- **Chapter 4** presents the datasets, the implementation of our model, the training methodology alongside the results of the experiments.
- **Chapter 5** draws the conclusions and proposes ideas for future work.

2.Related work

Emotion sentiment classification, especially on social media like Twitter, is a natural language processing (NLP) problem with valuable use cases on real-world data and it has been occupying many researchers in recent years (Bo Pang and Lillian Lee et al. (2002), Jansen et al. (2009), Pak et al. (2010), Kim et al. (2014), Kant et al. (2018), Ian D. et al. (2018), Bostan and Klinger et al. (2018), Gaind et al. (2019), Desai et al. (2020)).

Furthermore, after Vaswani et al. (2017) introduced multilingual transformer models, they have been in the spotlight and they have obtained great improvements for many NLP tasks on a variety of languages, especially for high-resource languages (Devlin et al. (2018), Liu et al. (2019), Conneau and Lample et al. (2019), Conneau et al. (2020)). Still, deliberate NLP tasks at low resource languages have been a challenge for the research community lately (Ranasinghe et al. (2020), Wang et al. (2020), Tela et al. (2020), Hedderich et al. (2020), Launcher et al. (2020)).

Next we review some examples of related work with this thesis.

Desai et al. (2020) presented HURRICANE EMO, an annotated dataset of perceived emotions spanning 15,000 tweets from multiple hurricanes. Tweets were annotated with fine-grained Plutchik-24 emotions, from which they analyzed implicit and explicit emotions and constructed Plutchik-8 binary classification tasks. Experiments with their dataset were demonstrated with both traditional neural models and pre-trained language models.

Kant et al. (2018) demonstrated that large-scale unsupervised language modeling combined with fine-tuning offers a practical solution to 'SemEval Task 1: E-c' multilabel emotion classification problem (Mohammad et al. 2018), based on the Plutchik wheel of emotions (*Figure 1.*). It is worth mentioning that the above dataset from 'SemEval Task 1:E-c' used in this thesis too. They achieved F1 scores such as *fear* (0.73), *disgust* (0.77), *anger* (0.78), *anticipation* (0.42), and *surprise* (0.37) by training an attention-based Transformer network on Amazon reviews and fine-tuned it on the training set.

Ranasinghe et al. (2020) applied cross-lingual contextual word embeddings in offensive language identification projecting predictions from English to Bengali, Hindi and Spanish. They reported results of 0.8415 F1 macro for Bengali, 0.8568 F1 macro for Hindi, and 0.7513 F1 macro for Spanish. Finally, they showed that XLM-R with transfer learning compares favorably to the best systems submitted to recent shared tasks on these three languages, confirming the robustness of cross-lingual contextual embeddings and transfer learning for this task.

Tela et al. (2020) proposed a transfer learning method to use an already existing English monolingual transformer, such as the XLNet model, to deal with NLP tasks of the low-resource language Tigrinya. Also, they released a new Tigrinya sentiment analysis dataset and a new XLNet model specifically for the Tigrinya language, TigXLNet. Interestingly, the results of the English XLNet model fine-tuned with the low-resource language Tigrinya were comparable to the results of monolingual TigXLNet which was pre-trained on Tigrinya text corpus. With only 10k examples of the given Tigrinya sentiment

analysis dataset, English XLNet outperformed BERT and mBERT by 10% and 7% at F1-score, respectively.

Hedderich et al. (2020) evaluated transfer learning and distant supervision on multilingual transformer models mBERT and XLM-RoBERTa, from English to three African languages Hausa, isiXhosa, and Yor`ub ́a, on both Named-entity recognition (NER) and topic classification. They studied realistic low-resource settings for the above African languages and they showed that even with a small amount of labeled data, a reasonable performance can be achieved.

Pires et al. (2019) studied the performance of multilingual BERT at zero-shot cross-lingual model transfer, by fine-tuning the model using task-specific supervised training data from one language, and evaluating that task in a different language. After a large number of experiments, they showed that transfer is possible even to languages in different scripts, and that transfer works better when the languages are typologically similar.

Launcher et al. (2020) obtained that cross-lingual transfer via multilingual transformers is considerably less effective in resource-lean scenarios and for distant languages. Their experiments included three lower-level tasks (POStagging, dependency parsing, NER) and two high-level semantic tasks (NLI, QA). They correlated transfer performance with the philological similarity between the source and the target languages, and with the size of pretraining corpora of target languages. Finally, they introduced few-shot transfer learning where they fine-tuned the model in the source language and then on a few target language instances. With this method, they showed that with a small amount of annotated target language instances, especially at low resource languages the transfer performance improved significantly.

It is worth mentioning that there aren't available related works that deal with zero-shot and transfer learning from English to the Greek language.

3. Methods

3.1 Key Concepts

In this section, we discuss the methods that we used for our task including the pre-trained model, XLM-RoBERTa (XLM-R), and our baseline models. XLM-R is a transformer-based multilingual masked language model (MLM) that is pre-trained on a text in one hundred languages, which obtains state-of-the-art performance on several NLP tasks. XLM-R is based on several key concepts:

3.1.1 Transformers

Vaswani et al. (2017) introduced the Transformers, based entirely on attention and replacing the recurrent layers with multi-headed self-attention. This attention mechanism processes the entire text input to learn contextual relations between words or sub-words. A Transformer consists of two parts, an encoder that reads the text input and produces a representation of it, such as a vector for each word, and a decoder that provides the translated text from that representation.

3.1.2 BERT

BERT (Devlin et al. (2018)), or Bidirectional Encoder Representations from Transformer, uses a "masked language model" (MLM) pre-training objective, that randomly masks some of the tokens from the input which are represented by [MASK]. The objective is to predict the original vocabulary id of the word based only on its context. More precisely, from a given set of tokens of an input sentence with size T, $x= [x_1,x_2,x_3, ...x_T]$, BERT first masks some tokens $y=[y_1,y_2,y_3, ...y_N]$ of the total given tokens, where $N < T$. Then the learning objective will be to predict the masked tokens with the:

$$\max_{\theta} \quad \log p_\theta(y|x) = \sum_{t=1}^{N} \log p_\theta(y_t|x)$$

3.1.3 RoBERTa

RoBERTa introduced by Liu et al. (2019), is a Robustly optimized BERT pretraining approach. They improved the performance of BERT by training the model longer, with bigger batches and more data. Also they removed the next sentence prediction objective (NSP) and they changed dynamically the masking pattern applied to the training data. Lastly, they additionally used the CommonCrawl News dataset.[3]

[3] http://web.archive.org/web/20190904092526/http://commoncrawl.org/2016/10/news-dataset-available/

3.1.4 XLM

XLM introduced by Conneau and Lample et al. (2019)) is a multilingual masked language model that leverages parallel data by proposing a translation language model (**TLM**). TLM extends the BERT MLM approach by using batches of parallel sentences instead of monolingual text streams. They randomly mask tokens in both the source and target languages and the goal of the TLM is to predict these masked tokens regardless of the language that they belong to.

Figure 2. shows the comparison of the masked language modeling (MLM), and the translation language modeling (TLM). The model also collects the language ID and the Positional Encoding which is the order of the tokens in each language. This procedure helps the model learn the relationship between tokens in different languages.

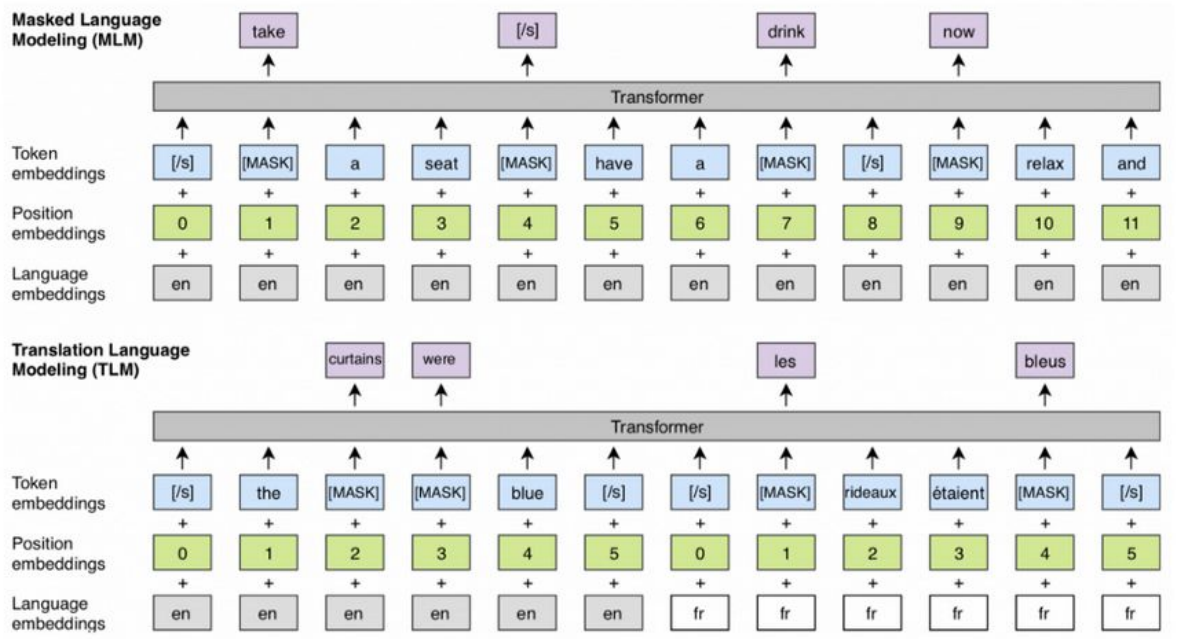

Figure 2: Masked Language Modeling (MLM) and Translation Language Modeling (TLM).

3.1.5 XLM-R

XLM-R, introduced by Alexis Conneau et al. (2020) followed the XLM approach and pushed the state-of-the-art results in many cross-lingual tasks, by increasing the amount of training data and encountered the trade-off of low-resource vs. high-resource languages. They didn't use the TLM objective from the XLM but they trained Roberta (Liu et al. (2019)) with the MLM objective on a huge multilingual dataset. They trained XLM-R on one hundred languages, using more than two terabytes of filtered CommonCrawl[4] data in contrast with

[4] https://commoncrawl.org/

previous works like BERT or XLM which were trained on Wikipedia[5]. With CommonCrawl they increased significantly the amount of data, especially for low-resource languages.

Through the XLM-R they sample streams of text from one hundred different languages, and they train the model in order to predict the masked tokens in the input. They applied sub-word tokenization directly on raw text data using SentencePiece introduced by Kudo and Richardson et al. (2018) alongside a unigram language model introduced by Kudo et al. (2018).

They didn't use word embeddings, like Lample and Conneau et al. (2019) with the XLM, but a large vocabulary size of 250K and they trained two models: XLM-R_{Base} and XLM-R. *Table 1* presents the tokenization used by each Transformer model, the number of layers L, the number of hidden states of the model H_m, the dimension of the feed-forward layer H_{ff}, the number of attention heads A, the size of the vocabulary V and the total number of parameters #params. In our project, we used the XLM-R_{Base}.

Model	**#lgs**	**tokenization**	**L**	**H_m**	**H_{ff}**	**A**	**V**	**#params**
XLM-R_{Base}	100	SPM	12	768	3072	12	250K	270M
XLM-R	100	SPM	24	1024	4096	16	250K	550M

Table 1: The parameters of XLM-R Base and XLM-R.

3.2 Zero shot-transfer learning

A method that was used in this thesis, is the zero-shot cross-lingual model transfer. Through this method we fine-tune the model using task-specific supervised training data from one language (usually high resource languages such as English), and evaluate that task in a different language (usually low resource languages such as Greek).

The fine-tuning procedure on the specific task (e.g emotion classification) is performed by adding one fully-connected layer on top of the XLM-R model and by training for a few epochs. This transfer learning approach eliminates the cost of pre-training a new model from scratch and achieves high performance even at low resource languages where we have a small amount of labeled data. The method is illustrated in *Figure 3*.

[5] https://www.wikipedia.org/

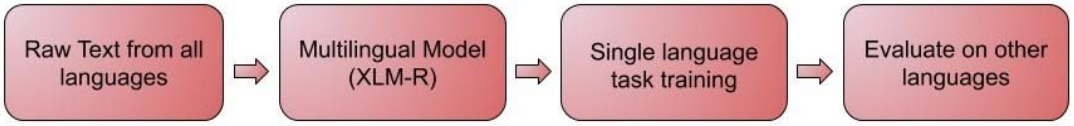

Figure 3: Zero-shot Learning Pipeline.

3.3 Baselines

In this section, we present the three models that were used as baselines in this project. The implementations of all of them are provided by the Python Library sci-kit-learn and they support multi-label classification tasks.[6]

3.3.1 KNN (K-Nearest Neighbor)

KNN is a supervised non-parametric classification algorithm that keeps the entire training set instead of splitting it into a training set and test set. It uses the training dataset to make future predictions by computing the similarity between an input sample and each training instance**.** The distance measures that can be used are Euclidean distance, Hamming distance, or Manhattan distance. The model selects **k** entries from the data which are closest to the new entry and through the majority vote classifies it to the most common class from the **k** entries.

3.3.2 Random Forests

Random Forests are also a supervised learning algorithm that (i) creates a decision tree for each sample, (ii) receives a prediction from each tree, and (iii) chooses the best solution through voting. The most important parameter of Random forests is the number of decision trees that participate in the procedure.

3.3.3 Logistic Regression

Logistic regression is a classification algorithm which is used to assign observations to a discrete set of classes. It transforms its output using the logistic sigmoid function in order to return a probability value. When we pass the inputs through a prediction function the classifier gives us a set of outputs or classes which are probabilities scores between 0 and 1. Then, we decide depending on whether they are above or below the threshold value, which classes will be labeled with one and which with zero.

[6] *https://scikit-learn.org/*

4. Experiments

4.1 Datasets

In this section, we will discuss the datasets that were used for our task.
These datasets are :

1. **SemEval-En**, the dataset of 'SemEval Task 1: E-c'[7] for the multilabel emotion classification problem (Mohammad et al. 2018). It is based on the Plutchik[8] wheel of emotions.
2. **Artificial-Gr**, an artificial dataset which is constructed by retrieving greek tweets, which are labeled with the nine main emotions
3. **ES (Evaluation Set),** a dataset of greek tweets constructed from scratch which was used as a gold benchmark for evaluating our models.
4. **Palo-Gr,** a dataset of greek tweets constructed from scratch for the needs of this project.
5. **Palo-Gr+,** the previous dataset enriched with comments that are classified as positive emotions.

Figure 4 shows the frequency of each category for each dataset

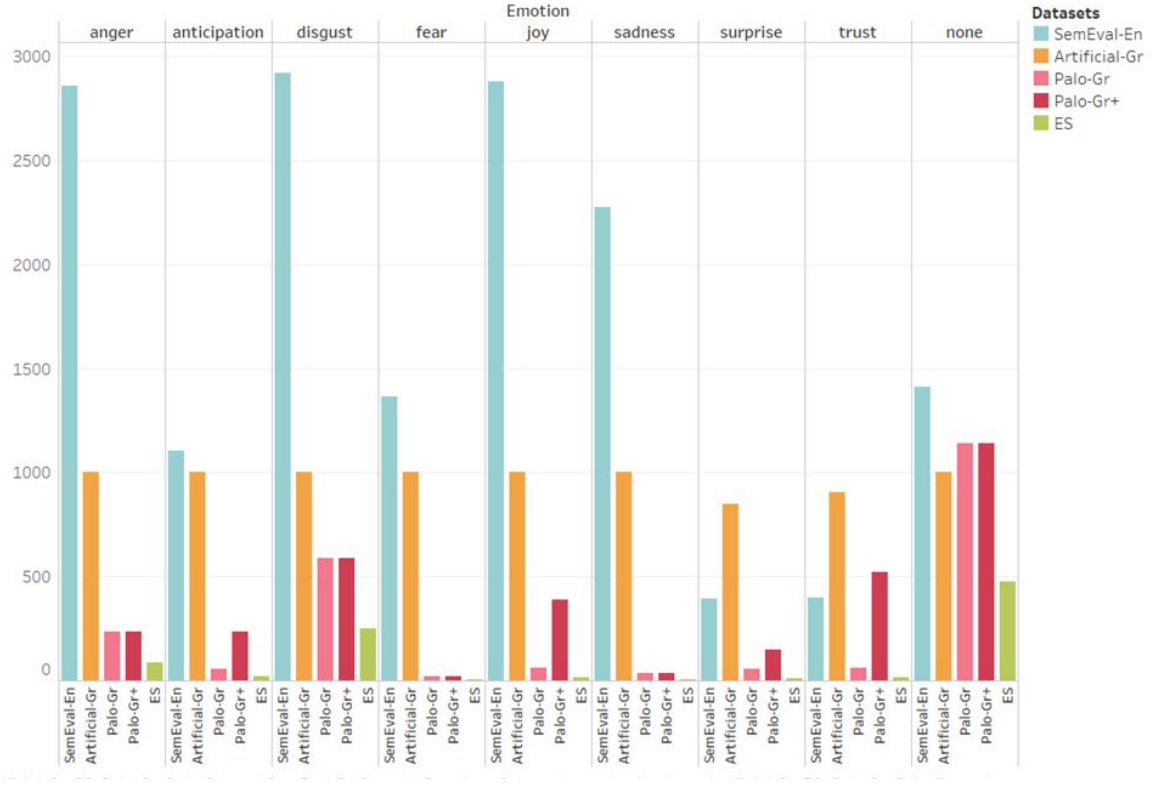

Figure 4: The support of each emotion per dataset.

Below we present in detail each of the previous datasets.

[7] *https://competitions.codalab.org/competitions/17751*
[8] *https://en.wikipedia.org/wiki/Robert_Plutchik*

4.1.1 SemEval(w/o)-En & SemEval-En

SemEval(w/o)-En is a multi-dimensional emotion dataset introduced by Mohammad et al. (2018) for the 'SemEval Task 1:E-c' problem: "Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter".
This dataset offers a training set of 6,857 tweets, a development set of 886 tweets, and a test set of 3259 tweets with binary labels for the eight Plutchik categories (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy), plus optimism, pessimism, and love. The categories are not mutually exclusive, i.e. a comment may belong to one or more categories. For example, the tweet: '*Don't be afraid to start. Be afraid not to start. #happiness*', belongs to three classes, fear, joy, and optimism.

In this project we used the primary eight Plutchik emotions plus the category 'none', to describe neutral tweets. To enrich the *SemEval(w/o)-En* dataset with tweets labeled as neutral, we created the *SemEval-En* by adding to the *SemEval(w/o)-En* 1,394 neutral timeline tweets of the British newspaper 'The Telegraph',[9] which were provided by the online community Kaggle.[10] The pre-processing of the *SemEval(w/o)-En* and the *SemEval-En* dataset was implemented by removing links and users, e.g '*@vanessa*', which are very common in tweets.

4.1.2 Artificial-Gr

We constructed the *Artificial-Gr* dataset from scratch, by retrieving greek tweets with Tweepy,[11] a Python library for accessing the Twitter API. In order to retrieve tweets for each emotion, we searched them by using specific words that could be used from the users due to their emotional state. For example, in order to collect tweets for the emotion *joy*, we searched tweets that contain words such as '*χαίρομαι*', which means '*I am happy*' in the Greek language.

After the removal of retweets, duplicates, links, and users ('*@user*'), we managed to store 1000 tweets for each of the categories *anger, anticipation, disgust, fear, joy,* and *sadness*, 846 tweets for *surprise*, and 907 for *trust*. *Table 2* shows which words were used to retrieve tweets for each emotion. It was a challenge to find words for each category, especially for the positive ones. At first glance, some obvious words have been omitted such as '*χαρά*', ('happiness') for the category *joy*. This happened because many tweets with the word '*χαρά*' didn't belong to the category *joy*, e.g. '*Η χαρά είναι ένα συναίσθημα που πρέπει να εκφράζεται στον ίδιο βαθμό όπως και τα υπόλοιπα'*, ('*Happiness is an emotion that must be expressed to the same degree as the rest.*') that belongs to the class *none*.

Despite the careful selection of words, we encountered issues like irony or sarcasm (which included in many tweets), and tweets with words like '*δεν*' ('*don't*'), which changed completely the initial emotion that they belong to. E.g The tweet '*Δεν εμπιστεύομαι τις ειδήσεις'* ('*I don't trust the news*') based on our method was classified to the category '*trust*'.

[9] *https://www.telegraph.co.uk/*
[10] *https://www.kaggle.com/*
[11] *https://www.tweepy.org/*

These problems, after all, point to us the necessity of manual annotation.

Words	Category
‘αίσχος’, ‘έλεος’, ‘ήμαρτον’, ‘αι σιχτιρ’, ‘γαμώ’, ‘νιώθω εξοργισμένος’, ‘νιώθω οργή’, ‘βλάκα’, ‘ηλίθιος’, ‘σιχαμα’	anger, disgust
‘περιμένω’, ‘αναμένω’, ‘προσμένω’	anticipation
‘φοβάμαι’, ‘τρομάζω’, ‘τρομακτικό’, ‘τρέμω’, ‘σκιάζομαι’	fear
‘χαίρομαι’, ‘είμαι χαρούμενος’, ‘πολύ χάρηκα’, ‘αχ ναιιι’, ‘ναιιι’, ‘τέλειοο’, ‘εκστασιασμένος’	joy
‘λυπάμαι’, ‘στεναχωριέμαι’, ‘θλίβομαι’, ‘θλίψη’, ‘απογοήτευση’	sadness
‘εκπλήσσομαι’, ‘έκπληξη’	surprise
‘εμπιστοσύνη’, ‘εμπιστεύομαι’	trust

Table 2: Words that are used for retrieving tweets for each category.

4.1.2 Evaluation Set (ES)

In this project, we constructed from scratch a ground truth dataset which was used as a gold benchmark for evaluation of models. The ES corpus includes greek tweets provided by the company Paloservices[12], whose platform is used for monitoring, measuring, and analyzing web mentions (like tweets) for companies, brands, people, or products. Data were annotated by two professional annotators from PaloServices and were classified into the following 10 categories: *anger, anticipation, disgust, fear, joy, sadness, surprise, trust, other*, and *none*.

As a first step, it was given to both annotators the same batch of 100 greek tweets for annotation, with the only instruction, not to communicate with each other. Cohen’s Kappa coefficient,[13] a metric with maximum value the number one and minimum the number zero , was used to measure inter-rater reliability and it was equal to 0.29, which is low. So we proceeded to the next step, the second round of annotation, where both annotators should annotate another batch of 100 same tweets, this time with the following strict guideline (based on Mohammand et al. (2018)).

We presented one tweet at a time to the annotators and we asked them two questions. The first was a single-answer multiple choice question:
Q1. Which of the following options best describes the emotional state of the tweeter?

[12] *http://www.paloservices.com/*
[13] *https://en.wikipedia.org/wiki/Cohen%27s_kappa*

(For each category, we also presented example tweets for the corresponding emotion).

– **anger** (also includes annoyance, rage)
e.g *"Εν τω μεταξύ όλοι δίνουν παράδειγμα την Παπαστράτος. Τα ξενοδοχεία πως δουλεύουν ρε μπετόστοκοι; Είδατε Κυριακή κλειστό ξενοδοχείο; Σκατά έχουν για μυαλό ρε μλκ τι να πω...#syriza_xeftiles #ΞΑΝΑΕΡΧΕΤΑΙ"*

– **anticipation** (also includes interest, vigilance)
e.g :*"Ελπίζω να καταφέρει να ανεβάσει ποιοτικά το νετφλιξ αν υπάρχει τέτοιο ενδεχόμενο"*

– **disgust** (also includes disinterest, dislike, loathing)
e.g:*"Παιδιά μια συμβουλή μακρυά από FORTHNET χαλαρα ότι από απαίσιο κυκλοφορεί σε Ίντερνετ"*

– **fear** (also includes apprehension, anxiety, terror)
e.g: *"Φοβάμαι πως η επόμενη φάση της πανδημίας στη χώρα άρχισε νωρίτερα από ότι υπολογίζαμε. Το φθινόπωρο τα πράγματα είναι σχεδόν σίγουρο ότι θα εξελιχθούν σε ένα νέο (χειρότερο) κύμα ή την διόγκωση του τωρινού, ακριβώς για τους λόγους που γράφεις."*

– **joy** (also includes serenity, ecstasy)
e.g:*"Αυτός που μου δίνει τους κωδικούς πλήρωσε ΕΠΙΤΕΛΟΥΣ το Νετφλιξ. Θα πάθω εγκεφαλικό απ τη χαρά μου."*

– **sadness** (also includes pensiveness, grief)
e.g*:"Με λύπη μου σας λέω , ότι αν είστε συνδρομητής @COSMOTE κ έχετε τεχνική βλάβη , ούτε άκρη θα βρείτε Σάββατο Κυριακο κ για την αποκατάσταση της, μπορεί να περιμένετε μια βδομάδα!!!!"*

– **surprise** (also includes distraction, amazement)
e.g:*"Υπέροχη νέα εφαρμογή Cosmote TV επιτέλους έχει και το Ε!"*

– **trust** (also includes acceptance, liking, admiration)
e.g:*"@SpyrosLAP: Αυτό είναι πολύ καλό. Ώρα του Υπ Παιδείας να πάρει τη χώρα μπροστά #Cyprus #Cyta @AnastasiadesCY #ΜΕΝΟΥΜΕΣΠΙΤΙ #StayAtHome"*

– **other**(sarcasm,irony,or other emotion)
e.g: *"ΟΤΕ ακούς; Τηλεφωνώ στο 13888 από την Παρασκευή, αλλά άκρα του τάφου σιωπή. Τί έπαθε ο γίγαντας των τηλεπικοινωνιών μας; @COSMOTE".*

– **none**
e.g:*"Αυτές είναι οι νέες σειρές και οι νέες ταινίες που έρχονται στο Netflix μέσα στο Δεκέμβριο! https://t.co/pxIpmDyZx1"*

The second question was a checkbox question, where multiple options could be selected:
Q2. In addition to your response to Q1, which of the following options further describes the emotional state of the tweeter? Select all that apply.

With the above guideline, the Kappa score of the primary emotions (answers of the Q1) improved to 0.36. Also, we measured the degree of agreement for the multi-label problem with Fleiss Kappa[14] score, also a metric with maximum value the number one and minimum the number zero, and it was equal to 0.26. Furthermore, from the second round of annotation, we observed that annotators often disagreed on tweets that contained news or announcements. E.g: *'@: 2 εκατ. ευρώ για την ενίσχυση των ελληνικών νοσοκομείων προσφέρει ο Όμιλος OTE https://t.co/MB2HgJ7M6A',* was classified in categories *none* and *joy* by each annotator respectively.

This observation led us to update the existing guideline with the following note: 'If the tweet involves news/announcement, it should be classified in the *none* class, assuming that the author does not have the emotion expressed by the news.
e.g: '*ΑΠΟΚΛΕΙΣΤΙΚΟ: Επίκαιρη Επερώτηση για NOVA και αθέμιτο ανταγωνισμό Μαρινάκη… καταθέτει ο ΣΥΡΙΖΑ! https://t.co/lJDdbZRzM0 μέσω του χρήστη @Piperata»*'.

With the updated guideline we proceeded to the next annotation round for the creation of the *ES*. At this part, it was given the same batch of 999 tweets to both annotators. We created a ground truth dataset using only the tweets that both annotators agreed. The Kappa score (for the multi-class task) of primary emotions was equal to 0.51 and the most important was that Fleiss Kappa of all emotions (for the multi-label task) was equal to 0.44. We kept 786 out of 999 tweets that annotators agreed on at least one emotion, and rejected 146 tweets with no agreement. Also, we didn't include 68 tweets labeled with the emotion 'other', which we decided not to use in our task. In *Table 3.,* we present the Kappa score and the Fleiss Kappa score for each step of the annotation.

Steps	**Kappa**	**Fleiss**
Round 1	0.29	-
Round 2	0.36	0.26
Ground Truth	**0.51**	**0.44**

Table 3*:* Kappa score and Fleiss Kappa score for each step of annotation.

4.1.3 Palo-Gr

Because of high Kappa and Fleiss scores, we created the *Palo-Gr* dataset with the same guideline by providing two different batches of 1000 greek tweets, one per annotator. From the multilabel annotated dataset of 2000 tweets, we excluded 135 tweets with the label '*other*', and the rest 1865 tweets, after the removal of links and users, were used for fine-tuning.

[14] https://en.wikipedia.org/wiki/Fleiss

4.1.4 Palo-Gr+

Palo-Gr+ consisted of *Palo-Gr* dataset plus 543 tweets annotated with positive emotions i.e *anticipation*, *joy, surprise, trust*. *Figure 5* shows the support of positive emotions for the *Palo-Gr* and the *Palo-Gr+* dataset.

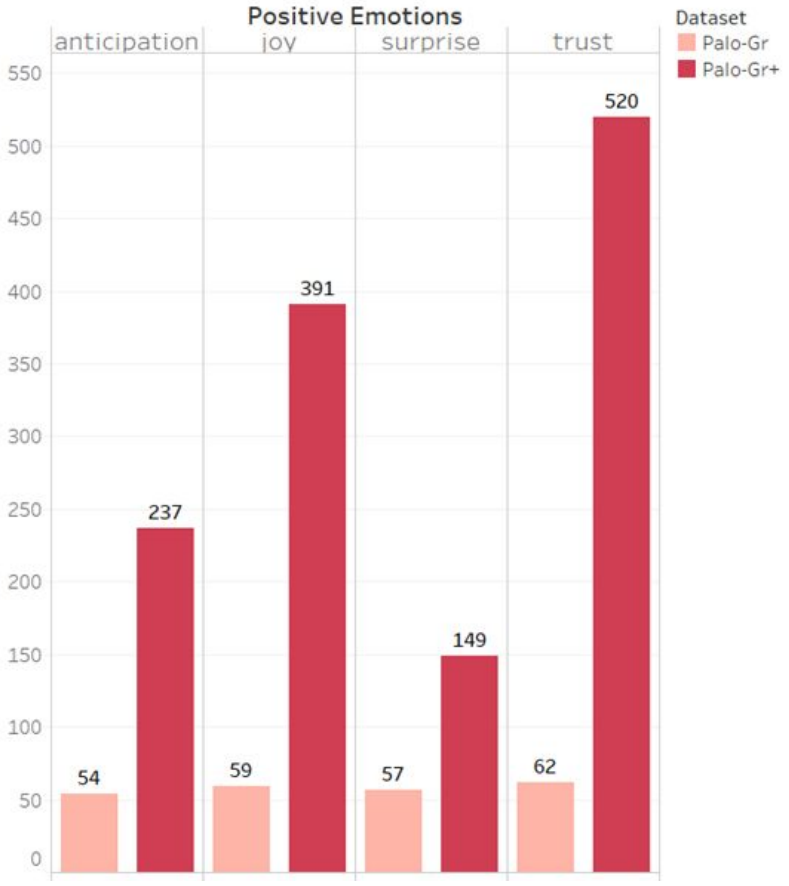

Figure 5: The frequency of positive emotions on Palo-Gr dataset and on Palo-Gr+ dataset.

Dataset	total	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
SemEval-En	**9137**	**31.3**	12.1	**32.0**	14.9	31.5	24.9	4.3	4.4	15.5
Artificial-Gr	**6313**	15.8	15.8	15.8	15.8	15.8	15.8	13.4	14.4	15.8
Palo-Gr	**1865**	12.7	2.9	**31.4**	1.0	3.2	1.9	3.1	3.3	**61.1**
Palo-Gr+	**2408**	9.8	9.8	24.3	0.8	16.2	1.5	6.2	21.6	**47.3**
ES	**786**	10.8	2.8	31.7	0.5	1.8	0.6	1.4	2.2	**60.6**

Table 4: The total number of tweets and the percentage of each emotion per dataset.

4.2 Implementation

We implemented our model by using the Hugging Face/Transformers library. Hugging Face/Transformers is a python-based library that provides an API to use transformer architectures including XLM-R.

In this project we used the TensorFlow model of XLM-RoBERTa base loaded by the name jplu/tf-xlm-roberta-base. The library develops three main classes: a configuration class, a tokenizer class, and a model class.

The configuration class: the configuration class contains relevant information regarding the model we used, such as the number of layers, the number of attention heads etc. *Figure 5* presents XLM-R's configuration file, for the pre-trained weights xlm-roberta-base-cased.

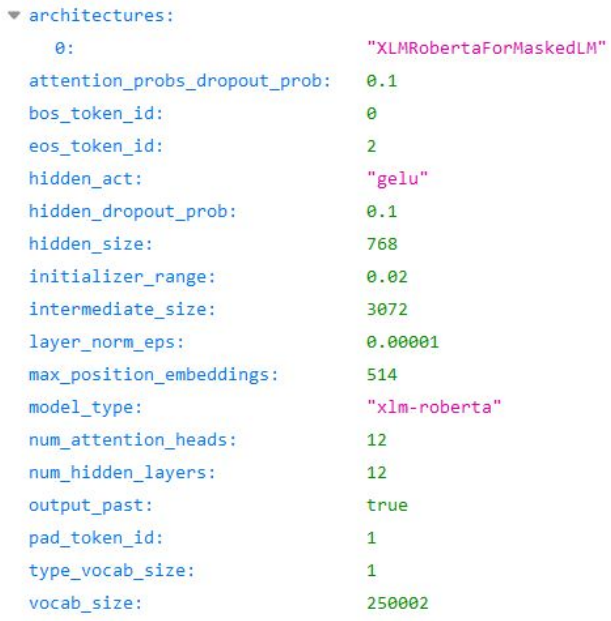

Figure 5: Configuration file of XLM-R base

The tokenizer class: the tokenizer class converts python strings into arrays or tensors of integers based on the vocabulary. It contains many useful features for the tokenization of a string into tokens. We used XLM-R's tokenizer which is based on SentencePiece.

The model class: the model class possesses the neural network modeling logic. We loaded the TensorFlow model named jplu/tf-xlm-roberta-base.

4.3 Preparation of the Data

After loading the tokenizer and the model from Hugging Face, we applied the function 'batch_encode_plus' to the tokenizer, which splits the sentence into tokens, adds special tokens at the beginning and the end of sequences (like *[SEP]*, *[CLS]*, *</s>* or *<s>* for instance) and maps the tokens to their IDs. So, it generates a dictionary that contains the input_ids, token_type_ids, and the attention_mask as a list for each input sentence. In our case we only wanted to generate the input_ids, so we set the arguments return_token_type_ids and return_attention_mask to False. Also, after measuring the max length of the sentences, we set the max length to sixty, and we padded and truncated all sentences to this constant length. For padding, this method adds the special tokens: *<pad>* at the end of each sentence. *Figure 6* presents the transformation of the initial sentence (step 1) to the tokenized sentence (step 2), and finally to inputs_ids (step 3). The tokenized sentence and the final array have a length of sixty tokens.

Then, we built an input pipeline for our model by leveraging the tensorflow_datasets[15] package for data loading. Tensorflow-dataset creates a source dataset from the input data, applies dataset transformations to preprocess the data and iterates over the dataset. This is helpful for saving memory during training because with an iterator, the entire dataset does not need to be loaded into memory. Using this method we batched the dataset in batches of 16 and also shuffled the training dataset.

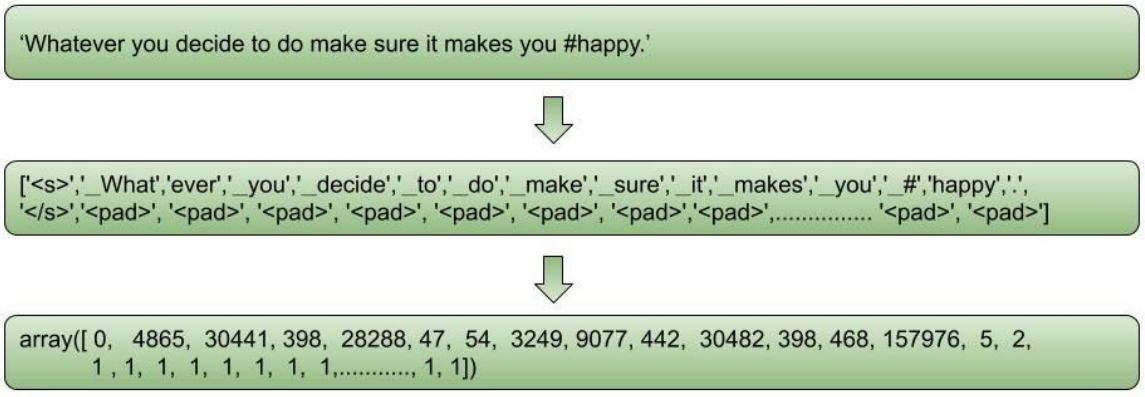

Figure 6: Steps of Tokenization of tweets

[15] *https://www.tensorflow.org/datasets/*

4.4 Model

We built a functional model by adding a Feed Forward Neural Network (FFNN) on top of the pre-trained XLM-RoBERTa. As *Figure 7* shows, we fed XLM-R with vectors (with the length of sixty values) that contain the input ids of the sentences, and subsequently, XLM-R fed the FFNN with its output i.e the context-aware embedding (length of 768) of the *[CLS]* token of each sentence. The number of nodes in the output layer is the same as the number of classes i.e eight. Since our task is a multi-label classification i.e the categories are not mutually exclusive, each node in the output layer uses as activation function the sigmoid. This predicts eight independent probabilities, between 0 and 1. Unlike the softmax activation function, probabilities are not constrained to sum to one, because the sigmoid looks at each raw output value separately.
The model is illustrated in *Figure 7*.

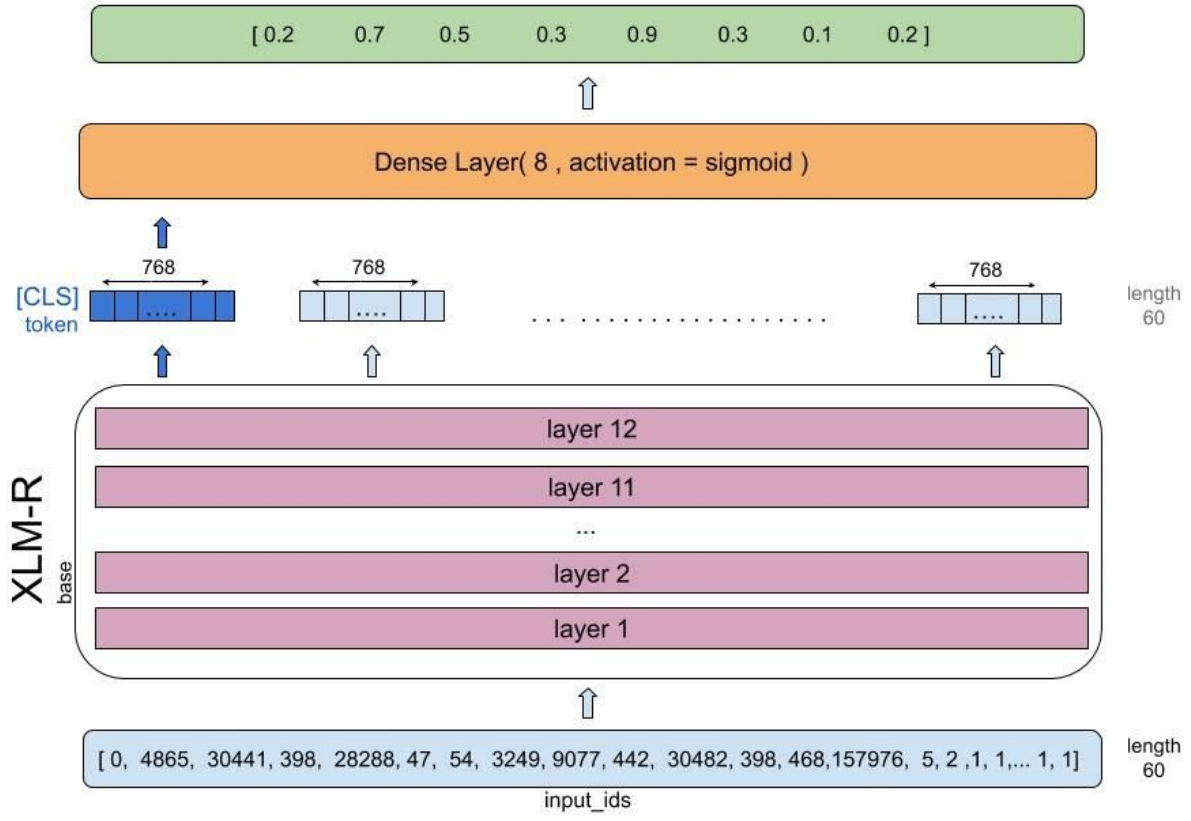

Figure 7: Architecture of our model.

4.5 Training

Binary cross-entropy loss function (CE) was used as the function to be minimized during training. CE also called sigmoid cross-entropy loss is a sigmoid activation plus a

cross-entropy loss. The equation below corresponds to the CE function, where y is the true class of an example encoded as a binary digit, and p is the probability predicted by the model.

$$CE(p,y) = -(ylog(p) + (1-y)log(1-p))$$

Furthermore, we used the Adaptive Moment Estimation optimization algorithm (ADAM) with learning rate equal to 1e-5 (0.00001). The data were split into batches of size 16. To avoid overfitting we applied the Early Stopping method with patience 7 and we monitored the validation loss. That means that during the training, we kept a record of the loss function on the validation data, and when we saw that there was no improvement on the validation set after 7 epochs we stopped, rather than going all the epochs. This strategy prevented the model from overfit the training set by stopping the training early before the absolute value of the weights became large.

4.6 Evaluation metric

Since our classes are highly imbalanced, evaluation was chosen to be the Area Under Precision-Recall Curves (AUPRC). The Precision-Recall curves (PRC) show the tradeoff between precision and recall for different thresholds. A high area under the curve indicates both high recall and high precision, where high precision corresponds to a low false-positive rate, and high recall corresponds to a low false-negative rate.

Precision is defined as the number of true positives over the number of true positives plus the number of false positives and recall as the number of true positives over the number of true positives plus the number of false negatives.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Although the measure that we trust is AUPRC, we also calculated the F1-score and AreaUnder the Receiver Operating Characteristic curve (AUROC) for each class. The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. The F1-score of a classification model is calculated as follows

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The ROC curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at several threshold settings. ROC corresponds to a probability curve and AUC indicates the degree or measure of separability, i.e tells how well a model can distinguish the classes.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + FN}$$

4.7 Baselines' Parameters

The tokenization method that was used for our baseline models is the CountVectorizer[16] provided by the sci-kit-learn library. We used word-frequency-based features, with max frequency equal to 50,000, and the ngram range equal to (1,1).

Also after tuning, we used specific parameters at the machine learning baselines *(Section 3.3)*. For the K-Nearest Neighbor (KNN) algorithm we used five neighbors, and for the Random Forest classifier we used 140 trees in the forest. Logistic Regression was implemented with default parameters.

4.8 Preliminary Experiments (with SemEval(w/o)-En)

In the preliminary experiments, our training dataset was the *SemEval(w/o)-En* ***(SE(w/o))*** *(Section 4.1.1)*. Also, since sixty percent of the *ES* dataset belongs to the category 'none'*(Table 4)*, and the training dataset doesn't include any neutral tweets, we used two datasets for evaluation: (i) the *ES (Section 4.2.1)* and (ii) the *ES* without the neutral tweets (ES-w/o). We present the following experiments:

- **XLM-R@SE(w/o)**
  We further pre-trained XLM-R on the English *SemEval(w/o)-En* (**SE(w/o)**) dataset and then we evaluated it on the two greek evaluation datasets, **ES** and **ES-w/o**. This can also be referred to as zero-shot learning.

- **XLM-R@SE(w/o)-Art**
  We further pre-trained XLM-R on the *SemEval(w/o)-En* (**SE(w/o)**) dataset, then we fine-tuned it on the *Artificial-Gr (**Art**)* dataset and we evaluated it on the **ES** and **ES-w/o** datasets.

- **XLM-R@SE(w/o)-Palo**
  Similarly, we further pre-trained XLM-R on the *SemEval(w/o)-En* (**SE(w/o)**) dataset, then we fine-tuned on the *Palo-Gr (**Palo**)* dataset and we evaluated it on the **ES** and **ES-w/o** evaluation datasets.

[16] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Figure 8 presents the AUPRC of the above experiments, for each category/emotion, on the ES, ES-w/o evaluation datasets. *Figure 9* illustrates the average AUPRC of all the emotions per system on the ES and ES-w/o and *Figure 10* illustrates the average AUPRC of all the systems per emotion.

		Emotions								
Systems	Evaluation sets	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
XLM-R@SE(w/o)	**ES**	0,22	0,04	0,47	0,00	0,04	0,01	0,01	0,02	0,65
	ES-w/o	0,39	0,19	0,89	0,01	0,12	0,02	0,04	0,06	-
XLM-R@SE(w/o)-Art	**ES**	0,20	0,03	0,39	0,00	0,17	0,01	0,11	0,02	0,69
	ES-w/o	0,36	0,08	0,82	0,01	**0,20**	0,02	**0,13**	0,06	-
XLM-R@SE(w/o)-Palo	**ES**	0,34	0,08	0,77	0,00	0,15	0,01	0,01	0,09	**0,86**
	ES-w/o	**0,42**	0,12	**0,90**	0,01	0,11	0,02	0,04	**0,12**	-

Figure 8: The **AUPRC** for the systems XLM-R@SE(w/o), XLM-R@SE(w/o)-Art, XLM-R@SE(w/o)-Palo per emotion on the ES, ES w/o evaluation datasets.

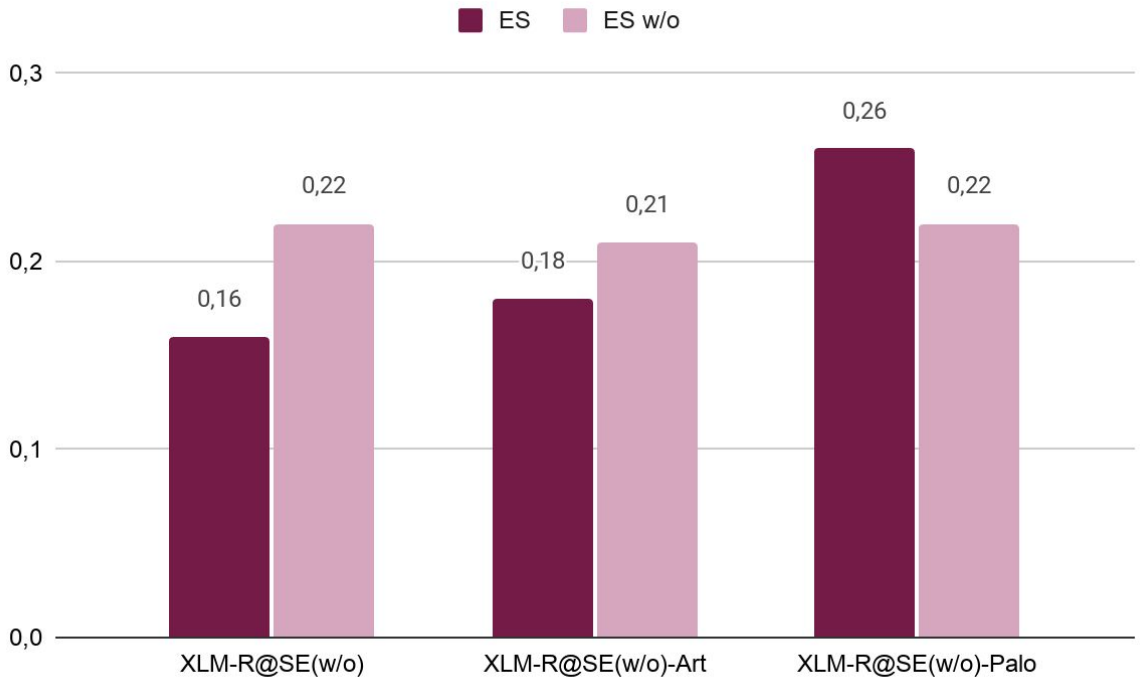

Figure 9: The average **AUPRC** of all the emotions per system on the ES and ES-w/o.

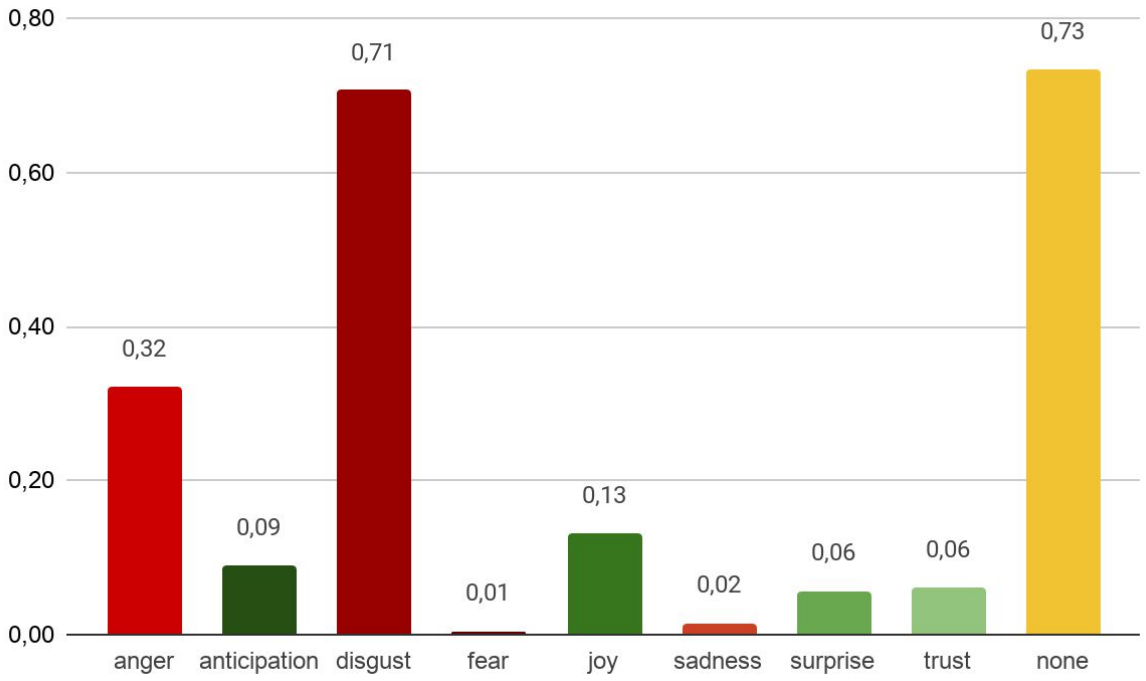

Figure 10: The average **AUPRC** of all the systems per emotion.

Although the metric we trust is the AUPRC, we additionally present the AUROC and the F1 score of the above experiments in *Figures 11,12,13 and Figures 14,15,16* respectively.

		Emotions								
Systems	Evaluation Sets	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
XLM-R@SE(w/o)	ES	0.71	0.63	0.71	0.50	0.73	**0.59**	0.50	0.50	0.52
	ES w/o	0.61	**0.70**	0.74	0.50	**0.76**	**0.59**	0.50	0.50	-
XLM-R@SE(w/o)-**Art**	ES	0.60	0.52	0.59	0.50	0.58	0.50	**0.55**	0.50	0.60
	ES w/o	0.56	0.51	0.56	0.50	0.58	0.50	**0.55**	0.50	-
XLM-R@SE(w/o)-**Palo**	ES	**0.78**	0.53	**0.91**	0.50	0.57	0.50	0.50	**0.53**	**0.87**
	ES w/o	0.65	0.53	0.78	0.50	0.57	0.50	0.50	**0.53**	-

Figure 11: The **AUROC** for the systems XLM-R@SE(w/o), XLM-R@SE(w/o)-Art, XLM-R@SE(w/o)-Palo per emotion on the ES, ES w/o evaluation datasets.

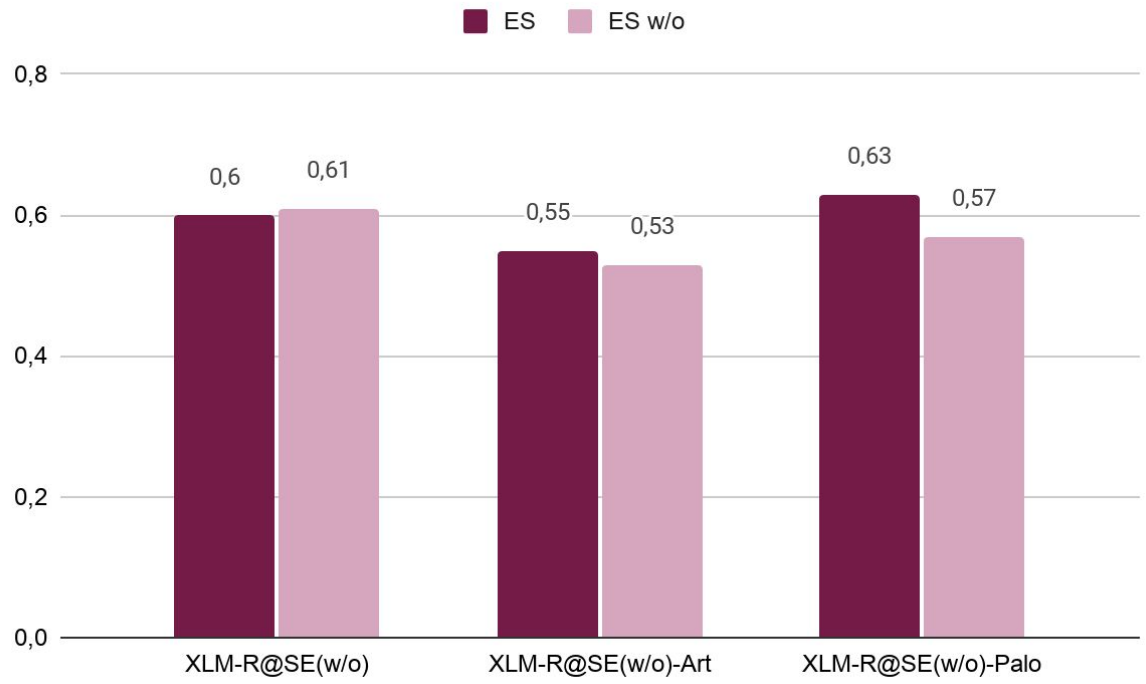

Figure 12: The average **AUROC** of all the emotions per system on the ES and ES-w/o.

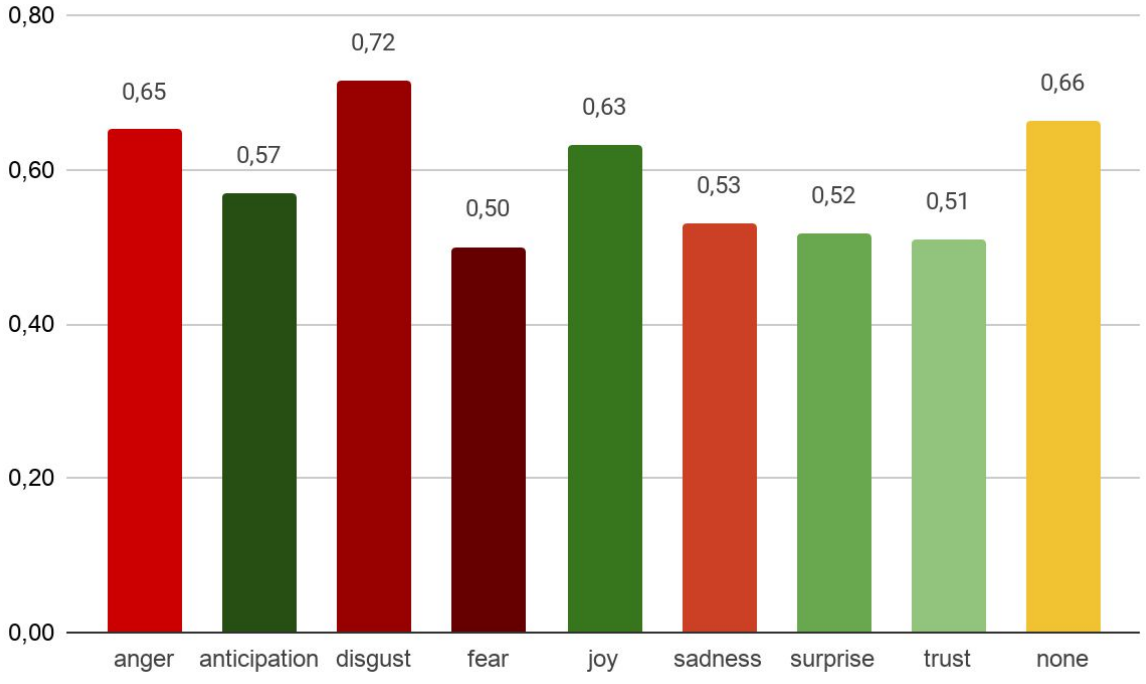

Figure 13: The average **AUROC** of all the systems per emotion.

Systems	Evaluation Sets	Emotions								
		anger	**anticipation**	**disgust**	**fear**	**joy**	**sadness**	**surprise**	**trust**	**none**
XLM-R@SE(w/o)	**ES**	0,39	0,09	0,58	0,00	0,08	0,04	0,00	0,00	0,33
	ES w/o	0,51	**0,37**	0,70	0,00	0,24	**0,09**	0,00	0,00	-
XLM-R@SE(w/o)-**Art**	**ES**	0,31	0,08	0,31	0,00	**0,27**	0,00	**0,18**	0,00	0,81
	ES w/o	0,32	0,08	0,31	0,00	**0,27**	0,00	**0,18**	0,00	-
XLM-R@SE(w/o)-**Palo**	**ES**	0,54	0,10	0,86	0,00	0,23	0,00	0,00	**0,12**	**0,91**
	ES w/o	**0,56**	0,10	**0,91**	0,00	0,23	0,00	0,00	**0,12**	-

Figure 14: The **F1 score** for the systems XLM-R@SE(w/o), XLM-R@SE(w/o)-Art, XLM-R@SE(w/o)-Palo per emotion on the ES, ES w/o evaluation datasets.

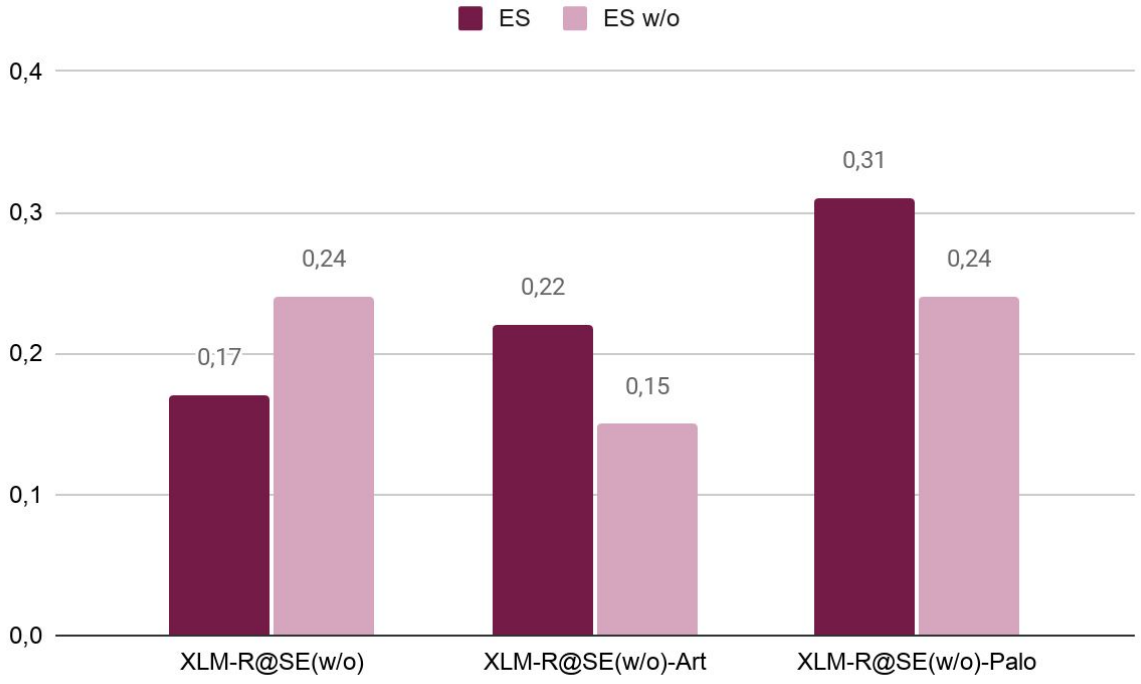

Figure 15: The average **F1 score** of all the emotions per system on the ES and ES-w/o.

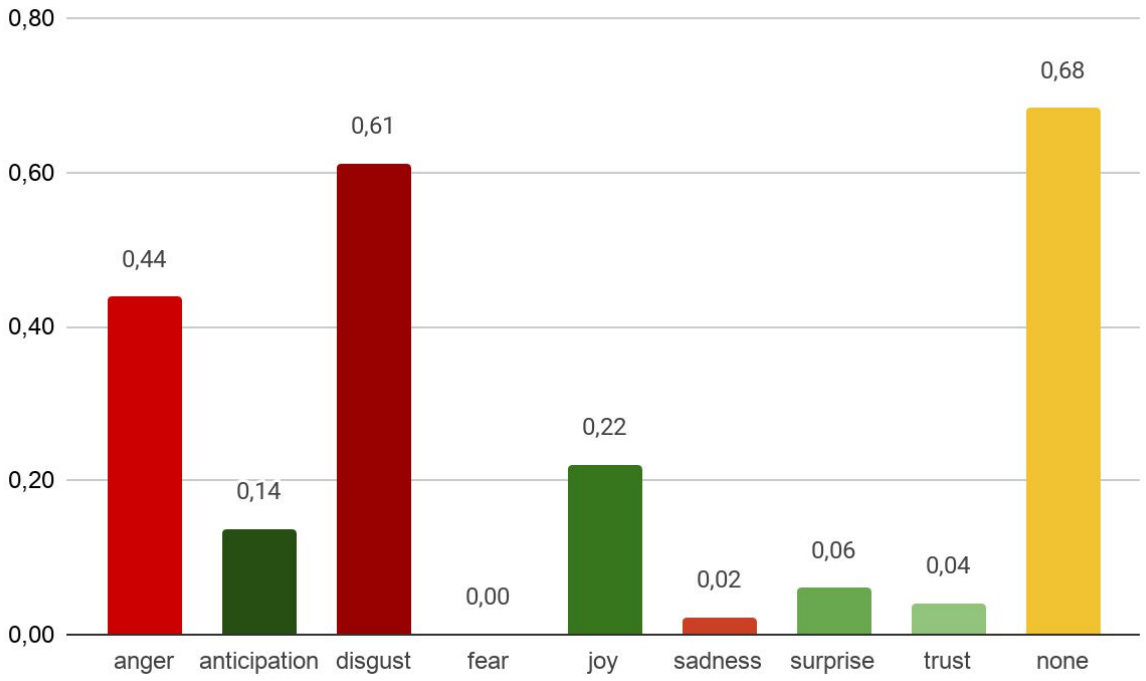

Figure 15: The average **F1 score** of all the systems per emotion.

4.9 Final Experiments

In subsequent experiments, we used the *SemEval-En* ***(SE)*** instead of *SemEval(w/o)-En (Section 3.3)* and we evaluated them only on the *ES* dataset which we used as a gold benchmark. Also, all the baseline models *(Section 3.3)* were trained with both *Artificial-Gr* ***(Art)*** and *Palo-Gr* ***(Palo)*** datasets. Following the previous symbolism, and representing the *Palo-Gr+* as ***Palo+***, *Random Forests* as ***RF***, *Logistic Regression* as ***LR***, our experiments were:

- XLM-R@SE
- XLM@SE-Art
- XLM-R@SE-Art-Palo
- XLM-R@SE-Palo
- XLM-R@SE-Palo+
- XLM-R@Palo
- XLM-R@Palo+
- RF@Palo
- RF@Art
- LR@Palo
- LR@Art
- KNN@Palo
- KNN@Art

4.10 Results

4.10.1 Emotion Classification

Figure 16 presents the AUPRC of all the final experiments, for each category/emotion, on the *ES* dataset. *Figure 17* illustrates the average AUPRC of all the emotions per system on the ES and *Figure 18* illustrates the average AUPRC of all the systems per emotion.

	Emotion								
Systems	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
XLM-R@SE	0,28	0,07	0,64	0,01	0,08	0,01	0,01	0,02	0,79
XLM@SE-Art	0,19	0,07	0,38	0,01	0,16	0,01	0,01	0,04	0,63
XLM-R@SE-Art-Palo	0,22	0,07	0,75	0,01	**0,31**	0,01	0,01	0,02	0,82
XLM-R@SE-Palo	**0,31**	0,03	0,78	0,01	0,11	0,01	0,01	0,08	**0,87**
XLM-R@SE-Palo+	0,20	**0,17**	**0,79**	0,01	0,19	0,01	0,01	**0,14**	0,86
XLM-R@Palo	0,21	0,03	0,72	0,01	0,02	0,01	0,01	0,02	0,82
XLM-R@Palo+	0,23	0,03	0,74	0,01	0,15	0,01	0,01	0,08	0,85
RF@Palo	0,11	0,03	0,41	0,01	0,09	0,01	0,01	0,05	0,64
RF@Art	0,12	0,03	0,32	0,01	0,02	0,01	0,01	0,02	0,61
LR@Palo	0,17	0,03	0,59	0,01	0,02	0,01	0,01	0,02	0,74
LR@Art	0,15	0,03	0,34	0,01	0,02	0,01	0,01	0,03	0,63
KNN@Palo	0,11	0,03	0,38	0,01	0,02	0,01	0,01	0,02	0,64
KNN@Art	0,12	0,03	0,32	0,01	0,02	0,01	0,01	0,02	0,64

Figure 16: The **AUPRC** of all the systems per emotion on the ES evaluation dataset.

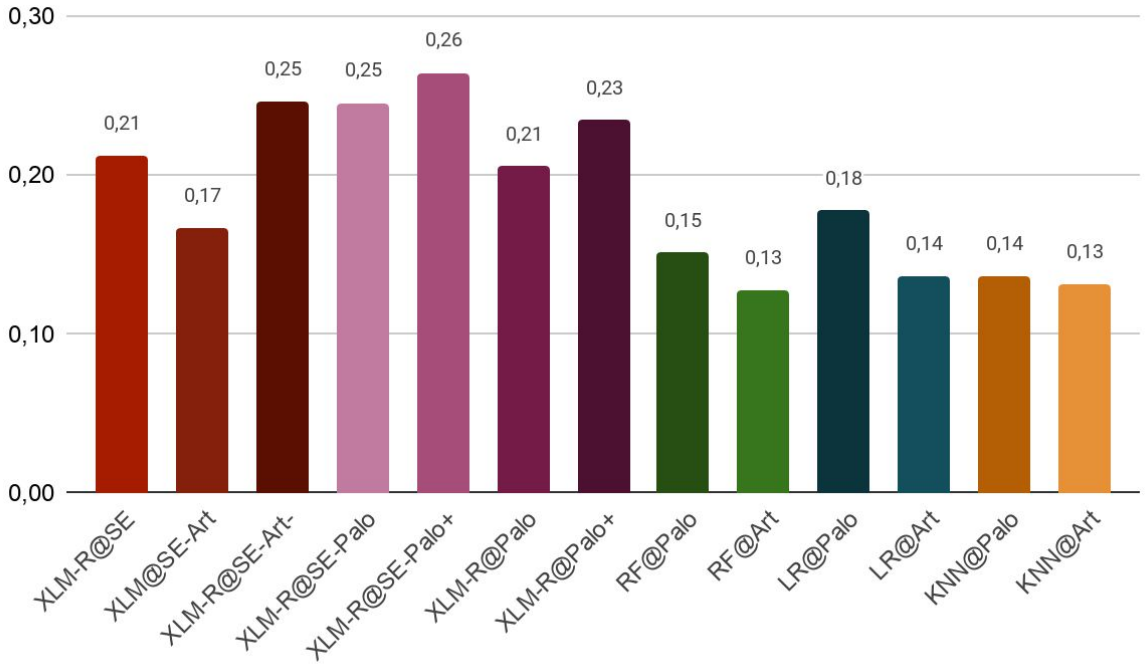

Figure 17: The average **AUPRC** of all the emotions per system on the ES.

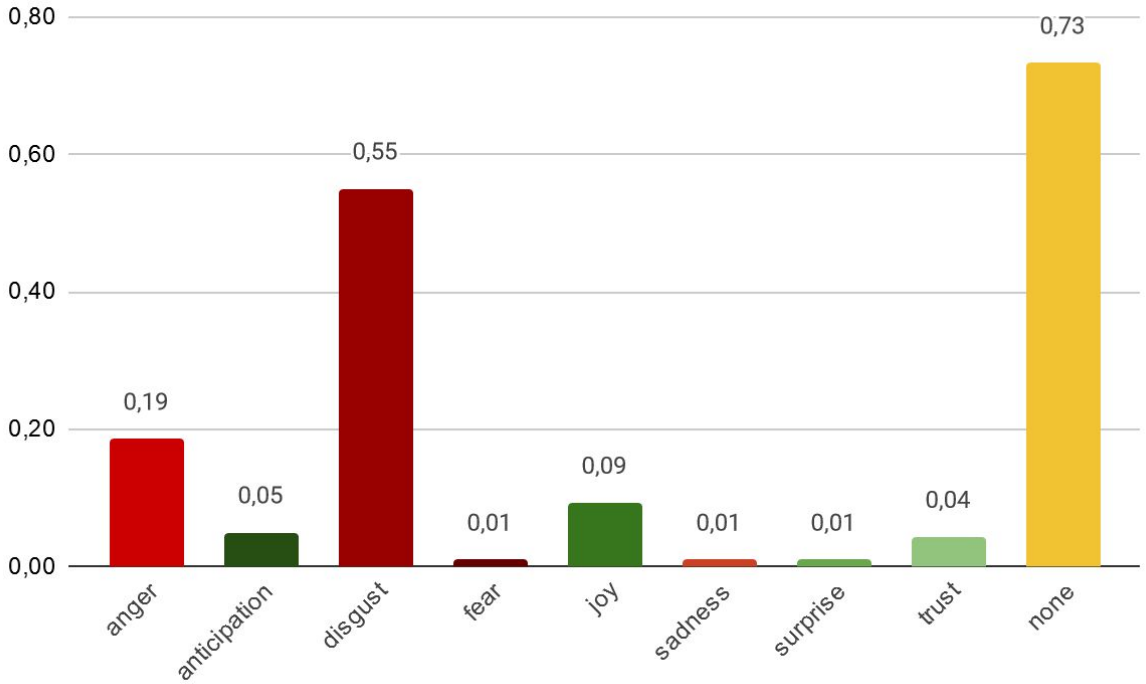

Figure 18: The average **AUPRC** of all the systems per emotion.

As we can observe from the above figures, for the categories *disgust* and *none* we had the best results, and for the categories *fear*, *sadness* and *surprise* none of the models seemed to work. If we look more precisely at the results of each model we can observe the follows.

- The results from the **XLM-R@SE** system seems to be quite impressive compared to the other models, especially if we consider that the model was trained on English corpus and evaluated on a greek dataset (zero-shot learning). From these results, we conclude that with this specific method, for the categories *disgust* and *none* we do not even need labeled data.
- Fine-tuning our model with the *Artificial-Gr (Art)* dataset did not only improve the results but it made them worse in most categories. An exemption is the category *joy* whose results improved using the *Artificial-Gr (Art)* dataset through the method **XLM-R@SE-Art-Palo**. Probably the results of the **XLM-R@SE-Art** are worse than the results of the **XLM-R@SE** since the dataset was constructed from scratch by retrieving tweets and it wasn't labeled by professional annotators. This led to several problems such as the fact that the model had learned to classify tweets with very specific words to the corresponding categories. For example, most tweets from the *Artificial-Gr (Art)* dataset which belonged to the class *joy* contained the word '*χαίρομαι*', as a result, the model learned that the comments that belong to the category *joy* are those that contain the specific word.
- In contrast with the *Artificial-Gr (Art)* dataset, the *Palo-Gr (Palo)* dataset was very beneficial for the improvement of our results. In all categories, the best performance was achieved by models that were fine-tuned on *Palo-Gr (Palo)* or *Palo-Gr+ (Palo+)* dataset. *Palo-Gr+ (Palo+)* dataset, which had higher support of positive tweets than *Palo-Gr (Palo)*, improved the results for positive categories such as *anticipation* and *trust*.
- In most cases, all of our models outperformed the baseline ones. Furthermore, the baselines that were trained with the *Palo-Gr (Palo)* outperformed the same baselines that were trained with the *Artificial-Gr (Art)*, which also shows the superiority of one dataset over the other.

Additionally, in *Figures 18, 19*, *20* and in *Figures 21, 22, 23* we present the F1 score and the AUROC of all the systems respectively.

	Emotion								
Systems	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
XLM-R@SE	0.46	0.19	0.73	0.00	0.17	**0.04**	0.00	0.00	0.80
XLM@SE-Art	0.27	0.22	0.20	0.00	0.25	0.00	0.00	0.15	0.75
XLM-R@SE-Art-Palo	0.38	0.15	0.82	0.00	**0.50**	0.00	0.00	0.00	0.90
XLM-R@SE-Palo	**0.51**	0.00	**0.86**	0.00	0.24	0.00	0.00	0.11	0.91
XLM-R@SE-Palo+	0.34	**0.38**	0.85	0.00	0.42	0.00	0.00	**0.35**	**0.92**
XLM-R@Palo	0.38	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.89
XLM-R@Palo+	0.40	0.00	0.82	0.00	0.38	0.00	0.00	0.26	0.91
RF@Palo	0.00	0.00	0.26	0.00	0.13	0.00	0.00	0.11	0.78
RF@Art	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.76
LR@Palo	0.23	0.07	0.64	0.00	0.08	0.00	0.00	0.06	0.84
LR@Art	0.13	0.00	0.06	0.00	0.00	0.00	0.00	0.07	0.77
KNN@Palo	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.77
KNN@Art	0.04	0.10	0.02	0.00	0.00	0.00	0.00	0.05	0.69

Figure 21: The **F1 score** of all the systems per emotion on the ES evaluation dataset.

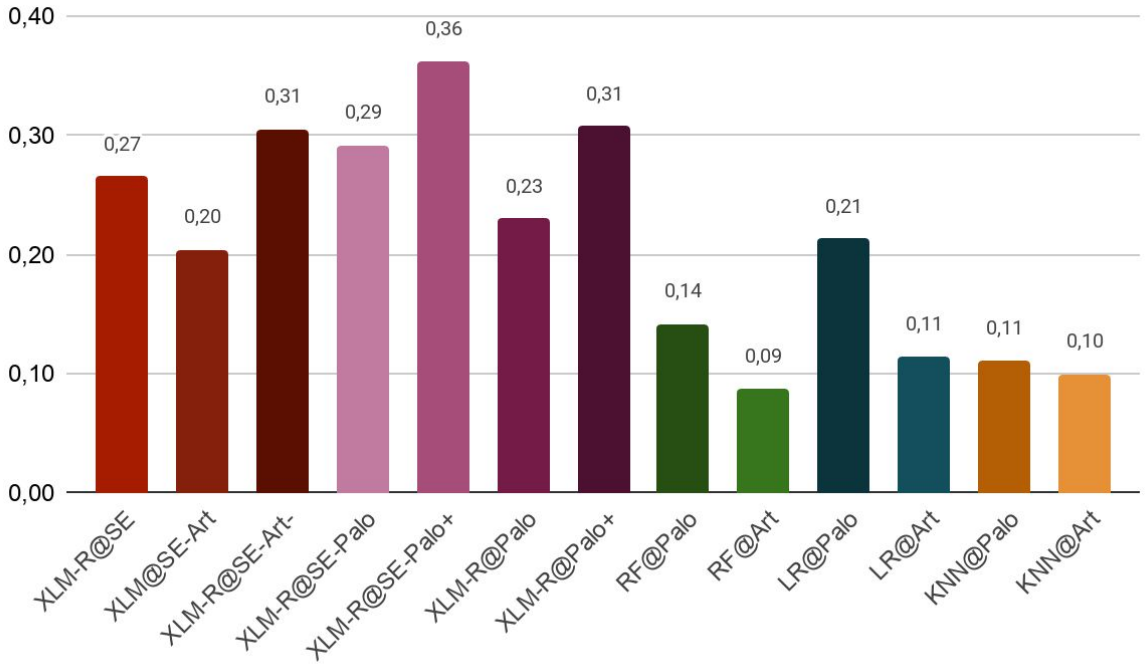

Figure 22: The average **F1 score** of all the emotions per system on the ES.

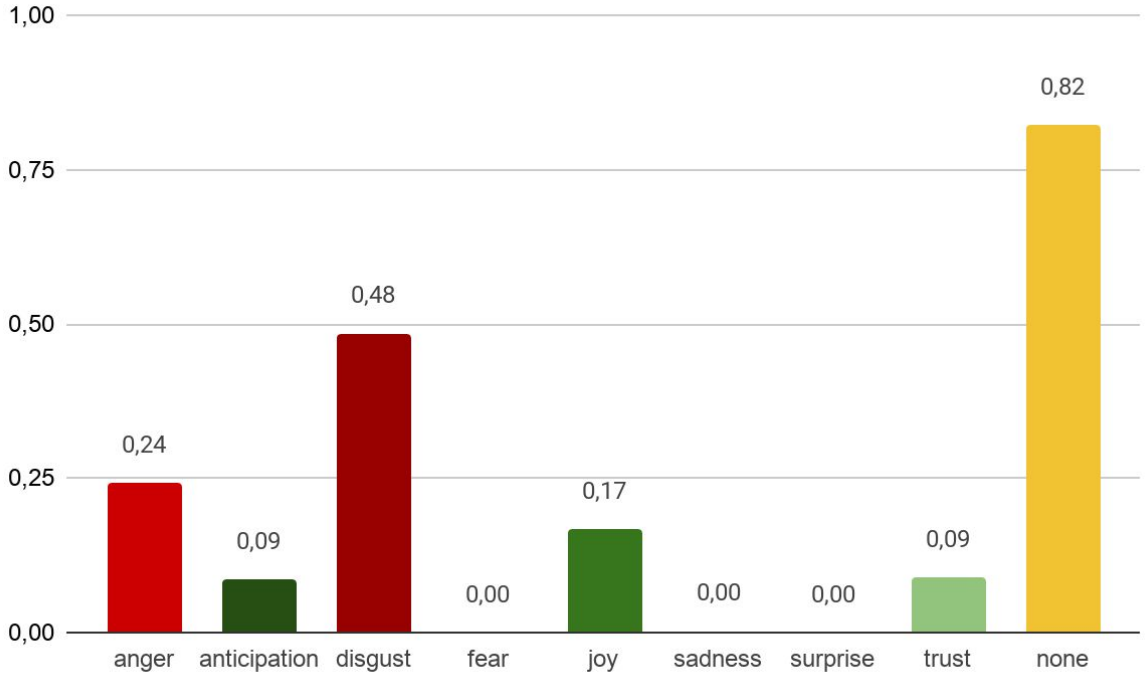

Figure 23: The average **F1 score** of all the systems per emotion.

	Emotion								
Systems	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	none
XLM-R@SE	**0.79**	**0.66**	0.80	0.50	**0.83**	**0.57**	0.50	0.50	0.77
XLM@SE-Art	0.58	0.58	0.55	0.50	0.57	0.50	0.50	0.60	0.55
XLM-R@SE-Art-Palo	0.64	0.54	0.86	0.50	0.68	0.50	0.50	0.50	0.83
XLM-R@SE-Palo	0.73	0.50	**0.91**	0.50	0.57	0.50	0.50	0.53	0.88
XLM-R@SE-Palo+	0.62	0.65	0.88	0.50	0.68	0.50	0.50	**0.70**	**0.88**
XLM-R@Palo	0.64	0.50	0.85	0.50	0.50	0.50	0.50	0.50	0.83
XLM-R@Palo+	0.66	0.50	0.86	0.50	0.71	0.50	0.50	0.64	0.86
RF@Palo	0.50	0.50	0.57	0.50	0.54	0.50	0.50	0.53	0.56
RF@Art	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.51
LR@Palo	0.57	0.52	0.73	0.50	0.53	0.50	0.50	0.52	0.74
LR@Art	0.53	0.50	0.52	0.50	0.50	0.50	0.50	0.52	0.54
KNN@Palo	0.50	0.50	0.56	0.50	0.50	0.50	0.50	0.49	0.56
KNN@Art	0.51	0.55	0.50	0.48	0.46	0.47	0.50	0.52	0.56

Figure 24: The **AUROC** of all the systems per emotion on the ES evaluation dataset.

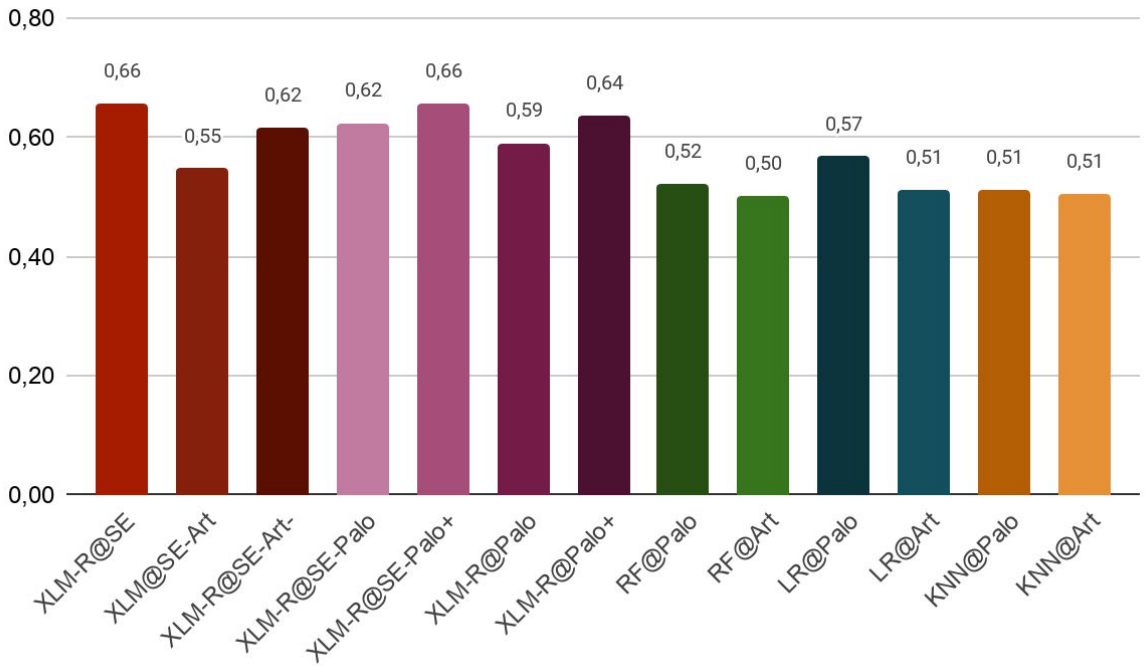

Figure 25: The average **AUROC** of all the emotions per system on the ES.

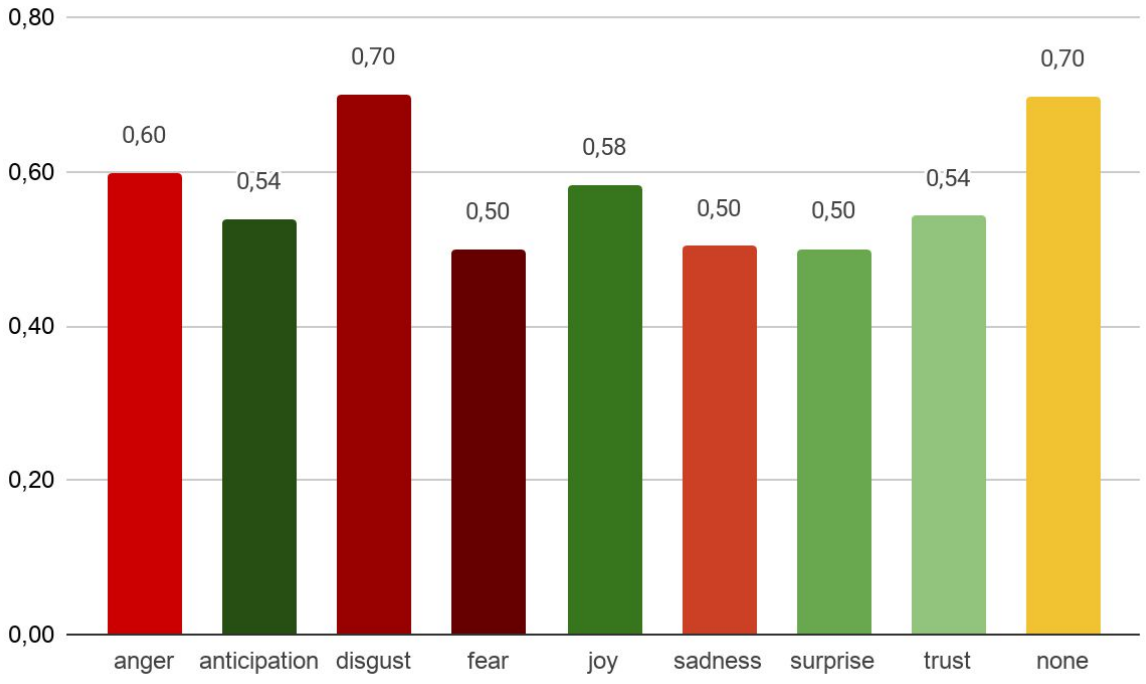

Figure 26: The average **AUROC** of all the systems per emotion.

4.10.2 Sentiment Classification

By using sentiment aggregation we also assessed our models *(Section 4.9)* for sentiment analysis i.e for the categories: neutral (*none*), positive (*joy, trust, anticipation, surprise*), and negative (*anger, disgust, fear, sadness*).

Figure 27 presents the AUPRC of all the final experiments, for each category/emotion, on the *ES* dataset for the classification task. *Figure 28* illustrates the average AUPRC of all the emotions per system on the ES and *Figure 29* illustrates the average AUPRC of (i) all the systems (*Systems(All)*) and (ii) of the systems except the baselines *(Systems(w/o-bas)* per emotion, also for the sentiment task.

Systems	**Negative**	**Positive**	**Neutral**
XLM-R@SE	0.66	0.15	0.79
XLM@SE-Art	0.38	0.09	0.63
XLM-R@SE-Art-Palo	0.76	0.15	0.82
XLM-R@SE-Palo	**0.80**	0.12	**0.87**
XLM-R@SE-Palo+	0.79	**0.22**	0.86
XLM-R@Palo	0.73	0.07	0.82
XLM-R@Palo+	0.76	0.12	0.85
RF@Palo	0.42	0.08	0.64
RF@Art	0.33	0.07	0.61
LR@Palo	0.59	0.08	0.74
LR@Art	0.35	0.07	0.63
KNN@Palo	0.39	0.08	0.64
KNN@Art	0.33	0.07	0.64

Figure 27: The **AUPRC** of all the systems per emotion on the ES evaluation dataset, for the sentiment task.

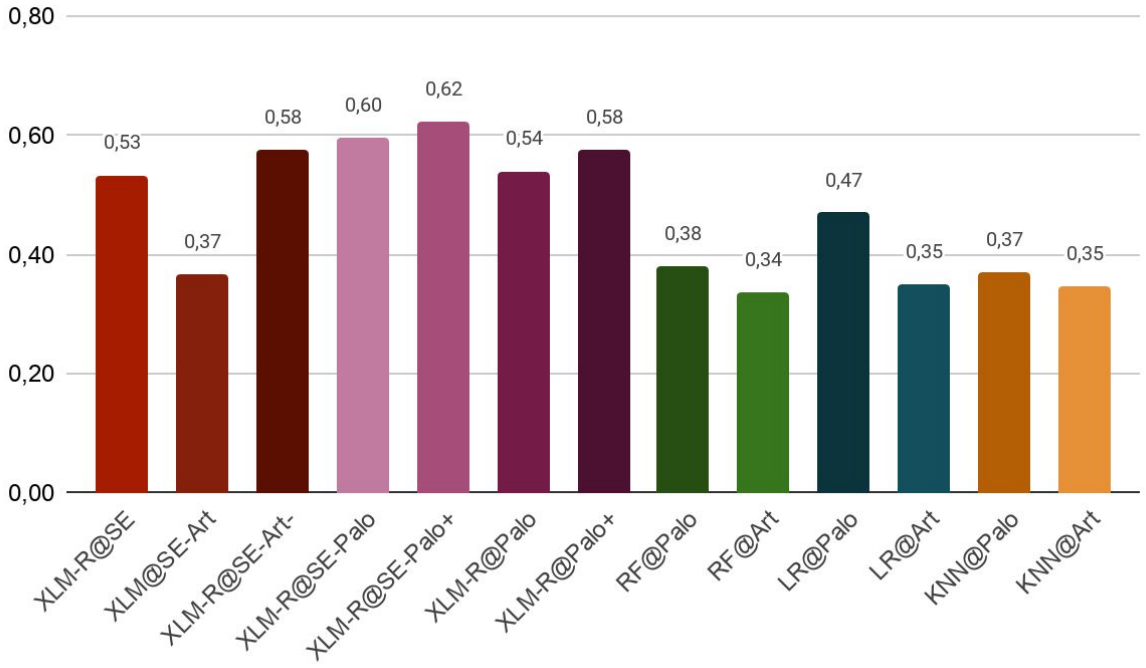

Figure 28: The average **AUPRC** of all the emotions per system on the ES, for the sentiment task.

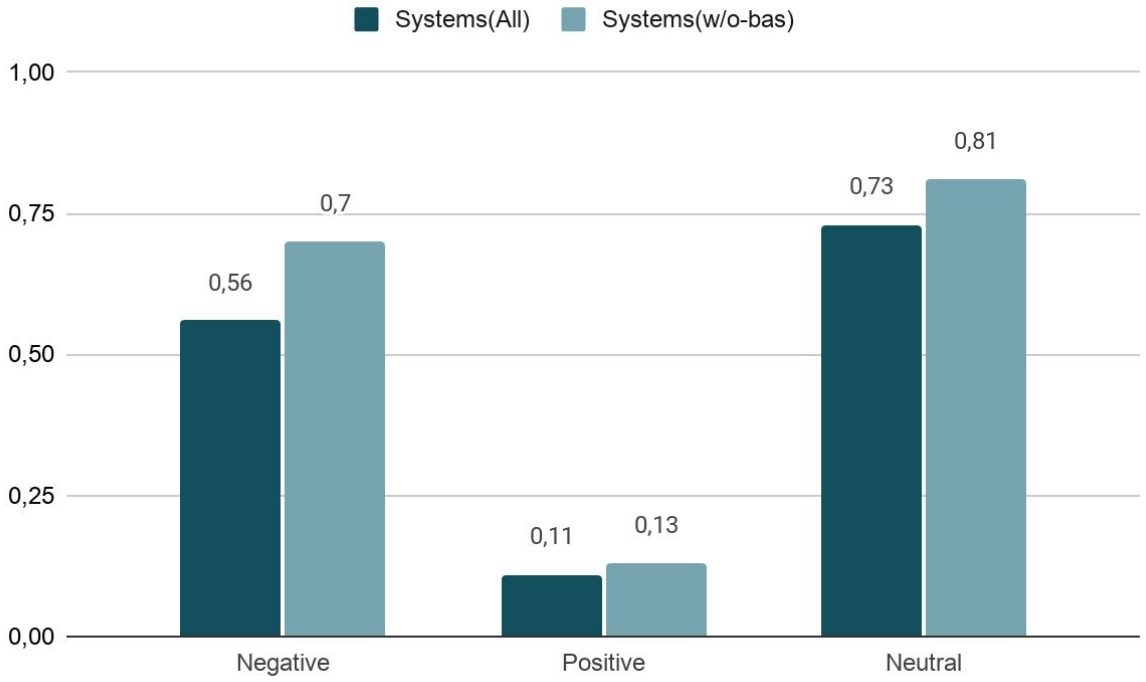

Figure 29: The average **AUPRC** of all the systems and the systems without the baselines per emotion,for the sentiment task.

As the results of the emotion classification had predisposed us, in the negative and neutral categories we have much better results than in the positive category. Since we implement sentiment aggregation, our observations for the models' behavior are the same as before. **XLM-R@SE**'s performance is really impressive especially if we consider that we used zero-shot learning. Also, models that were fine-tuned with the *Palo-Gr (Palo)* dataset achieved higher performance in the negative and neutral categories. In the positive category, the model which was fine-tuned with the *Palo-Gr+ (Palo+)* had the best results.
Additionally, in *Figures 30, 31*, *32* and in *Figures 33, 34, 35* we present the F1 score and the AUROC of all the systems for the sentiment task respectively.

Systems	**Negative**	**Positive**	**Neutral**
XLM-R@SE	0.74	0.31	0.80
XLM@SE-Art	0.21	0.18	0.75
XLM-R@SE-Art-Palo	0.82	0.22	0.90
XLM-R@SE-Palo	**0.87**	0.11	0.91
XLM-R@SE-Palo+	0.84	**0.41**	**0.92**
XLM-R@Palo	0.83	0.00	0.88
XLM-R@Palo+	0.83	0.23	0.91
RF@Palo	0.25	0.04	0.78
RF@Art	0.01	0.00	0.76
LR@Palo	0.63	0.10	0.84
LR@Art	0.09	0.06	0.77
KNN@Palo	0.22	0.09	0.77
KNN@Art	0.19	0.13	0.69

Figure 30: The **F1 score** of all the systems per emotion on the ES evaluation dataset, for the sentiment task.

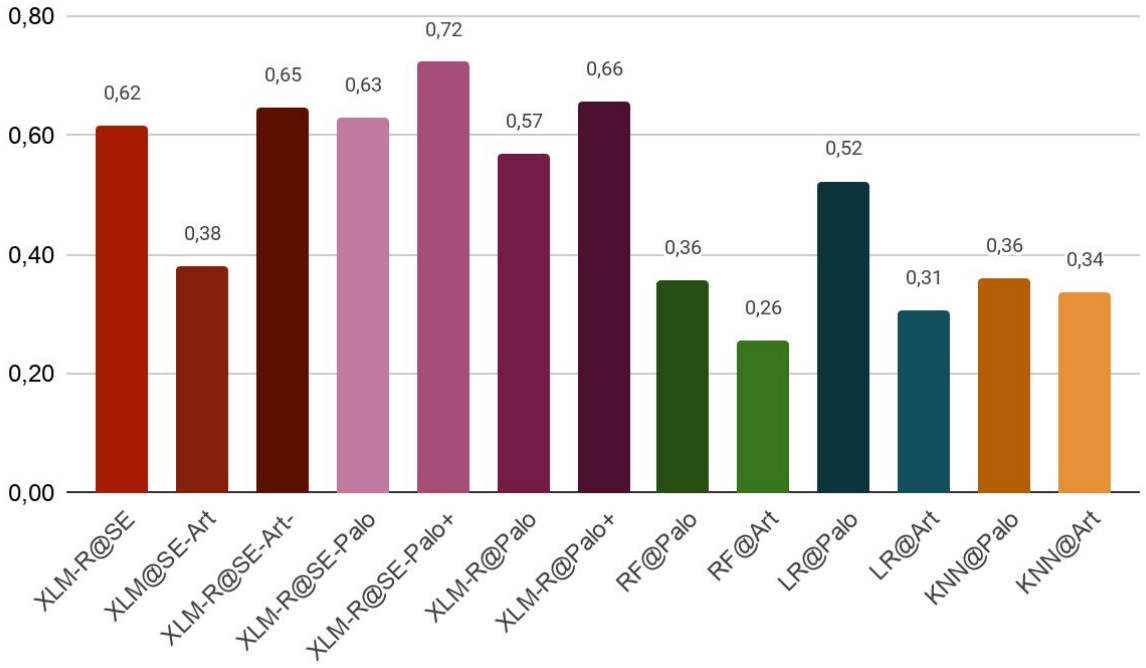

Figure 31: The average **F1 score** of all the emotions per system on the ES, for the sentiment task.

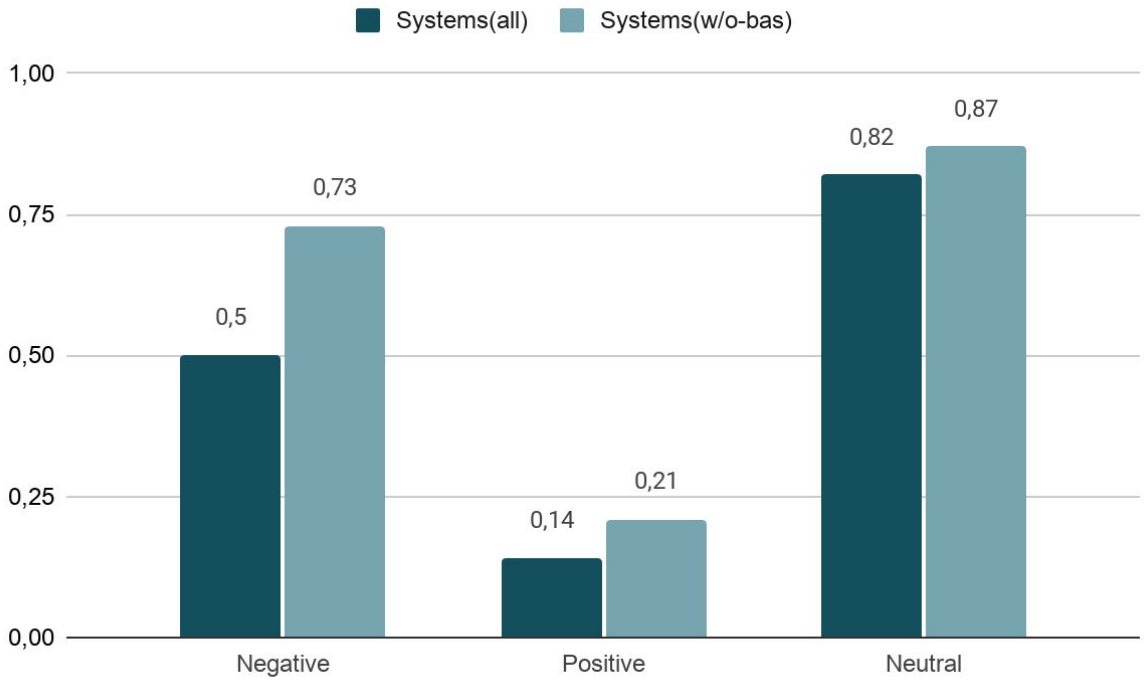

Figure 32: The average **F1 score** of all the systems and the systems without the baselines per emotion, for the sentiment task.

Systems	Negative	Positive	Neutral
XLM-R@SE	0.81	**0.70**	0.77
XLM@SE-Art	0.55	0.56	0.55
XLM-R@SE-Art-Palo	0.86	0.56	0.83
XLM-R@SE-Palo	**0.91**	0.53	**0.88**
XLM-R@SE-Palo+	0.87	0.66	**0.88**
XLM-R@Palo	0.86	0.50	0.83
XLM-R@Palo+	0.87	0.57	0.86
RF@Palo	0.57	0.51	0.56
RF@Art	0.50	0.50	0.51
LR@Palo	0.73	0.52	0.74
LR@Art	0.52	0.51	0.54
KNN@Palo	0.56	0.52	0.56
KNN@Art	0.51	0.53	0.56

Figure 33: The **AUROC** of all the systems per emotion on the ES evaluation dataset, for the sentiment task.

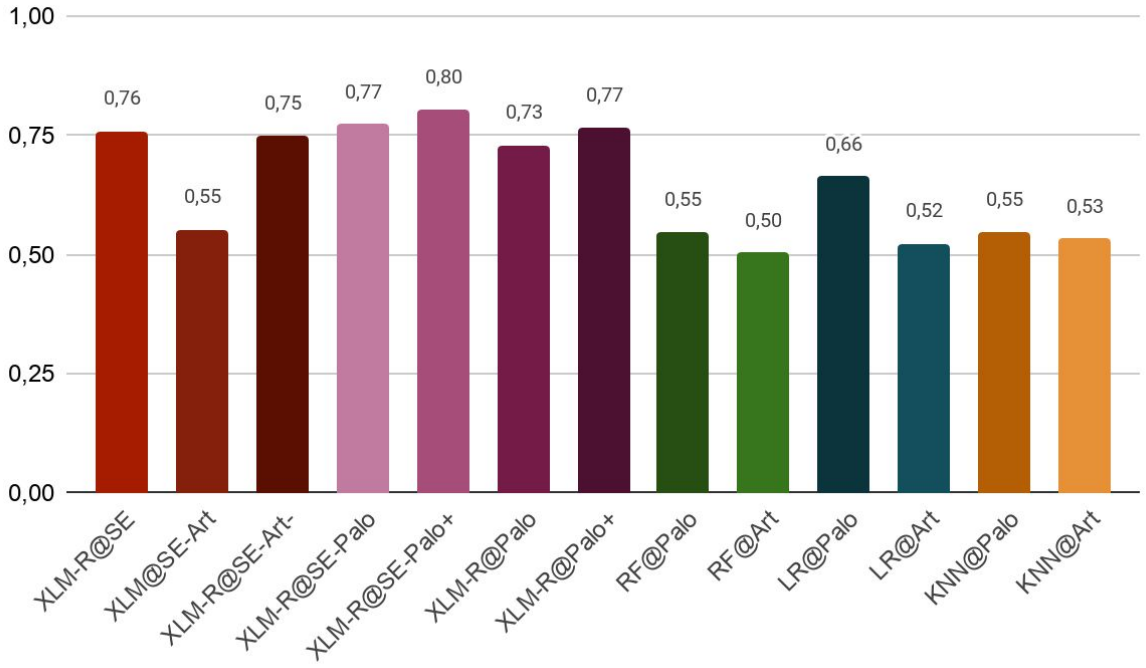

Figure 34: The average **AUROC** of all the emotions per system on the ES, for the sentiment task.

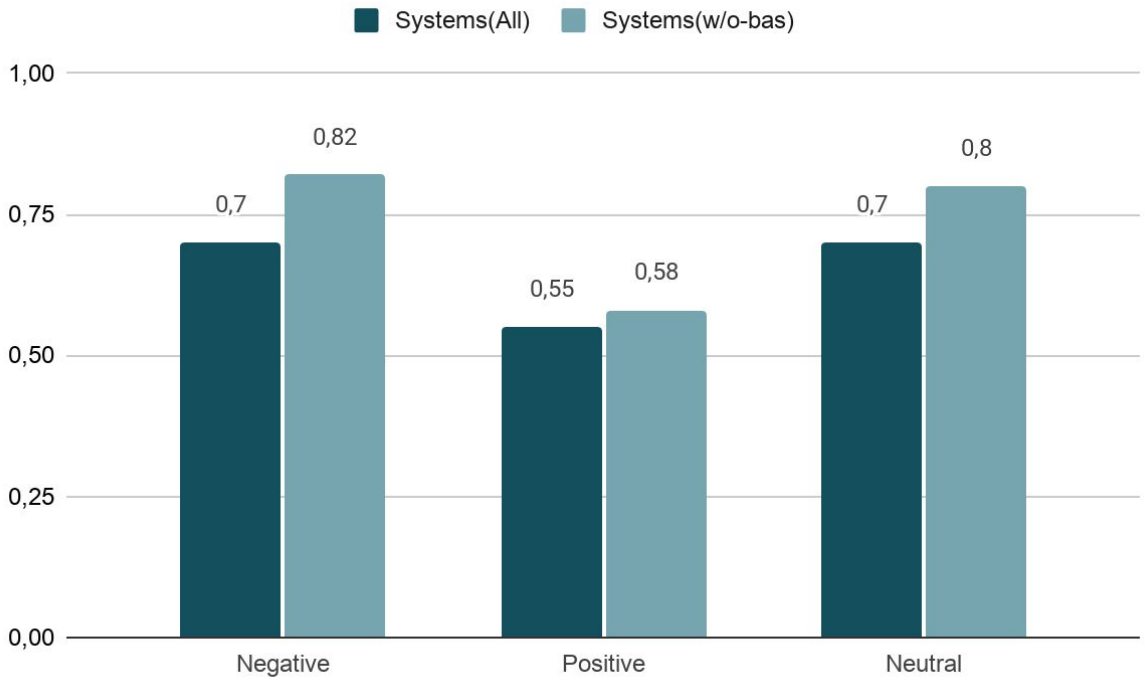

Figure 35: The average **AUROC** of all the systems and the systems without the baselines per emotion, for the sentiment task.

5. Conclusion

In this thesis, we discussed and we compared models to tackle the task of emotion classification in the Greek language. We used zero-shot and transfer learning from a high-resource language (English) to a low-resource language (Greek) with the pre-trained state-of-the-art multilingual transformer XLM-R. Fine-tuning was achieved through several datasets that correspond to the multi-label classification task of detecting what types of emotion, if any, a tweet contains. Initially, through the zero-shot learning, we trained our model on an English dataset with tweets and then we tested it in Greek tweets that were developed for this work. Subsequently, we fine-tuned our transformer on Greek tweets alongside: i) a dataset (Palo-Gr) that we annotated and we constructed from scratch in collaboration with the company PaloServices and ii) an artificial dataset that we created by streaming tweets using Tweepy. Unlike the artificial dataset, the Palo-Gr dataset despite its small size, it really improved the results compared to zero-shot learning through transfer learning from the pre-trained model XLM-R. We also compared all our methods of training and fine-tuning the XLM-R with several baselines. We present the performance of all the models on the emotion task and then for sentiment analysis by aggregating the results of the emotion analysis. As our results show, this method is highly promising for tasks that tackle low-resource language such as Greek, where few or even any labeled data are available.

5.1 Future Work

As future work, we plan to fine-tune the XLM-R on a new dataset, probably the already existing Palo+positives dataset enriched with more tweets that correspond to the categories with the lower results. For example, since the results in the category *fear* are poor which is related to the low support of the particular category in our dataset, the new one will be enriched with more tweets that correspond in the category *fear*.

Secondly, we intend to compare our models with the GREEK-BERT introduced by Koutsikakis, Chalkidis, Malakasiotis and Androutsopoulos et al. (2020), a monolingual pre-trained Transformer-based model for Greek language, similar to BERT, trained on 29 GB of Greek text with a 35k sub-word BPE vocabulary created from scratch.

Bibliography

Alexis Conneau and Guillaume Lample (2019). “Cross-lingual Language Model Pretraining”.
URL: https://arxiv.org/pdf/1901.07291.pdf

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). “Unsupervised Cross-lingual Representation Learning at Scale”.
URL: https://arxiv.org/pdf/1911.02116.pdf

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew (2019). “Huggingface’s transformers: State-of-the-art natural language processing”.
URL: https://arxiv.org/pdf/1910.03771.pdf

Mohammad, S. M. Bravo-Marquez, F. Salameh, M., and Kiritchenko, S. (2018). “ Semeval-2018 Task 1: Affect in tweets”. *In Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*.
URL: https://www.aclweb.org/anthology/S18-1001.pdf

Abrhalei Tela, Abraham Woubie, Ville Hautamaki (2020). “Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya”.
URL: https://arxiv.org/pdf/2006.07698.pdf

Tharindu Ranasinghe, Marcos Zampieri (2020). “Multilingual Offensive Language Identification with Cross-lingual Embeddings”.
URL: https://www.aclweb.org/anthology/2020.emnlp-main.470.pdf

Neel Kant, Raul Puri,Nikolai Yakovenko, Bryan Catanzaro (2018). “Practical Text Classification With Large Pre-Trained Language Models”.
URL: https://arxiv.org/pdf/1812.01207.pdf

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). “Attention Is All You Need”. *In 31st Annual Conference on Neural Information Processing Systems.USA*
URL: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). “ Bert: Pre-training of deep bidirectional transformers for language understanding”
URL: https://arxiv.org/pdf/1810.04805.pdf

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized Bert pretraining approach".
URL: https://arxiv.org/pdf/1907.11692.pdf

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, Ion Androutsopoulos (2020). "GREEK-BERT: The Greeks visiting Sesame Street"
URL: https://arxiv.org/abs/2008.12014

Plutchik, R. (1979). Emotions: A general Psychoevolutionary theory. 1
URL: https://link.springer.com/referenceworkentry/10.1007%2F978-3-319-28099-8_547-1

Saif M. Mohammad and Svetlana Kiritchenko, Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories
URL: http://saifmohammad.com/WebDocs/lrec2018-paper-tweet-emotion.pdf

Debora Nozza, Federico Bianchi, and Dirk Hovy (2020). "What the[MASK]? Making Sense of Language-Specific BERT Models".
URL: https://arxiv.org/pdf/2003.02912.pdf

Telmo Pires, Eva Schlinger, and Dan Garrette (2019). " How Multilingual is Multilingual BERT?".
URL: https://arxiv.org/pdf/1906.01502.pdf

Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382

Claudia Kittask, Kirill Milintsevich, and Kairit Sirts (2020). "Evaluating multilingual BERT for Estonian".
URL: https://arxiv.org/pdf/2010.00454.pdf

Michael A. Hedderich, David I. Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow (2020). "Transfer Learning and Distant Supervision for Multilingual TransformerModels: A Study on African Languages"
URL: https://arxiv.org/pdf/2010.03179.pdf

K. K, Z. Wang, S. Mayhew, and D. Roth (2019). "Cross-lingual ability of multilingual Bert: An empirical study."
URL: https://arxiv.org/pdf/1912.07840.pdf

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas (2020). " From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers"
URL: https://arxiv.org/pdf/2005.00633.pdf

Zihan Wang, Karthikeyan K, Stephen Mayhew, Dan Rot (2020). “Extending Multilingual BERT to Low-Resource Languages”
URL: https://arxiv.org/pdf/2004.13640.pdf

Laura-Ana-Maria Bostan,Roman Klinger (2018) . “An Analysis of Annotated Corpora for Emotion Classification in Text”
URL: https://www.aclweb.org/anthology/C18-1179.pdf

Ian D. Wood, John P. McCrae, Vladimir Andryushechkin, Paul Buitelaar (2018). “A Comparison Of Emotion Annotation Schemes And A New Annotated DataSet”
URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/61.pdf

Bharat Gaind, Varun Syal, Sneha Padgalwar (2019). “Emotion Detection and Analysis on Social Media”
URL: https://arxiv.org/pdf/1901.08458.pdf

Shrey Desai, Cornelia Caragea, Junyi Jessy Li (2020). “ Detecting Perceived Emotions in Hurricane Disasters”
URL: https://arxiv.org/pdf/2004.14299.pdf

Jansen, Zhang (2009). “Twitter Power: Tweets as Electronic Word of Mouth”
URL: (PDF) Twitter Power: Tweets as Electronic Word of Mouth