A STUDY OF TRANSFORMER MODELS FOR EMOTION CLASSIFICATION IN INFORMAL TEXT

by

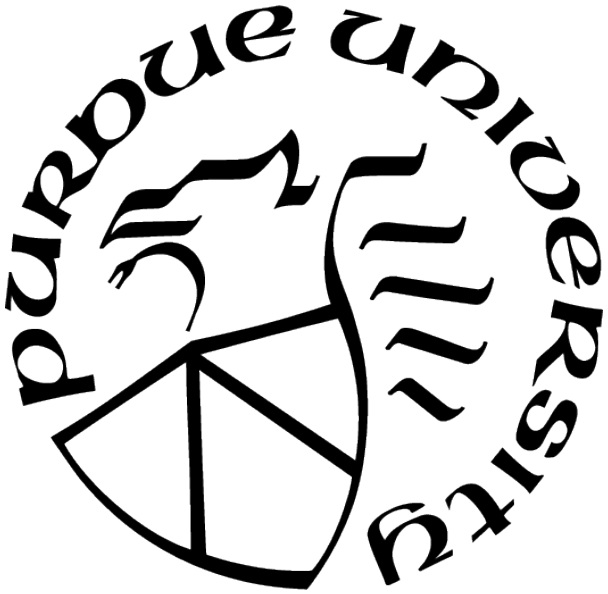
Alvaro Soares de Boa Esperanca

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science



Department of Electrical and Computer Engineering

Indianapolis, Indiana

December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Brian King, Co-Chair

Department of Electrical and Computer Engineering

Dr. Xiao Luo, Co-Chair

Department of Computer Information and Graphics Technology

Dr. Zhenming Ding

Department of Computer Information and Graphics Technology

Approved by:

Dr. Brian King

To Elizabeth *Liz* Meyers.

ACKNOWLEDGMENTS

Firstly, I would like to thank God for all the blessings and challenges that helped me grow through the process of developing this Thesis.

I want to thank Dr. Xiao Luo for her guidance and support throughout the development of this Thesis and her patience with me whenever I would hit a wall and did not know how to proceed. I would also like to thank Dr. Brian King for all continuous support throughout my entire academic career and the advice on moving forward on my journey to become a better engineer. I thank Dr. Zhenming Ding for being a part of my thesis committee and providing critical feedback on my work. Furthermore, I thank Sherrie Tucker for all the help and support and for ensuring I was always on track with my deadlines.

I want to thank my family and friends for all the love and support that helped me throughout the way. I want to extend a special thanks to my colleagues Mohammad Al-Merri, Nathaniel Cantwell especially, and Mauricio Ambrosio. My journey through graduate school was made much more interesting because of our collaboration and sharing of ideas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bi-directional Long Short Term Memory
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
SemEval	Semantic Evaluation
SOTA	State-of-the-art

ABSTRACT

Textual emotion classification is a task in affective AI that branches from sentiment analysis and focuses on identifying emotions expressed in a given text excerpt. It has a wide variety of applications that improve human-computer interactions, particularly to empower computers to understand subjective human language better. Significant research has been done on this task, but very little of that research leverages one of the most emotion-bearing symbols we have used in modern communication: Emojis. In this thesis, we propose several transformer-based models for emotion classification that processes emojis as input tokens and leverages pretrained models and uses them , a model that processes Emojis as textual inputs and leverages DeepMoji to generate affective feature vectors used as reference when aggregating different modalities of text encoding. To evaluate ReferEmo, we experimented on the SemEval 2018 and GoEmotions datasets, two benchmark datasets for emotion classification, and achieved competitive performance compared to state-of-the-art models tested on these datasets. Notably, our model performs better on the underrepresented classes of each dataset.

1. INTRODUCTION

Sentiment analysis is the branch of affective AI consisting of various methodologies for identifying emotional valence expressed in text. Over the past years, there has been an increase in the prevalence of sentiment analysis in research and industry. Use cases include identifying customer satisfaction by inferring the sentiment being expressed in product reviews [1] and determining job satisfaction from Voice of Employee surveys [2] and monitoring the emotional state of a large population by inferring the sentiment expressed in public communication platforms, particularly in the case of significant events such as political elections [3] and major health crises [4]. Identifying early signs of mental health conditions by identifying the sentiment expressed in published online content. Due to the increasing popularity of Emojis, there is great interest in analyzing and studying their usage in text content for sentiment analysis [5]–[7].

Despite the wide range of use cases and their successful outcomes, the interpretative capabilities of sentiment analysis are still limited. By definition, sentiment analysis predicts whether a text excerpt expresses sentiment ranging from positive to negative, often including neutral sentiment as the midpoint. That alone does not provide much actionable information or context regarding the type of emotion being expressed. For instance, sadness, disappointment, and anger are negative emotions, but they describe very different concepts.

Emotion analysis, also called Emotion Classification, expands on sentiment analysis by providing more meaningful and contextual predictions. While sentiment analysis predicts positive or negative valence, emotion analysis predicts whether a text excerpt expresses any discrete emotional states. Some of these emotions include joy, love, sadness, and anger. These provide more meaningful and actionable outcomes and provide a framework for understanding how humans express affect in text, thus providing researchers with the insight to teach machines to understand emotions better.

This thesis investigates multiple deep learning models that incorporate emoji processing in their architecture. We test these models on benchmark multilabel emotion classification datasets that use emojis and test these models on datasets that feature little to no emojis. The contribution of this thesis consists of novel emotion classification model architectures

that achieve competitive performance with SOTA models on the emotion classification task and a study of the effects of using emojis as input tokens in these same models. The remainder of this thesis is organized as follows: Chapter 2 summarizes the related work and introduces the foundational concepts for developing our models, Chapter 3 describes the datasets used in our experiments, Chapter 4 describes the architectures of the proposed models, Chapter 5 describes the experiments and its results, and finally Chapter 6 summarizes the main findings and outlines directions for future work.

2. BACKGROUND AND RELATED WORKS

The textual emotion or sentiment classification falls into two categories: lexicon-based approaches and machine learning-based approaches. The lexicon-based approaches utilize curated words and their associations to classify a text, whereas the machine learning-based approaches train a model to classify text. The defined lexicons are either used in rule-based models that rely on keyword frequency count [8] or keyword search [9] or as input features to machine learning-based models [10]. The machine learning-based approaches include those using either traditional or deep learning models. The traditional machine learning approach is similar to text classification, which includes steps as first to assign unicodes to the emojis, then apply feature selection, and classification algorithms, such as multinomial Naïve Bayes [11] for emotion classification or sentiment detection. With the advance of deep learning, the recent literature on emotion classification investigates various deep language models.

2.1 Lexicon-Based Emotion Classification

Affective lexicons have been extensively used since the early stages of affective AI research and still provide helpful linguistic features that aid in more contemporary methodologies. These lexicons usually consist of curated sets of words and their associated set of affect scores.

Lexicons like the Liu Lexicon [12] consists of a set of words with either a positive or a negative label with no measure of how positive or negative a word is, resulting in words having the same level of affect despite being inherently different (e.g., irate and annoyed are treated the same despite the prior being a more intense expression of anger than annoyed is). There are lexicons such as AFINN [13], and SentiWordNet [14] that do quantify how positive or how negative the word is, allowing for better distinction of expressed affect. However, it is limited to the valence score, i.e., whether a word is positive or negative in sentiment. Emotion classification models have used these lexicons, particularly TCS Research [10] and SeerNet [15], with some significant results.

EmoLex [16] is a lexicon that builds on top of the ones described prior and is better suited for the emotion classification task. This lexicon consists of words and binary labels

indicating the word's association with the primary emotion categories from the Ekman model of emotions (i.e., anger, sadness, contempt, disgust, surprise, and fear) and the word's association to either positive or negative sentiment. Much like Liu Lexicon, EmoLex does not quantify the association between a word and each emotion. These lexicons have also been used in emotion classification models such as NELEC [17], and SINAI [18].

Despite their prevalence, these lexicons still present some limitations. Namely the fact that most of these lexicons are a form of local representations of emotion. They fail to encode the context in which the words elicit specific emotions. Sarcastic texts would be a foil to these lexicon representations. Furthermore, words we usually assume not to have any emotional meaning could carry some depending on the context in which they are being used, and these lexicons would fail to identify that.

2.2 Deep Learning Model Foundations

The Deep Learning-based models that have been used in previous research have the following basic models as building blocks. Many of the models proposed in this thesis make use of these building blocks a well.

2.2.1 LSTM

LSTM[19] is an RNN well suited to process textual data. At each timestep i, an LSTM cell processes the ith token of the input sequence and the hidden state of the previous timestep to generate a hidden output state using internal gates that learn how to aggregate information from previous hidden states and the current input element. This process is unidirectional. In order to make it bidirectional, another LSTM is employed that processes the sequence in reversed order.

2.2.2 DeepMoji

DeepMoji [20] is a 2-layer BiLSTM with Attention [21]. It has been pretrained on the task of predicting the occurrence of emojis in an input text. This model has been shown to perform well on tasks such as emoji prediction and sarcasm detection achieving SOTA

performance. Moreover, other models have leveraged the knowledge learned by DeepMoji in their models with significant improvement to their performance, suggesting that DeepMoji generates suitable affective feature vectors and that our model can benefit from using its feature vectors as reference.

2.2.3 BERT

BERT is a model developed by Devlin, Chang, Lee, *et al.* [22] that has achieved SOTA performance in many benchmark NLP tasks ranging from sentiment analysis to natural language inference. As the name suggests, the model is based on the transformer[23] architecture consisting of an attention mechanism that learns contextual relations between words in a text excerpt. Unlike RNN-based or CNN-based models that encode a sequence of text unidirectionally or partially bidirectionally, Transformers do so in a pure bidirectional manner by processing the entire sequence at once and learning the word pair interactions via self-attention. In the case of BERT, this transformer architecture not only has many learned parameters but they were also learned from a vast corpus that allowed BERT to learn a rather generic and well-representing model of the English language.

2.3 Deep Learning Emotion Classification

NTUA-SLP [24] was the best performing model submitted to SemEval'18's Task 1 [25]. The authors trained a BiLSTM with deep attention where the input embeddings were Word2Vec[cite] word vectors whose dimensions were augmented with hand-picked affective features. The augmentation of the word vectors did not improve their performance on the multilabel classification task though it did improve their performance on regression tasks.

Seq2Emo [26] is one of the most recent deep learning-based models for emotion classification. Its architecture mimics a Seq2Seq model where an encoder BiLSTM network transforms a sequence of tokens into a sequence of emotion encodings. Their model does not use any other emotion or sentiment information such as lexicons or pretrained affective embeddings. Their performance is akin to that of NTUA-SLP despite the little additional information that it uses.

GoEmotions [27] is the most recent benchmark dataset for emotion classification. Demszky, Movshovitz-Attias, Ko, *et al.* have applied BERT to the dataset with some significant results. Notably, a standard pretrained BERT had already embedded much affective information, leading to quick learning and improved performance on other tasks when pretrained on the GoEmotions dataset. BERT based models have also been applied to emotion classification **alhuzali2021spanemo** by virtue of... The autoencoder-based approach has also been used to construct a latent variable representation from the latent emotion module to guide the prediction **fei2020latent**.

2.4 Emojis In Emotion Classification

Emojis are a pictorial representation of various concepts, including emotions, objects, and activities. Since their dawn in the early 2010s, emojis have become increasingly commonplace in our modern forms of electronic communication.

Hu, Guo, Sun, *et al.*

Hu, Guo, Sun, *et al.* conducted a study on the usage of emojis while focusing on the intent behind their widespread use [28] and found that emojis are used to express positive or negative sentiment, further increasing the amount of sentiment expressed in a text excerpt, and adjust the tone of a message to convey sarcasm, irony, humor, or closeness. Another significant finding of this study is that the authors found that positive or negative emojis can significantly change the overall sentiment of a seemingly neutral message by including an emoji, thus suggesting that emojis carry some crucial affect value. Ai, Lu, Liu, *et al.* conducted a similar study with the focus on understanding what leads an emoji to be more popular than others by analyzing the relationship between emojis and the context in which they are presented [29]. They found that emoji usage is characterized into two distinct functions: complementary and supplementary, where the complementary function emphasizes the meaning of a message by adding emojis. In contrast, the supplementary function replaces a word with an emoji that has the same meaning. Furthermore, the authors found that the most popular emojis are also the ones that convey the most sentiment.

Delobelle and Berendt argued that emojis were not used enough in NLP models [30]. They found that in most NLP research, emojis have either been underutilized or not utilized at all. The proper use of emojis can increase the performance of contextual models, with an observed increase of 5.85% in performance once emojis were used in their conversational model.

3. DATASETS

3.1 SemEval 2018 Task 1 - Affect in Tweets

SemEval 2018: Affect in Tweets [25] was a competition dedicated to affect-related NLP tasks such as emotion classification, emoji prediction, and sarcasm detection with datasets for each respective task. We use the multilabel emotion classification subset of the dataset in our experiments. We will now refer to this subset as the SemEval'18 dataset.

This dataset consists of tweets from 2016 to 2017 that contained terms related to *anger*, *fear*, *joy*, and *sadness* at different intensities. These are later labeled not only for the previously mentioned four emotions, but also the *anticipation*, *disgust*, *love*, *optimism*, *pessimism*, *surprise*, and *trust* emotions with neutral as no emotions. This labeling scheme leads to an overrepresentation of the basic emotions compared to the others, resulting in an inherent class imbalance in the dataset. These labels were attributed based on the categorical model of emotions (i.e., Plutchik [31] and Ekman [32] models). Given that the dataset consists of Tweets, we can expect to find Twitter-only tokens such as mentions and hashtags in our samples in addition to emojis.

The SemEval'18 dataset is split into train, validation, and test sets as described in Table 3.1. We can also observe the number of coincident classes per sample in Table 3.4.

Table 3.1. The number of occurrences of each emotion in each of the SemEval'18 dataset splits.

	Train	Valid	Test
Anger	2,544	315	1,101
Anticipation	978	124	425
Disgust	2,602	319	1,099
Fear	1,242	121	485
Joy	2,477	400	1,442
Love	700	132	516
Optimism	1,984	307	1,143
Pessimism	795	100	375
Sadness	2,008	265	960
Surprise	361	35	170
Trust	357	43	153
Split Size	6,838	886	3,259

3.2 CancerEmo

CancerEmo is a dataset designed for emotion classification focused on patients' experiences with different types of cancer, namely, breast, lung, and prostate cancer. This dataset is relatively novel and has not been studied extensively yet.

The dataset was collected from Online Health Communities (specifically csn.cancer.org), where patients and caregivers share their experiences with cancer ranging from undergoing medical procedures to side effects from their treatment. Unlike the previous datasets, the authors chose to collect and annotate only sentences instead of entire posts, preventing the use of posts that were longer and incorporating many topics and emotions simultaneously.

The emotion classes defined in this dataset are *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*. These classes are according to Plutchik's basic emotions model. It can also be observed that positive emotions are underrepresented in this dataset.

The data was collected from 2002 to 2018, and the resulting dataset is sampled from that period.

For each emotion class, the authors provide separate train, validation, and test splits, meaning that each emotion is trained individually, whereas the other datasets were designed to have the classes trained jointly.

Table 3.2. Dataset split sizes of the CancerEmo dataset. Each emotion has it's own subset of train, validation, and test splits.

	Train Size	Valid Size	Test Size
Anger	669	84	84
Anticipation	360	34	42
Disgust	735	90	66
Fear	4,310	539	539
Joy	4,834	604	605
Sadness	2,884	361	361
Surprise	614	102	110
Trust	1,509	189	189

Despite each class being trained individually, samples still share multiple labels. Table 3.2 shows the distribution of those labels per example, and we can observe that, like the other datasets, very few samples have more than three classes.

3.3 GoEmotions

GoEmotions is the largest manually annotated dataset of comments labeled for 27 emotion categories that can perform well in transfer learning settings. Given its novelty, there have not been many published studies using this dataset.

Reddit[1] is a collection of forums where users submit posts and submit comments and rate these posts. Posts are grouped by subreddits dedicated to any specific topic (e.g., /r/math for mathematics-related posts). The dataset was designed to reduce harmful and offensive language bias and have a balanced emotion distribution. In addition to that, the dataset was also designed not to overrepresent popular communities leading to more bias.

The dataset is annotated with an extensive taxonomy of emotions to get as much emotion coverage as possible. The labels are based on Cowen's[33] statistical emotion model and consist of the following: admiration, approval, annoyance, gratitude, disapproval, amusement, curiosity, love, optimism, disappointment, joy, realization, anger, sadness, confusion, caring, excitement, surprise, disgust, desire, fear, remorse, embarrassment, nervousness, pride, relief, and grief. With the emotions desire, disappointment, pride, realization, relief, and remorse being suggested by raters because of how frequently raters have identified them. The dataset has comments ranging from 2005 to 2019.

The GoEmotions dataset is split into train, validation, and test sets as described in Table 3.3. In addition to that, the table also shows the number of instances labeled with each emotion per split. We can observe that grief, nervousness, pride, and realization are severely underrepresented compared to the other classes, which could lead to these being the classes in which our models perform the worst.

[1]↑https://www.reddit.com

Table 3.3. GoEmotions Emotions per split.

	Train	Valid	Test
Admiration	4,130	488	504
Amusement	2,328	303	264
Anger	1,567	195	198
Annoyance	2,470	303	320
Approval	2,939	397	351
Caring	1,087	153	135
Confusion	1,368	152	153
Curiosity	2,191	248	284
Desire	641	77	83
Disappointment	1,269	163	151
Disapproval	2,022	292	267
Disgust	793	97	123
Embarrassment	303	35	37
Excitement	853	96	103
Fear	596	90	78
Gratitude	2,662	358	352
Grief	77	13	6
Joy	1,452	172	161
Love	2,086	252	238
Nervousness	164	21	23
Optimism	1,581	209	186
Pride	111	15	16
Realization	1,110	127	145
Relief	153	18	11
Remorse	545	68	56
Sadness	1,326	143	156
Surprise	1,060	129	141
Neutral	14,219	1,766	1,787
Split Size	43,410	5,426	5,427

3.4 Label Agreement

As a learning task, emotion classification is a challenging task because emotion perception is very subjective. One person might perceive one emotion while another person might perceive another emotion from the same piece of text. Moreover, given that we are only observing text, the perception of emotion becomes even more subjective, leading to difficulty labeling the datasets and training the emotion learning models.

Table 3.4. Summary of the datasets

		SemEval'18	GoEmotions	CancerEmo
Train Size		6,838	43,410	10,288
Valid Size		886	5,426	1,886
Test Size		3,259	5,427	1,872
Total Size		10,983	54,263	14,046
Number of Emotions		11	27+Neutral	8
Labels per Example				
	0	2.67%	0.00%	36.12%
	1	13.48%	83.75%	57.23%
	2	40.89%	14.97%	6.31%
	3	31.49%	1.21%	0.33%
	4+	11.46%	0.07%	0.02%

Table 3.5 shows the rater agreement scores for each of the previously described datasets. The scoring metric is different for each dataset, but we can compare them to some extent because they are on similar scales, especially in the case of the SemEval'18 and the GoEmotions datasets. These datasets achieve fair agreement at best, which is not much. However, as noted by Mohammad and Kiritchenko [34], that is sufficient and adequately incorporates the ambiguity of emotions though it makes it harder to quantify the performance of our models properly.

3.5 Emoji Usage

Emoji is used frequently in social media settings, and this can be reflected in our selection of datasets as well for the most part. The exception is the GoEmotions dataset because it is comprised of Reddit comments. There is an unspoken rule of sorts in Reddit that discourages the usage of emojis leading to the vast majority of users disregarding its usage entirely. Given that CancerEmo is based on sentences from a community forum, the usage of emojis is quite rare.

Table 3.5. Interrater agreement scores for each of the datasets

Emotions	SemEval'18 (Fleiss' Kappa, p=7)	CancerEmo (Kippendorff's Alpha, p=3)*	GoEmotions (Cohen's Kappa, p=4)
Admiration	-	-	0.47
Amusement	-	-	0.47
Anger	0.41	0.69	0.31
Annoyance	-	-	0.19
Anticipation	0.04	0.50	-
Approval	-	-	0.19
Caring	-	-	0.25
Confusion	-	-	0.27
Curiosity	-	-	0.37
Desire	-	-	0.25
Disappointment	-	-	0.18
Disapproval	-	-	0.23
Disgust	0.20	0.69	0.24
Embarrassment	-	-	0.22
Excitement	-	-	0.22
Fear	0.38	0.75	0.39
Gratitude	-	-	0.75
Grief	-	-	0.10
Joy	0.47	0.75	0.30
Love	0.21	-	0.56
Nervousness	-	-	0.14
Optimism	0.18	0.69	0.30
Pessimism	0.08	0.69	-
Pride	-	-	0.15
Realization	-	-	0.16
Relief	-	-	0.19
Remorse	-	-	0.36
Sadness	0.32	0.75	0.34
Surprise	0.07	0.69	0.33
Trust	0.04	0.69	-

p indicates the number of raters per sample on average.
*These are estimates based on how the author reported the agreement scores.

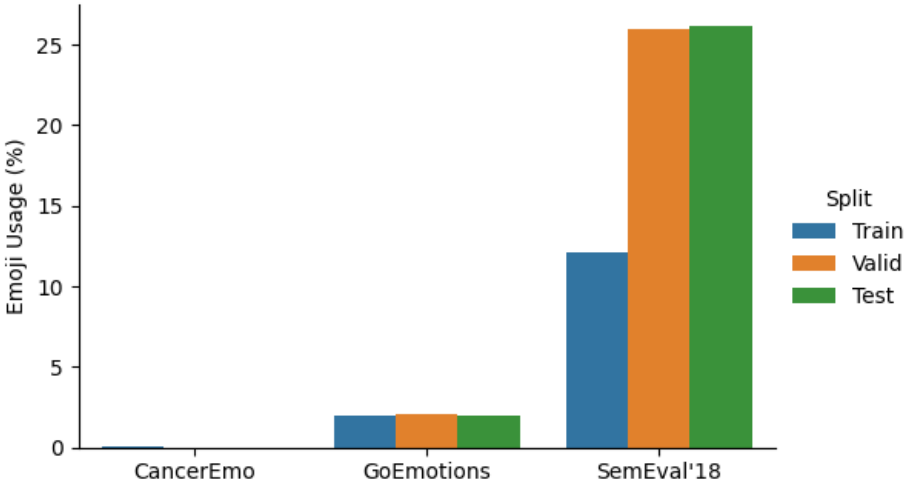


Figure 3.1. Emoji Usage in the Selected Datasets

Table 3.6. Emoji statistics of the datasets

	SemEval'18			GoEmotions		
	Total Samples	Samples w. Emojis	Avg. Emoji per Sample*	Total Samples	Samples w. Emojis	Avg. Emoji per Sample*
admiration	-	-	-	4,130	105	1.85
amusement	-	-	-	2,328	72	2.24
anger	2,544	269	1.96	1,567	18	2.11
annoyance	-	-	-	2,470	45	1.44
anticipation	978	104	2.60	-	-	-
approval	-	-	-	2,939	90	2.19
caring	-	-	-	1,087	33	1.79
confusion	-	-	-	1,368	13	1.31
curiosity	-	-	-	2,191	55	1.44
desire	-	-	-	641	22	1.73
disappointment	-	-	-	1,269	24	2.42
disapproval	-	-	-	2,022	38	1.55
disgust	2,602	268	1.99	793	15	1.27
embarrassment	-	-	-	303	7	1.57
excitement	-	-	-	853	29	1.41
fear	1,242	132	1.70	596	13	1.15
gratitude	-	-	-	2,662	91	1.36
grief	-	-	-	77	3	1.00
joy	2,477	418	2.07	1,452	38	1.26
love	700	157	2.14	2,086	87	1.32
nervousness	-	-	-	164	4	1.25
neutral	-	-	-	14,219	124	1.76
optimism	1,984	247	2.49	1,581	81	1.53
pessimism	795	101	1.68	-	-	-
pride	-	-	-	111	2	2.00
realization	-	-	-	1,110	53	1.23
relief	-	-	-	153	1	1.00
remorse	-	-	-	545	2	1.00
sadness	2,008	271	1.82	1,326	61	1.49
surprise	361	52	1.63	1,060	27	1.56
trust	357	29	1.86	-	-	-

* This is the average number of emojis used in the samples that actually use emojis.

4. MODELS

Let $\mathbf{x} = (x_1, x_2, \cdots, x_T)$ represent a sequence of tokens, including words and sometimes emojis. The task is to encode this sequence $\mathbf{x}$ and generate a feature vector that is used by a classification layer to predict the target labels defined as $\mathbf{y} = (y_1, y_2, \cdots, y_K) \in \{0, 1\}^K$, where $y_{\text{i}} = 1$ indicates that the ith emotion is being expressed in the input sequence and K indicates the number of emotion classes to predict.

4.1 Preprocessing

Each of the proposed models have similar preprocessing requirements for each of their encoding modules.

The BERT based models require the text to be processed as SentencePiece [35] tokens and uses BERT's built-in word embeddings to generate input token vectors. The BiLSTM based models require the text to be processed as word tokens with the addition of some preprocessing steps that preserve hashtags (e.g., #blessed), mentions (e.g., @BarackObama), and emojis. These tokens are then vectorized using GloVe [36] word vectors extended with emojis. The DeepMoji [20] model requires the text to be processed as word tokens and transformed into input embeddings using DeepMoji's own pretrained word vectors.

4.2 BERT+Emojis

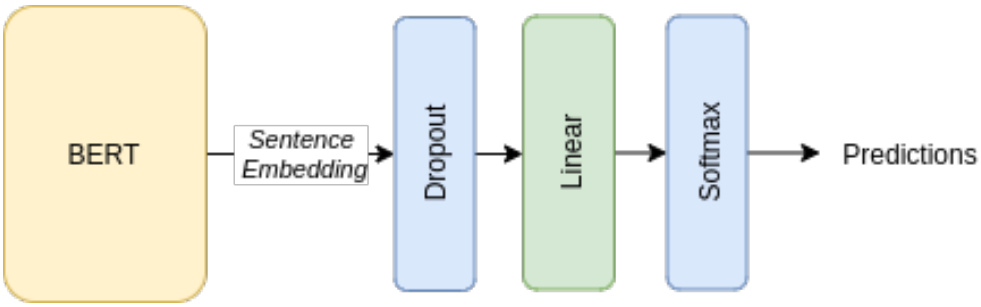


Figure 4.1. BERT architecture

BERT generates both sequence embedding and contextual token embeddings from the input sequence. As illustrated in Figure 4.1, the sentence embedding can be used as input to

a fully connected neural network for classification. We use sigmoid as the activation function because multiple emotions can co-occur.

Usually, BERT is used in downstream tasks by fine-tuning on the specific domain, but in our case, we add emojis as input tokens. Many research done in emotion classification disregards emojis from the training and inference process of their models. Given that emojis can be very representative of an individual's emotional state, we decide to process them as well. Specifically, we process the 50 most frequently occurring emojis in our datasets, as shown in Figure 4.2.

Figure 4.2. The 50 most frequent emojis in the datasets

4.3 BERT+BiLSTM

The preprocessing for this architecture is identical to the preprocessing done for the previous BERT architecture. Moreover, much like the previously described BERT architecture, we process emojis as input tokens as well.

As described prior, BERT generates both sentence embedding and contextual token embeddings. Rather than just using the sentence embedding as a feature vector, we combine

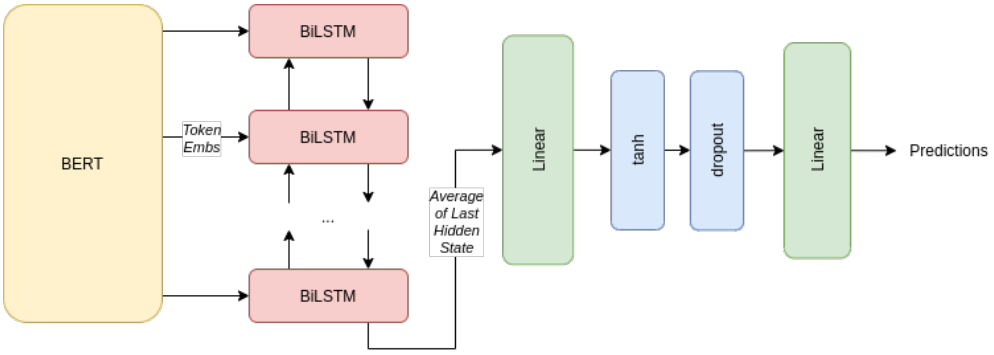


Figure 4.3. BERTBiLSTM architecture

the token embeddings using a BiLSTM and use the average of both directions of the output of the last LSTM cell as the feature vector, as illustrated in Figure 4.3.

4.4 BERTBiLSTM + DeepMoji

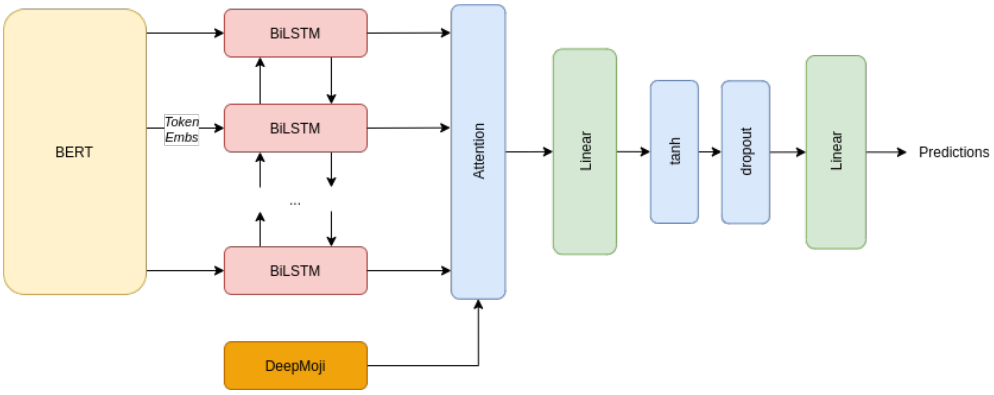


Figure 4.4. BERT+BiLSTM+DeepMoji architecture

This variation of the BERTBiLSTM architecture features the addition of DeepMoji features and an attention layer that aggregates the contextual embeddings and the DeepMoji affective features.

Let h_{i}^b be the BiLSTM output at timestep i and h_t^d be the DeepMoji affective feature vector. The attention scores for the BiLSTM outputs are defined as

$$s_{r,\mathrm{i}}^b = V^\top \tanh(W_q h_t^d + W_v h_{\mathrm{i}}^b) \tag{4.1}$$

$$\alpha_{r,\mathrm{i}}^b = \frac{\exp(s_{r,\mathrm{i}}^b)}{\sum_{\mathrm{j}=0}^{T} \exp(s_{r,\mathrm{j}}^b)} \tag{4.2}$$

where V, W_q, and W_v are learned parameters, $s_{r,\mathrm{i}}^b$ is the attention score for the i-th token embedding, and $\alpha_{r,\mathrm{i}}^b$ is the normalized attention score. The context vector c_t is computed as the attention-score-weighted sum of the BiLSTM outputs.

$$c_t^b = \sum_{\mathrm{i}=0}^{T} \alpha_{r,\mathrm{i}}^b h_{\mathrm{i}}^b \tag{4.3}$$

This context vector c_t^b is used as the input to the linear layers yielding the classifications.

$$\mathbf{y} = \sigma(W_2 \tanh(W_1 c_t + b_1) + b_2) \tag{4.4}$$

Given that DeepMoji was trained on emoji prediction and sarcasm detection, it has learned to generate decent affective representations from text hence our choice in doing so.

4.5 ReferEmo: Referential Emotion Encoder

The proposed model, as illustrated in Figure 4.5, encodes a sequence of tokens, including words and emojis, defined as $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, and generates a feature vector c_t used by a classification model to predict the target labels defined as $\mathbf{y} = (y_1, y_2, \cdots, y_K) \in \{0, 1\}^K$, where $y_{\mathrm{i}} = 1$ indicates that the ith emotion is being expressed in the input sequence. The architecture consists of three distinct layers: encoding layer, feature aggregation layer, and classification layer, which are preceded by a module-specific preprocessing step. This architecture borrows some of the ideologies of ensembles [37], with the major differences being the aggregation of feature vectors as opposed to predictions and the aggregation mechanism being learned as opposed to having a majority voting or a boosting scheme.

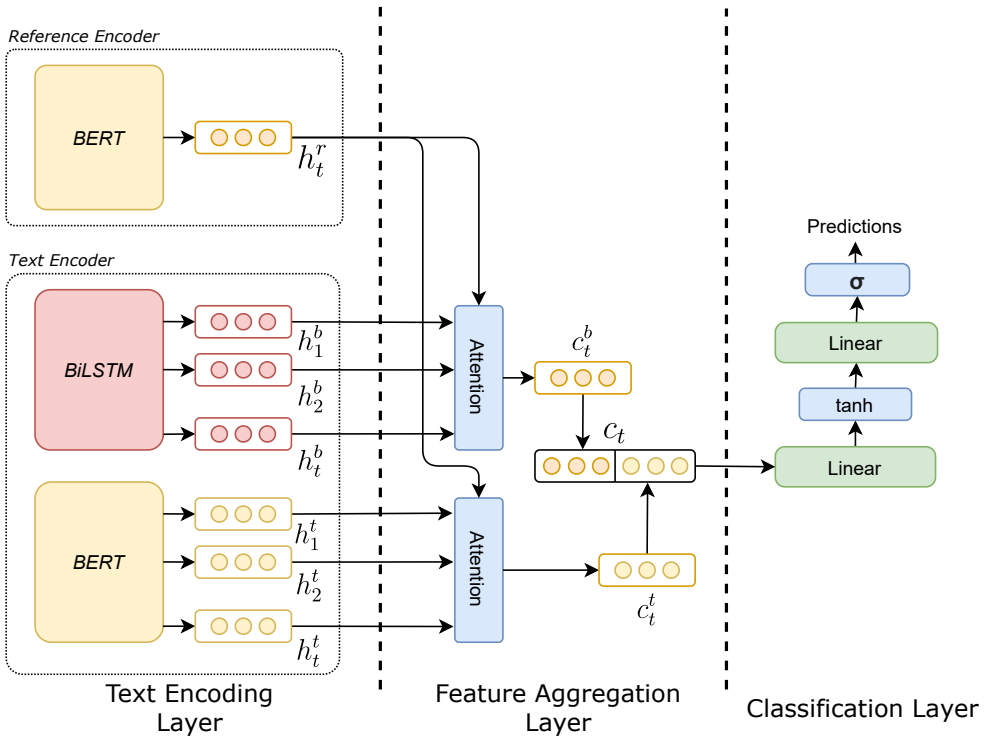


Figure 4.5. ReferEmo architecture

4.5.1 Encoding Layer

The encoding layer is responsible for creating the various feature vectors from the input sequence used by the upper layers. This layer consists of two types of encoders: a reference encoder and a text encoder.

Reference Encoder

The reference encoder is a model that generates an affective feature vector h_t^r. As the name suggests, this affective feature vector serves as a reference to enrich the word embeddings of the input sequence with affect knowledge. We use a pretrained DeepMoji[20] model as the reference encoder.

$$h_t^r = DeepMoji(\mathbf{x})$$

DeepMoji has been shown to perform well on tasks such as emoji prediction and sarcasm detection achieving state-of-the-art performance. Furthermore, other models have leveraged the knowledge learned by DeepMoji in their models with significant improvement to their performance. These facts suggest that DeepMoji generates good affective feature vectors and that our model can benefit from using its feature vectors as reference. Note that the use of the pretrained DeepMoji model is a deliberate design choice. Any model that can generate an affective feature vector can be used.

Text Encoder

This encoder generates a set of contextual token embeddings from the input sequence. These tokens embeddings are later enriched with the affective knowledge from the reference vector generated by the reference encoder.

We use a 2-layer BiLSTM that encodes an input sequence $\mathbf{x} = (x_1, \cdots, x_n)$ into the contextual token embeddings $h_1^b, \cdots, h_t^b$. These token embeddings are the intermediate outputs of the BiLSTM at timestep i and are defined as

$$h_{\mathrm{i}}^b = BiLSTM_{\mathrm{i}}(x_{\mathrm{i}})$$

Where h_{i}^b is the hidden state of the BiLSTM at timestep i. It summarizes all of the sequence information up to x_{i} from both the forward and backward directions.

In addition to the BiLSTM, we also use BERT as an additional text encoder. BERT is a transformer-based model that is pretrained on a large corpus for a masked language modeling task. BERT has shown state-of-the-art performance on many NLP tasks ranging from machine translation to sentiment analysis. Particularly, its performance on sentiment-related tasks suggests that BERT can assimilate how affectiveness is expressed in text. Preliminary experimental results have shown that the BiLSTM encoder yielded a better precision while the BERT encoder yielded a better recall. We use the two encoders to improve the sensitivity of the model while still maintaining high specificity. We define the token embeddings generated by BERT as the sequence $h_1^t, \cdots, h_t^t$ where h_{i}^t represents the embedding of the ith token in the sequence within the context of the entire input sequence.

4.5.2 Attention-based Feature Aggregation Layer

This layer receives as inputs the reference feature vector h_t^r and the token embeddings $[h_1^t, \cdots, h_t^t]$ and $[h_1^b, \cdots, h_t^b]$ from the previous layer and generates a context vector that aggregates the values of the sequence embeddings and the affective value of the reference vector.

In our architecture, we use attention as the aggregation mechanism. Given that there are two sets of token embeddings, two sets of attention scores are computed. One set of attention scores is between the reference vector and the BiLSTM token embeddings, and the other set is between the reference vector and the BERT token embeddings. Using attention not only generates a feature vector that better encodes longer sequences with long dependencies but also allows us to visualize the alignment scores between the reference feature vector and the token embeddings, thus providing information regarding which tokens carry the most affect value that allows us to assess the relationship between these input tokens and the classification label.

The attention scores with respect to the BiLSTM embeddings are defined as

$$s_{r,\mathrm{i}}^b = V^\top \tanh(W_q h_t^r + W_v h_\mathrm{i}^b) \tag{4.5}$$

$$\alpha_{r,\mathrm{i}}^b = \frac{\exp(s_{r,\mathrm{i}}^b)}{\sum_{\mathrm{j}=0}^{T} \exp(s_{r,\mathrm{j}}^b)} \tag{4.6}$$

where V, W_q, and W_v are learned parameters, $s_{r,\mathrm{i}}^b$ is the attention score for the ith token embedding, and $\alpha_{r,\mathrm{i}}^b$ is the normalized attention score. The attention scores with respect to the BERT embeddings are defined in an identical manner.

The context vector for each of the embeddings is computed as the attention-weighted sum of the token embeddings.

$$c_t^b = \sum_{\mathrm{i}=0}^{T} \alpha_{r,\mathrm{i}}^b h_\mathrm{i}^b \tag{4.7}$$

$$c_t^t = \sum_{\mathrm{i}=0}^{T} \alpha_{r,\mathrm{i}}^t h_{\mathrm{i}}^t \tag{4.8}$$

The final context vector is defined as the concatenation of c_t^b and c_t^t

$$c_t = [c_t^b; c_t^t] \tag{4.9}$$

4.5.3 Classification Layer

The last layer consists of a two-layer fully connected neural network with a *tanh* activation between the two layers and a sigmoid activation at the output layer. Since the model architecture is designed for multilabel classification tasks, using sigmoid as the activation of the output layer is the most suitable option.

$$\mathbf{y} = \sigma(W_2 \tanh(W_1 c_t + b_1) + b_2) \tag{4.10}$$

5. EXPERIMENTS AND RESULTS

5.1 Baselines

We trained each of the models described in chapter 4 and compare their performance on the SemEval'18, GoEmotions, and CancerEmo datasets with some of the better performing models found in literature. These models are:

- **NTUA-SLP**[24], an RNN-based model with domain-specific word embeddings and the top ranked model submitted to the SemEval'18 competition.
- **Seq2Emo**[26], an RNN-based encoder-decoder model that classifies emotions sequentially and the most current SOTA model for the emotion classification task.

Our comparison will be focused on how our models differ from the NTUA-SLP and the Seq2Emo models both at the summary level and at the emotion level. However, at the summary level, we also report the performance of other notable models for each dataset. Namely:

- **BERT** [22], a transformer-based model pretrained on a language modeling task.
- **LEMfei2020latent**, an auto-encoder that encodes emotions as latent variables.
- **BNetjabreel2019deep**, an RNN-based model that transforms the multi-label classification task into a binary classification task.
- **TCS Research**[10], an ensemble model that incorporates deep learning features with lexical features.
- **PlusEmo2Vec**[38], a model that generates domain-specific emotion embeddings with CNNs.

5.2 Evaluation Metrics

The models studied in this thesis were evaluated with the standard classification metrics of precision, recall, and F1 score [39], with the addition of the Jaccard score [40], which is frequently used to evaluate multilabel classification tasks.

Let TP, TN, FP, and FN be true-positive, true-negative, false-positive, and false-negative predictions. Precision and Recall are defined, respectively, as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

The F1 score is then defined as the harmonic mean of the precision and the recall.

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

The Precision, Recall, and F1 score are computed for each of the classes in our datasets. Reporting the per-class metrics allows us to have a better picture of the actual performance of our models in an unbundled manner. In addition to the individual class metrics, we compute the micro and macro averages of the F1 scores to have a more holistic understanding of the performance of our models.

The Jaccard score is an alternative to the accuracy score in a multilabel setting. It measures the overlap between the target labels of a dataset and the predicted labels of a model. Let T be the target label and P the predicted label. The Jaccard score for one sample is defined as

$$J(T, P) = \frac{|T \cap P|}{|T \cup P|} \tag{5.4}$$

The Jaccard Score for the entire dataset is defined as the average of the samples' Jaccard Scores, as illustrated in Equation 5.5, where S represents the dataset.

$$Jaccard = \frac{1}{|S|} \sum_{s \in S} J(T_d, P_d) \tag{5.5}$$

Table 5.1. Optimal hyperparameters for each of the models except BERT+Emojis

	BERT+BiLSTM	BERT+BiLSTM+DeepMoji	ReferEmo	Range
lstm_hidden	512	512	256	[2^7, 2^{10}]
num_layers	1	1	1	[1,3]
linear_hidden	128	128	128	[2^7, 2^10]
linear_dropout	0.25	0.25	0.5	[0.2, 0.7]
lr	5.45E-05	5.45E-05	1.15E-05	[1E-5, 1E-1]
optimizer	RMSProp	RMSProp	RMSprop	[Adam, AdamW, RMSprop, SGD]

5.3 Hyperparameter Tuning

We perform hyperparameter search using Optuna[41], a tool that aids in performing grid search to find the most optimal hyperparameters. The models that required hyperparameter search are BERT+BiLSTM, BERT+BiLSTM+DeepMoji, and ReferEmo. BERT+Emojis only had the learning rate as the hyperparameter and using the default value of $1E-5$ works well for SemEval'18 and GoEmotions. The optimal hyperparameters for the prior models are shown in Table 5.1, where lstm_hidden refers to the number of hidden units in an LSTM cell, num_layers refers to the number of LSTM layers to have in the architecture. linear_hidden refers to the number of hidden units in the Linear layer that performs the final classification, and linear_dropout refers to the dropout probability at the classification layer. LR refers to the learning rate and optimizer to the gradient descent optimizer to train the model. Note that our architectures seem to favor shallow LSTMs.

5.4 Classification Performance Analysis

5.4.1 SemEval'18 Performance

Table 5.2 shows the summarized performance of our models and the baseline models on the SemEval'18 dataset. We can see that ReferEmo and the other BERT-based models, except for BERT+BiLSTM+DeepMoji, perform better than the baselines, particularly in the F1 measures. For instance, NTUA-SLP's macro F1 and ReferEmo's F1 scores differ by 4%, while the micro F1 scores only differ by 1%, suggesting that our ReferEmo model can better classify underrepresented groups classes when compared to NTUA-SLP and Seq2Emo. Comparing the performance of our models to that of LEM, we observe that the macro

F1 score is relatively high compared to the other baselines while the micro F1 score is the lowest of the baselines, suggesting that there was an inherent tradeoff when classifying overrepresented and underrepresented classes. Our models do not have to do such tradeoffs, as can be seen in our results. Moreover, BERT+Emojis performs best at the summary level across the three metrics, hinting at emojis' representative power in emotion classification tasks. One interesting observation is that the model that incorporates DeepMoji performs the worst even though DeepMoji has performed well in some affect related tasks.

Table 5.2. Summary of the performance on the SemEval'18 dataset

	Micro F1	**Macro F1**	**Jaccard**
NTUA-SLP	0.70	0.53	0.59
Seq2Emo	0.70	0.52	0.59
LEM	0.67	0.56	-
BNet	0.69	0.56	0.59
TCS Research	0.69	0.53	0.58
PlusEmo2Vec	0.69	0.50	0.58
BERT+Emojis	0.72	0.59	0.60
BERT+BiLSTM	0.71	0.57	0.58
BERT+BiLSTM+DeepMoji	0.67	0.50	0.55
ReferEmo	0.71	0.57	0.58

In order to gain a better understanding of the performance of each of these models, we study the classification metrics, namely the precision, recall, and F1 score, for each emotion category, as shown in Table 5.3. NTUA-SLP and Seq2Emo perform the worst on the anticipation, surprise, pessimism, and trust categories. Not surprisingly, these classes have lower inter-rater agreement scores (as shown in Table 3.5). In addition to providing some minor improvement on some of the better performing categories of NTUA-SLP and Seq2Emo, the BERT-based models significantly improve the worse performing and underrepresented classes. Namely, anticipation and pessimism, with an improvement of 10%. We observe some improvement in the surprise and trust classes, though not significant. The low agreement score suggests that the samples labeled with surprise and trust might be too ambiguous for a model to learn correctly, and given that the number of samples with these classes is deficient, it is improbable to perform better in these classes without up-sampling these classes. Note

that the BERT+BiLSTM+DeepMoji model does not learn anything from the surprise and trust classes showing how difficult it is for a weaker model to learn from them.

5.4.2 GoEmotions

Table 5.4 shows the summarized performance of our models and the selected benchmark models for the GoEmotions dataset. We can see that NTUA-SLP performs worse on this dataset. This can be attributed to the fact that NTUA-SLP uses domain-specific embeddings, making it slightly more challenging to generalize on datasets of a different domain, as is the case for GoEmotions. Seq2Emo, however, only performs marginally better than BERT. Despite not performing better than Seq2Emo on the micro F1 measure, our models perform better on the macro F1 metric, thus suggesting that our models perform better on the underrepresented classes.

Table 5.5 shows the per emotion classification metrics much like in Table 5.3 for the SemEval'18 dataset performance. As expected, for NTUA-SLP and Seq2Emo, the performance for the underrepresented classes, namely grief, nervousness, pride, and relief, is low. What was not expected was the low performance in some classes with a significant number of samples such as annoyance, caring, and disapproval. The agreement score for these classes indicates lower ambiguity in the labels than in the SemEval'18 dataset, thus suggesting that these classes are inherently hard to learn.

The recall for the BERT-based models is much higher than that of the benchmarks. BERT-based models seem to have a higher sensitivity while having a slightly lower specificity. This tradeoff is to be considered if any of these models are to be deployed in any system. Note that, once again, the model with the added DeepMoji does not perform as well as the other ones suggesting that DeepMoji only hinders the performance of this model.

5.4.3 CancerEmo

Unlike the other datasets, each class in the CancerEmo dataset was trained independently from others despite being described as a multilabel dataset. To conform to previous research that used the dataset and make a fair comparison, we do the same thing. Therefore, Table

Table 5.3. Per emotion performance on the SemEval'18 dataset

		NTUA-SLP			**Seq2Emo**			**BERT+Emojis**			**BERT+BiLSTM**			**BERT+BiLSTM +DeepMoji**			**ReferEmo**		
Emotion	Support	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	1101	0.77	0.77	0.77	0.76	0.79	0.77	0.76	0.84	0.80	0.76	0.82	0.79	0.74	0.82	0.78	0.79	0.78	0.79
Anticipation	425	0.39	0.14	0.20	0.38	0.12	0.17	0.36	0.28	0.32	0.37	0.31	0.34	0.31	0.20	0.24	0.37	0.28	0.29
Disgust	1099	0.69	0.77	0.73	0.68	0.76	0.72	0.70	0.84	0.77	0.70	0.82	0.75	0.67	0.83	0.74	0.73	0.78	0.75
Fear	485	0.82	0.67	0.74	0.78	0.65	0.71	0.72	0.77	0.75	0.72	0.78	0.75	0.54	0.55	0.54	0.72	0.77	0.74
Joy	1442	0.86	0.83	0.84	0.85	0.83	0.84	0.83	0.88	0.85	0.82	0.87	0.84	0.79	0.86	0.82	0.81	0.89	0.85
Love	516	0.71	0.47	0.56	0.67	0.50	0.56	0.55	0.74	0.63	0.57	0.72	0.64	0.53	0.71	0.61	0.52	0.76	0.61
Optimism	1143	0.71	0.67	0.69	0.70	0.71	0.70	0.69	0.85	0.76	0.67	0.86	0.75	0.64	0.84	0.73	0.67	0.87	0.75
Pessimism	375	0.48	0.20	0.29	0.48	0.17	0.25	0.41	0.48	0.44	0.38	0.46	0.42	0.38	0.34	0.36	0.41	0.46	0.43
Sadness	960	0.77	0.63	0.68	0.75	0.63	0.69	0.67	0.76	0.71	0.64	0.75	0.69	0.60	0.71	0.65	0.67	0.75	0.70
Surprise	170	0.43	0.11	0.17	0.56	0.07	0.12	0.58	0.11	0.19	0.58	0.06	0.12	0.00	0.00	0.00	0.59	0.11	0.18
Trust	153	0.24	0.07	0.10	0.31	0.03	0.04	0.35	0.18	0.23	0.31	0.13	0.18	0.00	0.00	0.00	0.29	0.09	0.13

Table 5.4. Summary of the performance on the GoEmotions dataset

	Micro F1	**Macro F1**	**Jaccard**
NTUA-SLP	0.54	0.44	0.48
Seq2Emo	0.60	0.47	0.54
BERT	0.59	0.46	0.53
BERT+Emojis	0.57	0.50	0.39
BERT+BiLSTM	0.56	0.49	0.38
BERT+BiLSTM+DeepMoji	0.52	0.47	0.33
ReferEmo	0.56	0.48	0.53

5.6 shows the per emotion classification metrics with the average scores for each metric at the end of the table. The authors only reported the F1 scores for each model they tested hence the missing values for precision and recall for BERT, BiLSTM, and CNN.

As shown prior, this dataset does not have many emojis, with only the joy class having very few occurrences of emojis in the training set. This leads to no added improvement from leveraging emoji representations, as seen in similar metrics of the BERT and the BERT+Emojis models. ReferEmo suffers the most in terms of performance in all but the joy and fear classes. ReferEmo still maintains a somewhat reasonable recall performance. The precision, however, is the most affected metric suggesting that while still sensitive to the correct samples, the model fails to recognize them correctly. Given that CancerEmo is a medical domain dataset, it is possible that the vocabulary ReferEmo was trained does not provide the coverage of terms needed for it to perform well on this dataset leading to the decrease in the performance.

5.5 Label Ambiguity

Many examples were labeled ambiguously, making it difficult for a model to properly learn the relationships between the input tokens and the emotion classes. This ambiguity is even more prevalent when not even human annotators can agree with the gold standard label. The ambiguous labeling in Figure 5.1 shows some examples of these cases. One example shows that our model cannot identify 'love' and 'surprise', and the other example shows that

our model cannot identify 'trust'. This ambiguity, in turn, hinders the model performance, especially when it is a rather sensitive model like ReferEmo. These ambiguous labels are more present in the 'anticipation' and 'trust' classes of the SemEval'18 dataset than any other class. Part of the reason why the labels are so ambiguous is that the raters were given relatively relaxed conditions for attributing a label to an example [25]. In addition, examples are labeled with one primary and multiple secondary emotions. Some of the ambiguous labels might have secondary labels, though there is no way of confirming that.

5.6 Ambiguous Emoji Usage

Some of the examples in both datasets use emojis that lead to ambiguity when the model attempts to classify them. Figure 5.1 provides two examples to demonstrate the cases of ambiguous emoji usage. The first example in the Ambiguous Emoji Usage has a happy emoji, but it is labeled as 'anger' emotion. In the GoEmotions dataset, the happy emojis are commonly associated with more positive emotions such as love and joy. Similarly, some examples of the anticipation emotion in the SemEval'18 dataset use more negative emotions, such as the 'anger' emoji.

5.7 Experimenting with a Hybrid Model

When studying the performance of BERT+Emojis and ReferEmo, we can observe that their performances complement each other, particularly in the recall metric as shown in Tables 5.3 and 5.5. The recall is higher on the emotion classes, while the other model performs slightly worse. For instance, the 0.06 difference in the anger class as shown in Table 5.3 and the 0.33 difference in the desire class as shown in Table 5.5.

For that reason, we combine both the BERT and the ReferEmo models to make a hybrid model. This Hybrid model generates a feature vector that results from the concatenation of BERT+Emojis' sentence embedding and ReferEmo's context vector c_t.

The weights from the BERT and ReferEmo models are loaded and fine-tuned on the Hybrid model. The resulting model has slightly improved performance, specifically on the

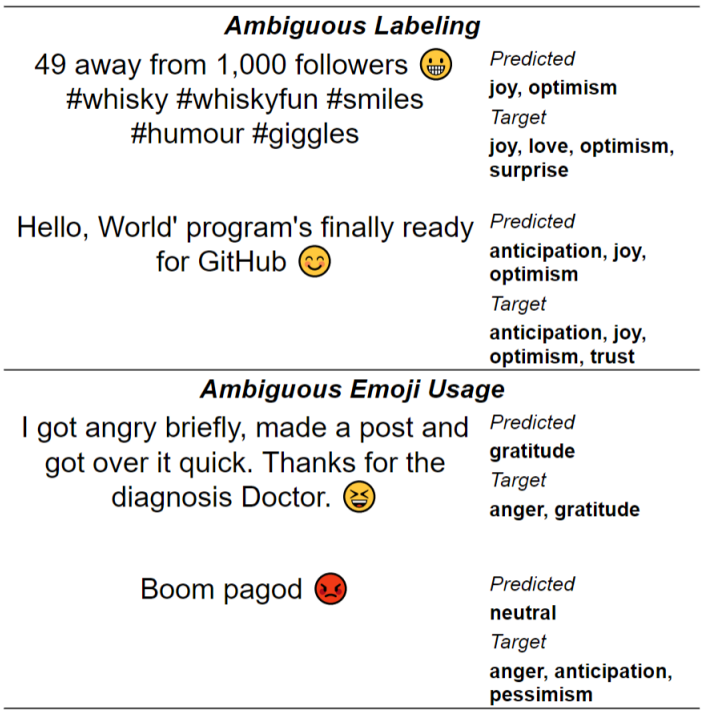


Figure 5.1. Sample tweets showing inherent ambiguity

GoEmotions dataset, as shown in Tables 5.7 and 5.8. However, the improvement is not necessarily significant enough to justify having a larger model comprising BERT and ReferEmo.

This suggests that pretraining a model on a language modeling task, much like BERT was, might fare better on the emotion classification task if it takes into account emojis as well.

Table 5.5. Per emotion performance on the GoEmotions dataset

		NTUA-SLP			**Seq2Emo**			**BERT+Emojis**			**BERT+BiLSTM**			**BERT+BiLSTM +DeepMoji**			**ReferEmo**		
Emotion	Support	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Admiration	4,130	0.63	0.67	0.65	0.66	0.65	0.65	0.64	0.78	0.70	0.64	0.78	0.70	0.63	0.70	0.66	0.62	0.75	0.68
Amusement	2,328	0.75	0.82	0.78	0.78	0.87	0.83	0.75	0.92	0.83	0.75	0.92	0.83	0.75	0.77	0.76	0.72	0.87	0.79
Anger	1,567	0.45	0.36	0.40	0.61	0.32	0.41	0.49	0.53	0.51	0.49	0.53	0.51	0.47	0.45	0.46	0.46	0.50	0.48
Annoyance	2,470	0.36	0.21	0.26	0.49	0.19	0.26	0.33	0.38	0.35	0.33	0.38	0.35	0.31	0.32	0.31	0.31	0.40	0.34
Approval	2,939	0.40	0.25	0.31	0.46	0.24	0.31	0.38	0.42	0.40	0.38	0.42	0.40	0.35	0.38	0.37	0.32	0.42	0.36
Caring	1,087	0.34	0.27	0.30	0.46	0.27	0.34	0.41	0.50	0.45	0.41	0.50	0.45	0.34	0.29	0.31	0.37	0.44	0.40
Confusion	1,368	0.38	0.36	0.37	0.52	0.29	0.37	0.40	0.56	0.46	0.40	0.56	0.46	0.34	0.41	0.37	0.32	0.54	0.40
Curiosity	2,191	0.48	0.41	0.44	0.50	0.43	0.46	0.48	0.71	0.57	0.48	0.71	0.57	0.49	0.54	0.51	0.46	0.64	0.54
Desire	641	0.52	0.34	0.41	0.61	0.33	0.42	0.62	0.49	0.55	0.62	0.49	0.55	0.48	0.34	0.40	0.53	0.46	0.49
Disappointment	1,269	0.27	0.16	0.20	0.42	0.16	0.23	0.38	0.32	0.35	0.38	0.32	0.35	0.32	0.26	0.29	0.30	0.31	0.30
Disapproval	2,022	0.32	0.26	0.29	0.43	0.24	0.31	0.38	0.42	0.40	0.38	0.42	0.40	0.33	0.33	0.33	0.33	0.46	0.38
Disgust	793	0.51	0.41	0.45	0.60	0.38	0.46	0.52	0.46	0.49	0.52	0.46	0.49	0.56	0.42	0.48	0.40	0.56	0.46
Embarrassment	303	0.42	0.32	0.35	0.60	0.31	0.41	0.67	0.38	0.48	0.67	0.38	0.48	0.47	0.24	0.32	0.42	0.40	0.41
Excitement	853	0.43	0.33	0.37	0.51	0.31	0.38	0.45	0.43	0.44	0.45	0.43	0.44	0.39	0.40	0.40	0.36	0.51	0.42
Fear	596	0.60	0.68	0.64	0.64	0.63	0.63	0.58	0.76	0.66	0.58	0.76	0.66	0.59	0.67	0.63	0.56	0.76	0.64
Gratitude	2,662	0.93	0.88	0.90	0.94	0.89	0.91	0.92	0.91	0.92	0.92	0.91	0.92	0.90	0.91	0.90	0.91	0.89	0.90
Grief	77	0.32	0.23	0.26	0.23	0.10	0.14	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.17	0.29	0.00	0.00	0.00
Joy	1,452	0.55	0.51	0.53	0.60	0.52	0.55	0.59	0.67	0.63	0.59	0.67	0.63	0.55	0.51	0.53	0.51	0.63	0.57
Love	2,086	0.74	0.80	0.77	0.77	0.83	0.80	0.75	0.87	0.80	0.75	0.87	0.80	0.74	0.74	0.74	0.72	0.85	0.78
Nervousness	164	0.38	0.23	0.28	0.63	0.17	0.25	0.47	0.35	0.40	0.47	0.35	0.40	0.28	0.22	0.24	0.32	0.38	0.35
Optimism	1,581	0.59	0.47	0.52	0.60	0.45	0.51	0.59	0.58	0.58	0.59	0.58	0.58	0.53	0.48	0.51	0.49	0.58	0.53
Pride	111	0.46	0.32	0.36	0.69	0.30	0.41	0.86	0.38	0.52	0.86	0.38	0.52	0.67	0.38	0.48	0.58	0.34	0.43
Realization	1,110	0.36	0.16	0.22	0.47	0.15	0.22	0.31	0.21	0.25	0.31	0.21	0.25	0.27	0.14	0.19	0.21	0.21	0.21
Relief	153	0.15	0.11	0.12	0.08	0.05	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.36	0.47	0.28	0.25	0.26
Remorse	545	0.53	0.64	0.58	0.57	0.53	0.52	0.54	0.84	0.66	0.54	0.84	0.66	0.60	0.71	0.65	0.58	0.83	0.68
Sadness	1,326	0.54	0.51	0.52	0.60	0.46	0.52	0.53	0.59	0.56	0.53	0.59	0.56	0.56	0.49	0.53	0.50	0.59	0.54
Surprise	1,060	0.56	0.46	0.51	0.62	0.39	0.46	0.52	0.57	0.54	0.52	0.57	0.54	0.52	0.45	0.48	0.50	0.55	0.52

Table 5.6. Summary of the performance on the CancerEmo dataset

	BERT			**BiLSTM**			**CNN**			**BERT+Emojis**			**BERT+BiLSTM**			**BERT+BiLSTM +DeepMoji**			**ReferEmo**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	-	-	0.68	-	-	0.67	-	-	0.58	0.67	0.66	0.62	0.64	0.63	0.59	0.68	0.68	0.65	0.20	0.50	0.29
Anticipation	-	-	0.70	-	-	0.53	-	-	0.54	0.86	0.86	0.86	0.86	0.86	0.86	0.84	0.83	0.83	0.77	0.57	0.48
Disgust	-	-	0.59	-	-	0.57	-	-	0.59	0.82	0.80	0.80	0.75	0.73	0.72	0.77	0.76	0.75	0.25	0.50	0.33
Fear	-	-	0.77	-	-	0.64	-	-	0.59	0.70	0.70	0.70	0.71	0.71	0.71	0.72	0.71	0.71	0.74	0.69	0.68
Joy	-	-	0.81	-	-	0.74	-	-	0.73	0.82	0.82	0.82	0.81	0.81	0.80	0.81	0.81	0.81	0.77	0.75	0.75
Sadness	-	-	0.71	-	-	0.64	-	-	0.63	0.73	0.72	0.72	0.71	0.70	0.70	0.72	0.71	0.71	0.69	0.64	0.62
Surprise	-	-	0.68	-	-	0.50	-	-	0.55	0.68	0.67	0.67	0.76	0.75	0.75	0.77	0.75	0.75	0.25	0.49	0.33
Trust	-	-	0.67	-	-	0.59	-	-	0.66	0.65	0.64	0.63	0.64	0.63	0.62	0.62	0.62	0.61	0.75	0.51	0.34
Average	-	-	0.71	-	-	0.61	-	-	0.61	0.74	0.73	0.73	0.73	0.73	0.72	0.74	0.74	0.73	0.55	0.58	0.48

Table 5.7. Per emotion performance comparison with the Hybrid model on SemEval'18

		BERT+Emojis			ReferEmo			Hybrid		
Emotion	Support	P	R	F1	P	R	F1	P	R	F1
Anger	1101	0.76	0.84	**0.80**	**0.79**	0.78	0.79	0.78	0.81	0.79
Anticipation	425	0.36	0.28	0.32	**0.37**	0.28	0.29	**0.37**	**0.34**	**0.35**
Disgust	1099	0.70	**0.84**	**0.77**	**0.73**	0.78	0.75	0.72	0.80	0.76
Fear	485	**0.72**	0.77	**0.75**	**0.72**	0.77	0.74	**0.72**	**0.79**	**0.75**
Joy	1442	**0.83**	0.88	**0.85**	0.81	**0.89**	**0.85**	**0.83**	0.87	**0.85**
Love	516	0.55	0.74	0.63	0.52	**0.76**	0.61	**0.59**	0.69	**0.64**
Optimism	1143	0.69	0.85	**0.76**	0.67	**0.87**	0.75	**0.70**	0.83	**0.76**
Pessimism	375	**0.41**	**0.48**	**0.44**	**0.41**	0.46	0.43	0.40	0.45	0.42
Sadness	960	**0.67**	**0.76**	**0.71**	**0.67**	0.75	0.70	**0.67**	0.75	**0.71**
Surprise	170	0.58	0.11	0.19	**0.59**	0.11	0.18	0.47	**0.14**	**0.22**
Trust	153	**0.35**	0.18	**0.23**	0.29	0.09	0.13	0.26	**0.21**	**0.23**

Table 5.8. Per emotion performance comparison with the Hybrid model on GoEmotions

		BERT+Emojis			ReferEmo			Hybrid		
Emotion	Support	P	R	F1	P	R	F1	P	R	F1
Admiration	4,130	**0.64**	**0.78**	**0.70**	0.62	0.75	0.68	**0.64**	0.77	**0.70**
Amusement	2,328	0.75	**0.92**	**0.83**	0.72	0.87	0.79	**0.76**	0.91	**0.83**
Anger	1,567	**0.49**	**0.53**	**0.51**	0.46	0.50	0.48	**0.49**	0.52	0.50
Annoyance	2,470	0.33	0.38	0.35	0.31	**0.40**	0.34	**0.34**	**0.40**	**0.36**
Approval	2,939	**0.38**	0.42	0.40	0.32	0.42	0.36	**0.38**	**0.44**	**0.41**
Caring	1,087	**0.41**	0.50	0.45	0.37	0.44	0.40	0.40	**0.53**	**0.46**
Confusion	1,368	**0.40**	**0.56**	**0.46**	0.32	0.54	0.40	0.37	**0.56**	0.44
Curiosity	2,191	0.48	**0.71**	**0.57**	0.46	0.64	0.54	**0.49**	0.70	**0.57**
Desire	641	**0.62**	0.49	**0.55**	0.53	0.46	0.49	0.61	**0.51**	**0.55**
Disappointment	1,269	**0.38**	0.32	**0.35**	0.30	0.31	0.30	0.33	**0.34**	0.34
Disapproval	2,022	**0.38**	0.42	**0.40**	0.33	**0.46**	0.38	**0.38**	0.42	**0.40**
Disgust	793	**0.52**	0.46	**0.49**	0.40	**0.56**	0.46	0.50	0.47	**0.49**
Embarrassment	303	**0.67**	0.38	0.48	0.42	0.40	0.41	0.65	**0.41**	**0.50**
Excitement	853	**0.45**	0.43	**0.44**	0.36	**0.51**	0.42	0.42	0.45	0.43
Fear	596	0.58	0.76	0.66	0.56	0.76	0.64	**0.62**	**0.77**	**0.69**
Gratitude	2,662	**0.92**	**0.91**	**0.92**	0.91	0.89	0.90	0.91	**0.91**	0.91
Grief	77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Joy	1,452	**0.59**	0.67	**0.63**	0.51	0.63	0.57	**0.59**	**0.68**	**0.63**
Love	2,086	**0.75**	**0.87**	**0.80**	0.72	0.85	0.78	0.74	0.86	**0.80**
Nervousness	164	**0.47**	0.35	0.40	0.32	0.38	0.35	0.38	**0.43**	**0.41**
Optimism	1,581	**0.59**	0.58	**0.58**	0.49	0.58	0.53	0.55	**0.59**	0.56
Pride	111	**0.86**	**0.38**	**0.52**	0.58	0.34	0.43	**0.86**	**0.38**	**0.52**
Realization	1,110	**0.31**	0.21	**0.25**	0.21	0.21	0.21	0.27	**0.23**	**0.25**
Relief	153	0.00	0.00	0.00	**0.28**	**0.25**	**0.26**	0.00	0.00	0.00
Remorse	545	0.54	**0.84**	0.66	**0.58**	0.83	**0.68**	0.54	0.82	0.65
Sadness	1,326	**0.53**	0.59	0.56	0.50	0.59	0.54	**0.53**	**0.61**	**0.57**
Surprise	1,060	0.52	**0.57**	0.54	0.50	0.55	0.52	**0.53**	**0.57**	**0.55**

6. CONCLUSION

This thesis investigated multiple deep learning models that incorporate emoji processing in their architecture. We tested these models on benchmark emotion classification datasets that use emojis, namely SemEval'18 and GoEmotions. In addition to that, we tested on an emotion dataset that does not contain emojis, CancerEmo. We found that the performance of these models is comparable to state-of-the-art models and that for the most part, they benefit from processing emojis as input tokens. The improvement is not only notable on the datasets that make the most use of emojis, but also on the ones in which emojis are scarce thus showing the representational power of emojis in this particular task.

One potential direction for future work is to pre-train our text encoders on a language modeling task that also processes emojis in order for the models to better learn how emojis are used in general, not just on the specific dataset it's being trained on. This would provide consistent performance on the tasks.

Another potential direction for future research is pretraining the models in a different domain dataset and train on a target dataset to study how emotions are expressed in different domains and which ones are shared between them. For instance, pretraining on the Reddit domain dataset and training on the Twitter domain dataset. Moreover, training the model on a medical focused dataset to not only better understand patient's expression of emotions, but also understand how these expressions of emotion differ from the general domain.

Recently, the emoji standard started to support the selection of skin tones to many commonly used emojis. Some possible direction for future work is expanding our models to properly address these emojis whose skin tone can change. It would also be interesting to study how the perception of emotion changes with a variation on the emoji's skin tone and what type of bias are learned from how we use emojis.

REFERENCES

[1] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, p. 5, Jun. 2015, ISSN: 2196-1115. DOI: 10.1186/s40537-015-0015-2.

[2] R. Gelbard, R. Ramon-Gonen, A. Carmeli, R. M. Bittmann, and R. Talyansky, "Sentiment analysis in organizational work: Towards an ontology of people analytics," en, *Expert Systems*, vol. 35, no. 5, e12289, 2018, ISSN: 1468-0394. DOI: https://doi.org/10.1111/exsy.12289.

[3] K. L. Kermanidis and M. Maragoudakis, "Political sentiment analysis of tweets before and after the Greek elections of May 2012," *International Journal of Social Network Mining*, vol. 1, no. 3-4, pp. 298–317, Jan. 2013, Publisher: Inderscience Publishers, ISSN: 1757-8485. DOI: 10.1504/IJSNM.2013.059090.

[4] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention," *JMIR Public Health and Surveillance*, vol. 3, no. 2, Jun. 2017, ISSN: 2369-2960. DOI: 10.2196/publichealth.7157.

[5] R. Kelly and L. Watts, "Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships," *Experiences of technology appropriation: Unanticipated users, usage, circumstances, and design*, vol. 2, 2015.

[6] N. Ljubešić and D. Fišer, "A global analysis of emoji usage," in *Proceedings of the 10th Web as Corpus Workshop*, 2016, pp. 82–89.

[7] H. Miller, D. Kluver, J. Thebault-Spieker, L. Terveen, and B. Hecht, "Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[8] O. Udochukwu and Y. He, "A Rule-Based Approach to Implicit Emotion Detection in Text," en, in *Natural Language Processing and Information Systems*, C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais, Eds., vol. 9103, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 197–203, ISBN: 978-3-319-19580-3 978-3-319-19581-0. DOI: .

[9] E. M. van den Broek-Altenburg and A. J. Atherly, "Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season," en, *Applied Sciences*, vol. 9, no. 10, p. 2035, Jan. 2019, Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/app9102035.

[10] H. Meisheri and L. Dey, "TCS Research at SemEval-2018 Task 1: Learning Robust Representations using Multi-Attention Architecture," en, in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 291–299. DOI: 10.18653/v1/S18-1043.

[11] T. LeCompte and J. Chen, "Sentiment analysis of tweets including emoji data," in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2017, pp. 793–798.

[12] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[13] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[14] F. Sebastiani and A. Esuli, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, pp. 417–422.

[15] V. Duppada, R. Jain, and S. Hiray, "SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets," en, *arXiv:1804.06137 [cs]*, Apr. 2018, arXiv: 1804.06137.

[16] S. Mohammad and P. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, CA: Association for Computational Linguistics, Jun. 2010, pp. 26–34.

[17] P. Agrawal and A. Suri, "NELEC at SemEval-2019 Task 3: Think Twice Before Going Deep," *arXiv:1904.03223 [cs]*, Apr. 2019, arXiv: 1904.03223.

[18] F. M. Plaza-del-Arco, M. D. Molina-González, M. Martin, and L. A. Ureña-López, "SINAI at SemEval-2019 Task 3: Using affective features for emotion classification in textual conversations," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 307–311. DOI: 10.18653/v1/S19-2053.

[19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[20] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615–1625, 2017, arXiv: 1708.00524. DOI: 10.18653/v1/D17-1169.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv:1409.0473 [cs, stat]*, May 2016, arXiv: 1409.0473.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," en, *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805.

[23] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[24] C. Baziotis, N. Athanasiou, A. Chronopoulou, *et al.*, "NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning," en, in *arXiv:1804.06658 [cs]*, arXiv: 1804.06658, Apr. 2018.

[25] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," en, in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1–17. DOI: 10.18653/v1/S18-1001.

[26] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou, and O. Zaïane, "Seq2Emo: A Sequence to Multi-Label Emotion Classification Model," en, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 4717–4724. DOI: 10.18653/v1/2021.naacl-main.375.

[27] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," *arXiv:2005.00547 [cs]*, Jun. 2020, arXiv: 2005.00547.

[28] T. Hu, H. Guo, H. Sun, T.-v. T. Nguyen, and J. Luo, "Spice up Your Chat: The Intentions and Sentiment Effects of Using Emoji," *arXiv:1703.02860 [cs]*, Mar. 2017, arXiv: 1703.02860.

[29] W. Ai, X. Lu, X. Liu, N. Wang, G. Huang, and Q. Mei, "Untangling Emoji Popularity Through Semantic Embeddings," en, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, Apr. 2017, Number: 1, ISSN: 2334-0770.

[30] P. Delobelle and B. Berendt, "Time to Take Emoji Seriously: They Vastly Improve Casual Conversational Models," en, *arXiv:1910.13793 [cs]*, Oct. 2019, arXiv: 1910.13793.

[31] R. Plutchik, "A general psychoevolutionary theory of emotion," en, in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds., Academic Press, Jan. 1980, pp. 3–33, ISBN: 978-0-12-558701-3. DOI: 10.1016/B978-0-12-558701-3.50007-7.

[32] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, May 1992, ISSN: 0269-9931. DOI: 10.1080/02699939208411068.

[33] A. Cowen, D. Sauter, J. L. Tracy, and D. Keltner, "Mapping the Passions: Toward a High-Dimensional Taxonomy of Emotional Experience and Expression," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 69–90, Jul. 2019, Publisher: SAGE Publications Inc, ISSN: 1529-1006. DOI: 10.1177/1529100619850176.

[34] S. M. Mohammad and S. Kiritchenko, "Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories," en, p. 12,

[35] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," *arXiv:1808.06226 [cs]*, Aug. 2018, arXiv: 1808.06226.

[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," en, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[37] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, ser. MCS '00, Berlin, Heidelberg: Springer-Verlag, Jun. 2000, pp. 1–15, ISBN: 978-3-540-67704-8.

[38] J. H. Park, P. Xu, and P. Fung, "PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and #hashtags," en, *arXiv:1804.08280 [cs]*, Apr. 2018, arXiv: 1804.08280.

[39] N. Chinchor, "MUC-4 evaluation metrics," in *Proceedings of the 4th conference on Message understanding*, ser. MUC4 '92, USA: Association for Computational Linguistics, Jun. 1992, pp. 22–29, ISBN: 978-1-55860-273-1. DOI: 10.3115/1072064.1072067.

[40] D. J. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants," EN, *Science*, Oct. 1960, Publisher: American Association for the Advancement of Science.

[41] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701.