A study of Emotion Detection on COVID-19 related Misinformation on Twitter

Dimitrios Mylonas and Matthew Purver

School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.

*Corresponding author(s). E-mail(s): dimitriosmylonas@outlook.com; m.purver@qmul.ac.uk;

Abstract

Increase in social media usage during the COVID-19 pandemic has also given rise to misinformation and fake news. Misinformation and fake news cause powerful emotional reactions in the public, which can potentially aid their propagation. Analysing the emotional state of the public towards fake news can help us gain insight about the rapid spread of misinformation and ways to combat it. Emotions are complex and emotion detection remains a difficult task. This paper explores models for multi-label emotion classification which utilise the BERT architecture as well as emotion lexicon information to accurately analyse the emotional responses of the public to known COVID-19 misinformation tweets. Experiments on the SemEval 2018, Task 1 dataset demonstrate the model's strength, achieving a state of the art result on subtask E-c for English tweets. Using the model, we find that misinformation provokes significant emotional responses from people on Twitter, especially negative emotions like anger and disgust, and it also creates mistrust towards potentially influential accounts such as government officials.

1 Introduction

Social media websites, such as Twitter, have made it easier than ever to share views and opinions with others. With the rise of deep learning, the NLP community has taken a great interest in the information that can be extracted from this large amount of data. Due to their short length (280 characters), tweets have a very different structure than typical text. Furthermore, they are usually in informal tone, and make uses of hashtags and emojis, creating further challenges for research. However, the writing style of tweets, results in emotionally rich pieces of text, that make them useful in analysing sentiment and emotion.

Large scale crises such as the COVID-19 pandemic [1], greatly affect the emotional states of the public. Furthermore, due to the nature of COVID-19, social media usage has dramatically increased, resulting in a vast amount of discourse online. This has made COVID-19 a great case study for NLP researchers, with widely available, emotionally rich textual data.

Amidst this crisis, and with the increased amount of information that comes with it, cases of misinformation have also spread online. Although lacking an accepted definition, misinformation is broadly defined as information that is false [2]. Many studies such as [3, 4], have taken an interest in analysing and combating COVID-19 misinformation using NLP techniques. It has been claimed that misinformation propagates the spread of COVID-19 [5] and that the most successful misinformation plays into people's emotions [6].

Misinformation can potentially have major negative effects, especially in the emotional states of the public. Even when spread without any malicious intent, it could lead to the formation of a negative opinion towards the propagator by the public. This is especially important and harmful when it comes to the reputation of authoritative and influential figures on social media. There is a clear benefit for policy makers and researchers in understanding the emotional states triggered by such misinformation.

Emotions are a fundamental part of human behaviour and communication, arguably shaping every decision we make on a daily basis. They are also extremely complicated, with humans being able to experience hundreds of different emotions. According to categorical theories of emotion [7, 8], emotions are correlated to others, with some emotions appearing more frequently together than others. Due to these properties, a solid understanding of the public's emotional state could help us better understand the effect that misinformation has on people and the motivations behind some of their behaviour.

Our main objective is to develop an accurate emotional detection tool to investigate the role of emotions in misinformation spreading. In this paper we make the following contributions:

1. Inspired by the success of Deep Language Models across many NLP tasks, we explore different models for multi-label emotion classification, fine-tuned on the SemEval 2018, Task 1, E-c dataset [9]. We achieve state of the art results on this challenge dataset.
2. We fine tune a further model for detection of the 8 main Plutchik emotions using the SemEval and CovidEmo [10] datasets in the context of COVID-19.
3. Using the Twitter API, we create a dataset of 38000 replies to 1500 tweets known to contain misinformation[1]. These misinformation tweets are taken from the publicly available dataset for COVID-19 Misinformation on Twitter [3].
4. We deploy our emotion classifier to analyse the emotional responses of the public to tweets that contain misinformation. This helps us further understand the rapid spread of misinformation online and generate insights about the emotional reactions of people to such content.

2 Related Works

2.1 Misinformation and Emotion

Emotion plays an important role in how misinformation propagates and how successful it is, and multiple works have studied the relationship between fake news and emotion.

In [11], the authors attempt to understand the psychological processes behind people's susceptibility to fake news. They find that emotionally provocative headlines lead to diminished truth discernment by the participants. The only exception to this is the emotion of anger, which leads to an increased truth discernment instead.

A number of studies have also looked at how emotion can aid fake news detection. [12] explore a multi-task approach for fake new detection from premise-hypothesis pairs, combining information about a texts novelty, emotion and sentiment. Similarly, [13] demonstrate the performance improvement in fake news detection by combining BERT word-embedding features and emotional features. However, this study only looks at the emotions present in the original article, and not in the responses of the public.

A study which chooses to focus on which emotions fake news evoke in the reader is [14]. They propose that there is a relationship between the emotions present in the original article and the emotions in the comments aroused in the crowd. They conclude that this proposed ”dual emotion” feature outperforms state of the art emotional features, and is easily compatible with existing fake news detectors to improve their performance. However, unlike our study, this study does not analyse the accuracy of their emotion detector nor does it investigate the specific emotions present in the text.

2.2 Emotion Detection

Emotions are usually represented in either a discrete or dimensional form. In models such as

[1] In accordance with Twitter terms and conditions the dataset contains only the Tweet IDs of the tweets and will be publicly released after the acceptance of the paper

Ekman's [8] and Plutchik's [7], emotions are represented as either present or not. In the Ekman model there are 6 discrete emotions (happiness, sadness, anger, disgust, surprise, and fear) and in the Plutchik model there are 4 pairs of opposite emotions (joy vs. sadness, surprise vs. anticipation, trust vs. disgust, and anger vs. fear).

In dimensional models such as in the Circumplex of Affect proposed by [15], where emotions are 3-dimensional, with valence, arousal and dominance scores representing each emotion.

Emotion recognition (ER) is a branch of Sentiment Analysis (SA) concerned with extracting fine-grained emotions from text. Due to the difficulty of creating labelled datasets and the ambiguity of emotions present in text in general, most research has focused on classification of overall sentiment or on a limited amount of emotions.

Before the advent of deep learning, SA/ER systems were built by extracting hand crafted features or using lexicon features [16] and feeding them into classical machine learning algorithms such as Support Vector Machines (SVM) [17, 18].

With the rise of deep learning, the need for such laborious methods was replaced by methods the neural networks ability to learn directly through data. Advances in language models have culminated in pre-trained models such as BERT [19] dominating the tasks of ER and SA and achieving state of the art in a number of other NLP tasks. Some examples of models using BERT for emotion detection are [20] and [21].

2.3 SemEval

A major contribution towards the progress of NLP has been made by the SemEval workshops, organising tasks and datasets with the mission to advance state of the art in semantic analysis.

SemEval tasks in the past have focused on sentiment analysis (SA) or emotion recognition (ER). The most notable ones, focusing on tweets, are the SemEval 2017 Task 4 [22] and the SemEval 2018 Task 1 [9], concentrating in SA and ER respectively.

In sub-task 5 of SemEval 2018 Task 1, 10983 English tweets are labeled with 11 emotions (Plutchik emotions and love, optimism, and pessimism). Emotion categories do not occur in the same proportions, with emotions such as anticipation, pessimism and trust being under-represented. The aim is to build multi-label classification systems, with 75 teams taking part in the competition.

Notable teams include:

1. The winners of the competition, NTUA-SLP [23], make use of a Bi-LSTM architecture, with a multi-layer self attention mechanism and word2vec [24] word embeddings, trained on 550M tweets. They also pre-process tweets using the ekphrasis tool [25] and chose to pre-train on the semantically similar dataset of SemEval 2017 on tweet sentiment analysis.
2. Second place, TCS Research [26], who combined features from deep learning models, lexicons and pre-trained models, passing them into a Support Vector Classifier.

Since the end of the competition, a number of different works have improved on these approaches. EmoGraph, presented in [27], is a system which uses graph neural networks to better model the dependencies between different emotions. Their analysis shows that under-represented emotion classes such as ”trust”, can greatly benefit from the correlation with other emotions.

Building on the prowess of BERT in NLP tasks [28] propose Domain Knowledge BERT (DK-BERT), a BERT LM system fine-tuned on data from the target domain (in this case tweets). They find that supplementing BERT with extracted domain-specific features improves performance by one percent over previous state of the art BERT approaches.

More recently, achieving state of the art performance is SpanEmo, by [29]. This model casts multi-label prediction as a span-prediction problem. The model takes into consideration both the input sentence and a label set of emotion classes for selecting the span of emotion as an output.

2.4 Emotion Detection during COVID-19

The COVID-19 virus has majorly disrupted daily life, triggering a wide range of emotional reactions from the public. Social media platforms such as Twitter have seen increased engagement, with people sharing their feelings and opinions during this global crisis. Subsequently, Twitter has become a useful source of data to analyse the public opinion and sentiments.

A number of studies have worked to make COVID-19 databases available. [30] have collected and open sourced over 4 million tweets a day since the start of the pandemic. Another dataset, COVIDSenti [31], contains 90000 tweets, labelled for positive, negative, or neutral sentiment.

Multiple studies have focused on detecting emotion during COVID-19. [32] conduct a study on sentiment of people in China, and [33] present a similar study on Saudi Arabian tweets. [34] propose EmoBERT, a BERT based model for multi-label emotion recognition, creating an analysis on emotions of the public in London, UK. Similarly, [10] utilise a number of different BERT models for emotion recognition during covid, and create a Plutchik emotions labelled dataset of tweets.

With an increased amount of information online, misinformation and panic have also spread. [35] show that panic created by postings online spreads faster than the COVID-19 virus. Similarly, [3] show that tweets containing misinformation propagate faster than tweets that do not. They also create a dataset of 1500 tweets that are known to contain misinformation. We use this dataset to study the emotional response of the public towards those tweets, by analysing the replies directed towards them.

3 Methodology

3.1 Datasets

The main experimentation is carried out on SemEval 2018, Task 1, E-c [9]. This multi-label emotion classification dataset contains 10983 English tweets divided into training set (6838 tweets), validation set (886 tweets), and testing set (3259 tweets). The tweets are labelled with 11 different emotion categories (the 8 Plutchik emotions along with 3 additional emotions love, optimism, and pessimism). In line with [9] we use Jaccard Accuracy, Micro F1, and Macro F1 as our evaluation metrics.

To fine-tune the model on emotion-labeled COVID-19 tweets, we utilise the CovidEmo dataset [10]. The currently publicly available dataset contains 1600 English tweets, annotated with the 8 Plutchik emotions and is temporally distributed across 18 months of the Covid-19 pandemic. We split this dataset in a 62/8/30 split and concatenate it with the SemEval training, validation and testing sets.

The Twitter analysis section of the paper was carried out on a self collected dataset of tweets replying to known misinformation Twitter posts. Using the dataset from [3], which contains around 1500 professionally fact-checked COVID-19 related tweets that are false or partially false, we create a dataset containing the replies to those tweets. The replies were fetched using the Twitter API, collecting only tweets in English, which resulted in a dataset of 38000 tweets.

3.2 Implementation Details

3.2.1 BERT

Motivated by the success of deep language models across multiple NLP tasks, we choose to utilise variants of BERT [19], namely $BERT_{base}$ and $BERT_{large}$. $BERT_{base}$ is a 12 attention head transformer model, containing 110 million parameters, and resulting in 768-dimensional representations. $BERT_{large}$, containing 24 attention heads, 340 million parameters and results in a 1024-dimensional representation.

Furthermore, due to the nature of our task, we choose to use 2 variants of BERTweet [36], a BERT model pre-trained on over 850 million tweets. Namely we use, $BERTweet_{base}$, a 135 million parameter model, and $BERTweet_{large}$, a 350 million parameter model.

All of the pre-trained models are obtained from the HuggingFace website [37].

3.2.2 Data Processing

Before being passed into BERT, the data is first tokenised using the respective tokenisers and truncated/padded to 64 tokens, a small length chosen as tweets are restricted in their size. A classification $[CLS]$ token and a sentence separator $[SEP]$ token were added to the start and end of the input respectively. Attention masks are also created to help the model identify which tokens are padding. For all sets, the data shuffled, and batched with a batch size of 16, and for the training set it is repeated 3 times, which was found to have a positive effect.

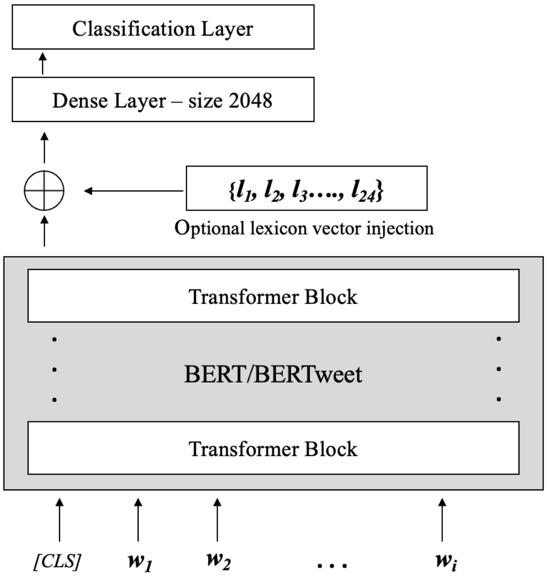

Fig. 1 Our model architecture, with the optional injection of lexicon features.

When using variants of BERTweet, the same method of tokenisation was used but an additional step of pre-processing the data was carried out. This pre-processing step was in accordance with [36]. Tokens were made lower case, emojis were replaced by their :code: equivalents, URLs were replaced by a "HTTPURL" token, and users were kept anonymous by replacing all mentions with "@USER".

3.2.3 Model Architecture

Fig. 1 displays our model architecture . We carry out a standard fine tuning method on the pre-processed data. Each tokenised input sequence is passed into the chosen encoder BERT. Subsequently, we use the $[CLS]$ token output (a representation of full sentence features) from the final BERT layer to feed into a fully-connected layer with 2048 nodes, which then feeds into a final classification layer with as many nodes as possible emotion labels.

As we are focusing on a multi-label classification problem, we choose to minimise the Binary Cross-Entropy (BCE) loss. For labels y and predictions p, the BCE loss is defined as

$$Loss_{BCE} = -(y\log(p) + (1-y)\log(1-p)) \quad (1)$$

We calculate a separate loss for each emotion class c per observation o and sum the result, leading to

$$Loss_{BCE} = -\sum_{c=1}^{M} y_{o,c}\log(p_{o,c}) \quad (2)$$

We use the stochastic optimiser Adam [38] throughout our experiments. For the fully connected layer of the model we test 3 different activation functions. Namely, Sigmoid

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (3)$$

Rectified Linear Unit (ReLU) [39]

$$Relu(z) = max(0, z) \quad (4)$$

and Gaussian Error Linear Unit (GELU) [40]

$$Gelu(z) = z * \Phi(z) \quad (5)$$

where $\Phi(z)$ is the cumulative distribution function of Gaussian distribution.

The last classification layer, uses a sigmoid function as in Eq. 3.

3.2.4 Cyclical Learning Rates

Inspired by the work in [41], which shows that the use of Cyclical Learning Rates (CLR) is beneficial to the learning process, we choose to test the effectiveness of CLR in our model. CLR is an optimiser policy where learning rates "oscillate" between two bounds linearly through cycles. The parameter stepsize in CLR is the number of iterations in a half cycle.

[42] argue that saddle points give rise to the main difficulties in minimising the loss (rather than local minima), as they slow down the learning process because of their small gradients. By varying the learning rate between two bounds we can more rapidly get out of such saddle points, while retaining the benefits of a small learning rate for converging to the minimum. Another benefit of CLR is that, unlike adaptive learning rates, they incur no major computational cost.

By training the model with flat learning rates in a wide range, we estimate that the optimal learning rate lies within $2e-5$ and $2e-7$ and therefore choose these values as the bounds for CLR. Following [41], which states that optimal step sizes are 2-10 times the number of iterations in an epoch, we experiment with $stepsize =$

$\{2, 4, 8\} * epoch$ for our models. We use Tensorflow [43] to implement our system, and train our models with 3 epochs and batch size of 16.

3.3 Lexicon Information

To mitigate the lack of labelled data for emotion detection, we explore the addition of lexicon information into BERT, to aid our models performance. Similar methods have found success in fields like abuse detection [44] and sarcasm detection [45].

To integrate lexicon information, we follow a method similar to [46]. We create a lexicon vector using 3 different open-source emotion lexica, NRC Emotion [47], NRC Hashtag Emotion [48], and the Multidimensional Lexicon of Emojis [49]. Through a lookup for each word/emoji in a target sentence, lexicon values are obtained and averaged over the total words in the sentence. The resulting vector is 24-dimensional, representing 8 emotion scores for each of the lexica. A score of 0 for all emotions is assigned to words/emojis not appearing in any of lexica.

We inject the lexicon vector directly into BERT during the fine tuning process on emotion detection. We concatenate the $[CLS]$ token with our lexicon vector and pass this joined vector through the pre-classification Dense layer leading to the classification layer.

3.4 Experiments

The first part of experimentation focuses on improving performance on the SemEval Task 1, E-c dataset. As mentioned we explore 5 different BERT models in combination with our architecture mentioned in Section 3.2.3. During this stage we choose the optimal parameters to maximise the models performance and we also explore the performance after injecting lexicon information.

In the second part, we restrict our SemEval dataset to the 8 Plutchik emotions and combine it with the COVIDEmo dataset. This is done to create an emotion classifier with not only domain knowledge of Twitter, but also context of COVID-19. Here we keep the architecture and hyper-parameters of our best performing model on the full SemEval dataset, to train this new classifier from scratch. We test its robustness on a test set that contains both SemEval and CovidEmo data points, as mentioned in Section 3.1.

Lastly, we apply our emotion classifier to perform emotion detection on our dataset of replies to misinformation tweets from Section 3.1, as well as on the misinformation tweets themselves, and analyse the results.

4 Results

4.1 Metrics

First, we define the metrics used throughout the paper. Following the standard set by the SemEval 2018 Task 1 [9], we report our results for Micro-F1, Macro-F1, and Jaccard accuracy.

Given predictions and gold labels, a number of True Positives (TP), False Positives (FP), and False Negatives (FN) can be found. The F1 score, is a method to evaluate model performance by using the values for $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$ of the system. F1 score is defined as

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{6}$$

For multiple classes c, we can find an average F1 score, using macro or micro averaging. The macro method weighs classes equally by calculating an average on each class and summing. On the other hand, micro averaging accounts for class imbalances by calculating the sums of TP, FP, and FN first and calculating the micro-F1 score using those.

We also use Jaccard similarity/accuracy throughout this paper. Jaccard similarity is a metric to judge the similarity between two sets. In our case, given a set of predicted labels P and true labels Y for a sentence, Jaccard similarity is the size of their intersection divided by their union.

$$Jaccard(P, Y) = \frac{\mid P \cap Y \mid}{\mid P \cup Y \mid} \tag{7}$$

Our aim is to maximise this value, and so the thresholds we choose for classification of a positive case of an emotion reflect that. For this task, there is not a crystal clear preference between the sensitivity and the specificity and so we strike for a balance. It can be argued, that since emotions occur in opposite pairs, it is unlikely that multiple emotions will occur together (more than 3-4

Table 1 Results of experimentation on SemEval test set with $BERT_{base}$ and $BERT_{large}$.

Model	Learning rate	Micro-F1	Macro-F1	Jaccard
$BERT_{base}$	$2e-5$	0.685	0.559	0.560
	$2e-6$	0.705	0.554	0.586
	CLR stepsize=8	0.706	0.572	0.586
	CLR stepsize=4	0.711	0.572	0.592
	CLR stepsize=2	0.703	0.523	0.584
$BERT_{large}$	$2e-5$	0.475	0.236	0.320
	$2e-6$	0.706	0.555	0.587
	CLR stepsize=8	0.709	**0.575**	0.589
	CLR stepsize=4	0.710	0.574	0.590
	CLR stepsize=2	**0.712**	0.559	**0.593**

Table 2 Results of experimentation on SemEval test set with $BERTweet_{base}$ and $BERTweet_{large}$.

Model	Learning rate	Micro-F1	Macro-F1	Jaccard
$BERTweet_{base}$	$2e-5$	0.598	0.472	0.478
	$2e-6$	0.663	0.472	0.540
	CLR stepsize=8	0.696	0.539	0.579
	CLR stepsize=4	0.686	0.510	0.569
	CLR stepsize=2	0.639	0.429	0.512
$BERTweet_{large}$	$2e-5$	0.728	0.596	0.610
	$2e-6$	**0.733**	0.593	0.614
	CLR stepsize=8	0.732	**0.598**	0.618
	CLR stepsize=4	**0.733**	0.594	**0.619**
	CLR stepsize=2	0.732	0.593	0.617

emotions per sentence), and so minimising our False Negatives is essential. Our chosen thresholds reflect that, usually being around 0.35 to classify an emotion as positive.

4.2 SemEval

Here, we present the results of the experimentation on the SemEval 2018 dataset.

First, we compare the performance of $BERT_{base}$ and $BERT_{large}$. Our models are built and trained as described in Section 3.2 and we use a sigmoid activation for fully connected layer. We explore the use of flat learning rates and CLR. The results can be seen in Table 1.

There are a few main takeaways from this part of experimentation.

- $BERT_{large}$ appears to be performing better than $BERT_{base}$, achieving the best results in all metrics, although not by a great margin. This is expected as $BERT_{large}$ contains 3 times the parameters and so can model the data with a higher complexity. However, this model takes a significantly longer time to fine tune.
- Training the models using CLR almost always increases performance in comparison to a flat learning rate for both $BERT_{base}$ and $BERT_{large}$.
- There is no clear CLR step size that performs better than others. We observe that each of the 3 best metrics achieved are spread between the 3 step sizes evenly.

Next, we explore the performance of $BERTweet_{base}$ and $BERTweet_{large}$ in a similar fashion, with a sigmoid activation for the fully connected layer. The results can be seen in Table 2. The observations are the following

- The *large* variant of BERTweet outperforms *base* again, this time by an obvious margin.
- For all learning rates, $BERTweet_{large}$ performs at a similar level, with CLR improving on the metrics only marginally. The positive effects of CLR are however still present, both here and in the $BERTweet_{base}$ results.
- Even though $BERTweet_{base}$ has twitter domain knowledge, it does not outperform $BERT_{base}$. This is not the case however for the *large* variants. $BERTweet_{large}$ here is dominant over its vanilla BERT counterpart.

It is evident that $BERTweet_{large}$, is superior over other variants of BERT. Although there is a small variation between the performance for different learning rates, we choose to adopt CLR with stepsize 4 for the rest of the paper, as it achieves the best results.

Here, we report the results of the experimentation of Lexicon features. The different models we test are a $BERTweet_{large}$ model with the NRC Emotion and NRC Hashtag Emotion lexicon features, which we call $BERTweet_{word}$, and another with the additional Multidimensional Lexicon of Emojis added, which we call $BERTweet_{word+emoji}$. The results are seen in Table 3.

We observe that $BERTweet_{word}$ performs slightly better than the other variants, however, it is clear that the difference between these models is not significant to justify that lexical injection is beneficial to the model. We therefore choose to abandon the use of lexicon features in the rest of the paper.

Table 3 Results of experimentation on SemEval test set with $BERTweet_{large}$ and lexicon features

Model	Micro-F1	Macro-F1	Jaccard
$BERTweet_{large}$	0.733	**0.594**	0.619
$BERTweet_{word}$	**0.736**	**0.594**	**0.620**
$BERTweet_{word+emoji}$	0.733	0.590	0.618

Table 4 Results of experimentation on SemEval test set with $BERTweet_{large}$ and 3 different activation functions.

Activation Function	Micro-F1	Macro-F1	Jaccard
Sigmoid	0.733	0.594	0.619
ReLU	**0.738**	0.595	**0.624**
GELU	0.736	**0.597**	0.623

Table 5 Results of emotion classification on SemEval-2018 test set.

Model	Micro-F1	Macro-F1	Jaccard
$NTUA$	0.701	0.528	0.588
$BERT_{base} + DK$	0.713	0.549	0.591
$SpanEmo$	0.713	0.578	0.601
$TCSResearch$	0.693	0.530	0.582
$EmoGraph$	0.707	0.563	0.589
$BERTweet(ours)$	**0.738**	**0.595**	**0.624**

Our last step of improving the model, involves testing different activation functions for the fully connected layer before our classification layer. So far, the Sigmoid function has been used, but we compare its performance against the GELU and ReLU activation functions. The results are seen in Table 4. We can see that the ReLU activation function makes an improvement of 0.5 percent and so we choose to adopt it moving forward.

4.3 COVID-19 BERTweet

Using the best performing architecture on SemEval, $BERTweet_{large}$ with ReLU activation function and trained using CLR with step size 4, we train a new classifier using the SemEval dataset combined with the COVIDEmo dataset.

The test set, as described in Section 3.1, now contains both the test set of SemEval as well as 400 tweets that are COVID-19 specific. We test if the classifier will retain robustness trained on this new data.

The metrics produced by this new model on the test set are a Micro-F1 score of 0.732, a Macro-F1 of 0.594 and a Jaccard accuracy of 0.652. These results indicate that this model has retained performance even with this new dataset added. However, a high score such as this is expected, as 3 of the 11 categories have now been discarded, 2 of them (love and pessimism) being minority classes.

Looking at the Area Under Curve (AUC) of the Retriever Operating Characteristic (ROC) curve for each of the 8 classified emotions, and comparing it with the same values for our SemEval model we can see that performance is retained on an individual emotion basis. These results are seen in Table 6.

It is evident that performance decreases slightly when adding COVIDEmo, even though there are more training samples. This could imply that emotion co-occurrence, has a large effect on how well the classifier performs, and that the lack of 3 emotion categories leads to a decreased overall performance. For example the absence of pessimism labels could be responsible for the decrease in AUC score for sadness and fear, as they are emotions that are likely to co-occur. This classifier is nonetheless proficient at detecting emotion given how difficult a task it is, and so it is used in this paper to analyse emotion on Twitter.

5 Analysis

5.1 SemEval

The model we have created outperforms previous approaches on multi-label emotion detection on the SemEval dataset by a Jaccard Similarity of 2.3%, a Micro-F1 score of 2.5% and Macro-F1 score of 1.7%. The full comparison of our method against previous methods can be seen in Table 5.

Our method is most similar to $BERT_{base}+DK$ [28], as it also attempts to tackle the challenge by introducing domain specific knowledge to the classifier to aid with the heterogeneity of informal language of tweet data. In $BERT_{base} + DK$, after pre-processing tweets in a special manner, a convolutional token pattern detector is used to sift through tweets to obtain domain knowledge, which is then integrated within BERT to supplement its domain-specific knowledge.

In our method, instead of concatenating general BERT features and extracted domain features, we use BERTweet, which is initially trained on Twitter specific tokenised data. This means that our model contains intrinsic knowledge about the target domain from the start, and this results in a greater performance across all metrics. This is in line with the performance benefits found in the original BERTweet study [36], which showed

Table 6 AUC ROC curve scores for each emotion class in both SemEval and SemEval+COVIDEmo test sets.

Model	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Love	Optimism	Pessimism
SemEval	**0.936**	0.763	**0.917**	**0.940**	**0.953**	**0.916**	**0.804**	0.790	0.921	0.908	0.842
SemEval+COVIDEmo	0.935	**0.794**	0.910	0.928	0.952	0.903	0.790	**0.807**	None	None	None

7% increases in performance on gender classification and sentiment analysis on Twitter datasets against a simple BERT model.

Adding lexicon information on the other hand did not provide any performance benefits to our model. A similar result is seen in [46], where they achieve no added performance injecting lexicon information directly into the transformer for emotion detection in Dutch. A likely explanation is that using only 3 affect lexica, did not provide any additional knowledge to the LM that it did not already possess. In a similar vein, the failure of lexicon features could be that a 24-dimensional vector is not significant enough to have any affect compared to the 1024-dimensional $[CLS]$ vector of $BERT_{large}$.

5.2 Emotion Detection on Twitter Misinformation

In this section we perform an analysis on the emotions present in the dataset for COVID-19 Misinformation on Twitter [3]. and the self-collected dataset containing around 38000 English replies from the original 1500 Misinformation tweets. We perform emotion classification on both the datasets which we will refer to as original and reply datasets respectively.

The proportions of emotions occurring after labelling using our BERTweet model fine tuned on COVID-19 data, can be seen in Fig. 2. Comparing the emotions present in both the datasets, it is clear that emotion is not as prevalent in the original tweets. This is also found by [3] using a lexicon approach to classifying emotion. This is possibly because most misinformation tweets in this dataset are from news outlets, politicians and other high profile accounts. Tweets from such sources tend to stay more ”factual” and use formal language, and so it is intuitive that they do not contain as many emotions overall as the replies to them by the public.

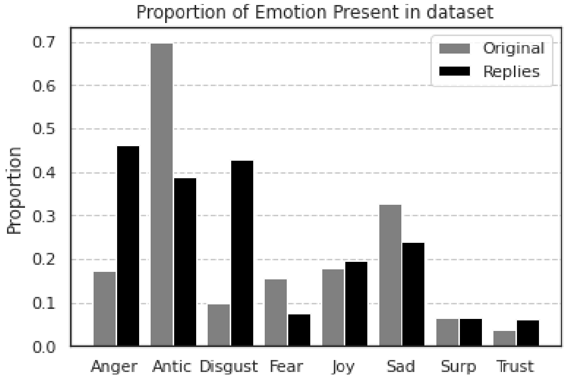

Fig. 2 Proportions of emotions occurring in all tweets for both the replies to misinformation and the original dataset.

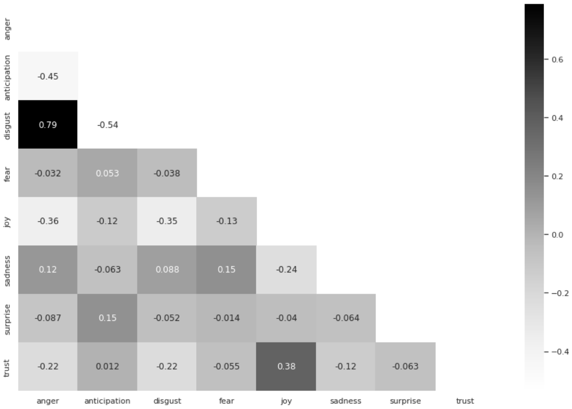

Fig. 3 Values for the Pearson Correlation between each emotion in the Reply dataset.

5.2.1 Anger and Disgust

Focusing on the reply dataset, we can see that negative emotions prevail in the replies towards misinformation tweets. More specifically ”anger” and ”disgust” appear in 46% and 43% of the tweets respectively. By observing Fig. 3, which shows the Pearson Correlation values for all emotion combinations, we can also see that anger and disgust are the most positively correlated emotions with $r = 0.79$, meaning they are most likely to be present together in a tweet.

Comparing with the original tweets which spread COVID-19 misinformation, anger and disgust occur 17% and 10% respectively, and so it is evident that the posts are not conveying these feelings in the post, yet they are provoking them from the public. Without looking at the replies individually, it is not clear whether this emotional response is towards the author for misleading and spreading misinformation or it is simply the expression of feeling towards the claim posted.

5.2.2 Trust and Surprise

An important observation is that the least occurring emotion in the replies to misinformation is "trust", occurring only in 6.1% of tweets. This suggests that the public is generally sceptical of these tweets, and that perhaps the anger and disgust is directed towards the author and not in reaction to the content of the tweet. It is certainly the case that a lot of the tweets are expressing mistrust in this way, with phrases like "fake news", "liar", "lies" appearing more than 1500 times collectively throughout the replies tweet dataset.

Given that some misinformation is trying to trigger an emotional response by presenting a false and sometimes exaggerated claim, it is expected that "surprise" might be an emotion that is present towards such a tweet. However, this is not the case, with "surprise" only appearing in 6.4% of tweets, slightly more than trust. This may suggest that since a lot of replies are mistrusting, they are not surprised by the claim because they do not believe it in the first place.

5.2.3 Sadness and Fear

"Sadness" and "fear" are two emotions that are prevalent in the original misinformation tweets appearing 33% and 16% respectively, yet not as common in the responses, appearing only 24% and 7%. The analysis of the original misinformation tweets in [3], shows that the most common hashtags concern the COVID-19 virus (i.e. #coronavirus and #covid19) or stopping said virus (i.e. #stopcorona).

We can infer then that the the high proportion of sadness and fear emotions in original tweets, is indicative of the message that the tweets are trying to convey. However, the low levels of fear and sadness in the responses, indicate that those tweets were not successful in provoking these emotions. Fear and sadness is another emotion pair that has a significant positive correlation compared to other emotion pairs.

5.2.4 Joy and Anticipation

The final pair to consider are the emotions of "joy" and "anticipation". Even though both are generally "positive" feelings, they are not positively correlated with each other, with a $r = -0.12$. Joy, is an emotion which we expect to see, as misinformation is not always negative, and it could trigger joy in people based on their views. It is highly possible that the news and politician accounts that posted the misinformation, receive replies from people that follow those accounts, and are therefore more likely to share similar political views.

Anticipation is the 3rd highest proportion emotion in the replies set and 1st highest in the original tweet set, reaching 70%. The tweets come from an early time of COVID-19, January-July 2020, a time of uncertainty and novelty for everyone in the world. It is expected that in such a situation, a lot of news and emotional responses are anticipatory, whether this is expecting new measures to take place, looking forward to a lockdown being lifted, or simply anticipating the end of the pandemic.

In [3], the authors find that confirmed misinformation propagates faster than other tweets (by tracking number of retweets within dataset timeframe). Given our analysis we can see why the statement that "successful misinformation plays into people's emotion" [6] stands. The misinformation tweets provoke unnatural levels of anger and disgust, and certainly increasing the public's engagement with the tweet.

A lot of research is justly concerned with tackling misinformation. As we have seen, it can incite negative emotional responses in the public, and mislead people. It is certainly worrying that the emotion "trust" is the least commonly present in replies to authoritative accounts present in this dataset such as politicians. As [3] state "In order for authorities to maintain information sovereignty, users – in this case typically citizens – need to trust the authorities.". In crisis situations such as a global pandemic, where compliance with government regulations and cooperation with each other is necessary, a low level of trust, could be detrimental to the well being of many. It is the duty of policy makers and researchers to work to understand the nature of misinformation and the emotional reactions it provokes, and to use this knowledge to improve the management of global crises.

6 Conclusion and Future Work

In this paper, we have explored methods for multi-label emotion classification, and successfully created a model that outperforms current methods on the SemEval 2018 Task 1 dataset on all metrics. We have also used our system to analyse the responses of the public towards misinformation, finding that they provoke emotions of mistrust, anger and disgust in the public, justifying their high rates of propagation.

A limitation of this work is the absence of a more detailed analysis into the relationship between specific tweet and reply sets. Analysing the language and emotions present for each misinformation tweet and its replies separately can unlock a lot more useful insights into the nature of misinformation on Twitter.

Another limitation is the the lack of improvement from injecting lexicon features, which could be resolved by attempting to create larger lexicon vectors by combining more lexica, or attempting a meta-learner approach as in [46], where predictions from BERT and predictions from lexicon features are added into a SVM classifier.

7 Conflict of Interest Statement

Mylonas receives no funding relating to research. Purver received funding for this research from a number of sources: the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103) and the project Sovrag (Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration, J5-3102); the UK EPSRC via the project Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1); and the project RobaCOFI (Robust and adaptable comment filtering), which indirectly received funding from the European Union's Horizon 2020 research and innovation action programme via the AI4Media Open Call 1 issued and executed under the AI4Media project (Grant Agreement no. 951911). This paper reflects only the authors' views; the EC and the AI4Media project are not responsible for any use that may be made of the information contained herein.

References

[1] Catrin Sohrabi, Zaid Alsafi, Niamh O'neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International journal of surgery*, 76:71–76, 2020.

[2] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[3] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104, 2021. ISSN 2468-6964. doi: https://doi.org/10.1016/j.osnem.2020.100104.

[4] Jackie Ayoub, X. Jessie Yang, and Feng Zhou. Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58:102569 – 102569, 2021.

[5] Laurie Garrett. Covid-19: the medium is the message. *The lancet*, 395(10228):942–943, 2020.

[6] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy-making, 2017.

[7] Robert Plutchik. A general psychoevolutionary theory of emotion. theory, research, and experience. In *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: https://doi.org/10.1016/B978-0-12-558701-3.50009-0.

[8] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[9] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in

tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001.

[10] Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea, and Junyi Jessy Li. Emotion analysis and detection during covid-19. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 6938–6947, 2022.

[11] Bence Bago, Leah R. Rosenzweig, Adam J. Berinsky, and David G. Rand. Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion*, 0(0):1–15, 2022. doi: 10.1080/02699931.2022.2090318. URL https://doi.org/10.1080/02699931.2022.2090318.

[12] Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. A multitask learning approach for fake news detection: Novelty, emotion, and sentiment lend a helping hand. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. doi: 10.1109/IJCNN52387.2021.9534218.

[13] Andrew L. Mackey, Susan Gauch, and Kevin Labille. Detecting fake news through emotion analysis. In *Proceedings of eKNOW 2021: the 13th International Conference on Information, Process, and Knowledge Management*, page 65–71. IARIA, 2021. ISBN 978-1-61208-874-7.

[14] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3465–3476, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450004. URL https://doi.org/10.1145/3442381.3450004.

[15] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. doi: 10.1037/h0077714.

[16] Saif Mohammad and Peter 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*, pages 26–34, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[17] Alec Go. Sentiment classification using distant supervision. 2009.

[18] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50 (1):723–762, may 2014. ISSN 1076-9757.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

[20] Kisu Yang, Dongyub Lee, Taesun Whang, Seolhwa Lee, and Heuiseok Lim. Emotionx-ku: Bert-max based contextual emotion classifier, 2019.

[21] Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. Emotionx-idea: Emotion bert - an affectional model for conversation. *ArXiv*, 2019.

[22] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088.

[23] Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1037.

[24] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[25] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment

analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[26] Hardik Meisheri and Lipika Dey. TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 291–299, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1043.

[27] Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. Emograph: Capturing emotion correlations using graph networks. *ArXiv*, 2020.

[28] Wenhao Ying, Rong Xiang, and Qin Lu. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5541.

[29] Hassan Alhuzali and Sophia Ananiadou. Spanemo: Casting multi-label emotion classification as span-prediction. In *EACL*, 2021.

[30] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[31] Usman Naseem, Imran Razzak, Matloob Khushi, Peter W Eklund, and Jinman Kim. Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015, 2021.

[32] Xuehua Han, Juanle Wang, Min Zhang, and Xiaojie Wang. Using social media to mine and analyze public opinion related to covid-19 in china. *International Journal of Environmental Research and Public Health*, 17 (8), 2020. ISSN 1660-4601. doi: 10.3390/ijerph17082788.

[33] Aseel Addawood, Alhanouf Alsuwailem, Ali Alohali, Dalal Alajaji, Mashail Alturki, Jaida Alsuhaibani, and Fawziah Aljabli. Tracking and understanding public reaction during COVID-19: Saudi Arabia as a use case. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.24.

[34] Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra I Cristea, and Lei Shi. Detecting fine-grained emotions on social media during major disease outbreaks: Health and well-being before and during the covid-19 pandemic. In *AMIA Annual Symposium Proceedings*, volume 2021, page 187. American Medical Informatics Association, 2021.

[35] Anneliese Depoux, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. The pandemic of social media panic travels faster than the covid-19 outbreak, 2020.

[36] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, 2019.

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[39] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.*, 20(3–4):121–136, sep 1975. ISSN 0340-1200. doi: 10.1007/BF00342633.

[40] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016.

[41] Leslie N. Smith. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer*

Vision (WACV), pages 464–472, 2017.

[42] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *NIPS*, 2015.

[43] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[44] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.5.

[45] Avinash Kumar, Vishnu Teja Narapareddy, Pranjal Gupta, Veerubhotla Aditya Srikanth, Lalita Bhanu Murthy Neti, and Aruna Malapati. Adversarial and auxiliary features-aware bert for sarcasm detection. In *8th ACM IKDD CODS and 26th COMAD*, CODS COMAD 2021, page 163–170, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450388177. doi: 10.1145/3430984.3431024.

[46] Luna De Bruyne, Orphee De Clercq, and Veronique Hoste. Emotional RobBERT and insensitive BERTje: Combining transformers and affect lexica for Dutch emotion detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 257–263, Online, April 2021. Association for Computational Linguistics.

[47] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3): 436–465, 2013.

[48] Saif M. Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015. ISSN 1467-8640. doi: 10.1111/coin.12024.

[49] Rebecca Godard and Susan Holtzman. The multidimensional lexicon of emojis: A new tool to assess the emotional content of emojis. *Frontiers in Psychology*, 13, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.921388.