# CancerEmo ♥: A Dataset for Fine-Grained Emotion Detection

**Tiberiu Sosea and Cornelia Caragea**
Computer Science Department
University of Illinois at Chicago
tsosea2@uic.edu, cornelia@uic.edu,

## Abstract

Emotions are an important element of human nature, often affecting the overall wellbeing of a person. Therefore, it is no surprise that the health domain is a valuable area of interest for emotion detection, as it can provide medical staff or caregivers with essential information about patients. However, progress on this task has been hampered by the absence of large labeled datasets. To this end, we introduce CancerEmo ♥, an emotion dataset created from an online health community and annotated with eight fine-grained emotions. We perform a comprehensive analysis of these emotions and develop deep learning models on the newly created dataset. Our best BERT model achieves an average F1 of 71%, which we improve further using domain-specific pretraining.

## 1 Introduction

Life-threatening diseases such as cancer and AIDS make people extremely vulnerable and stir a diverse range of feelings and emotions in them, e.g., from fear to trust or joy and from anger to surprise or sadness. These feelings and emotions shape a person's behavior, beliefs, and actions, and many turn to online health communities to share their health concerns and emotions. Recent research shows that this form of sharing is very beneficial to a patient's progress and well-being. For example, Qiu et al. (2011) show that cancer patients feel better and change to positive attitudes when they interact with others during or after the disease. Pollak et al. (2007) show that less anxiety and depression lead to better adherence to cancer care therapies.

The online sharing of emotions in online health communities on topics such as treatment, medication, side effects, moods, and the disease itself, has resulted into a large amount of user-generated content in the form of discussions. This together with the fact that people find it easier to express themselves and reveal personal details in health forums, rather than in a face-to-face context (Kummervold et al., 2002), make online health communities a great place to examine and study patients' emotions at a large scale using computational models.

However, despite that emotion detection has started to emerge in the health domain, the lack of large annotated datasets in the field greatly hinders the capabilities of supervised techniques and limits an understanding of fine-grained expressions of emotions at a large scale. For example, available datasets contain only about $1,000$ sentences annotated with Ekman's six basic emotions. Since some emotions appear very rarely in the annotated set, only the most frequent ones joy and sadness are analyzed (Khanpour and Caragea, 2018).

In this paper, we explore fine-grained emotion detection in online health communities and present a large dataset for this task. Specifically, we introduce CancerEmo ♥, a health-related dataset, composed of $8,500$ sentences annotated with emotions taken out of $25,000$ sentences sampled from an online cancer survivors network. This network, which is designed for patients suffering from cancer, and their caregivers, friends, and families, contains several discussion boards grouped by cancer type, where users can start a discussion thread or comment to messages in an existing thread. We construct our dataset from the breast, lung, and prostate cancer discussion boards, since there are higher stakes involved for patients with this type of disease. For example, breast cancer is the most common women cancer with about 18% of all women's cancers (McPherson et al., 2000); lung cancer is the leading cause of death among men and second among women (Torre et al., 2016), while prostate cancer is the third leading cause of cancer deaths in the United States (Haas et al., 2008). Our dataset is fine-grained, being annotated with



*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8892–8904,
November 16–20, 2020. 

| | |
|---|---|
| SADNESS | I just cant stand seeing her like this. |
| ANTICIPATION<br>TRUST | If they get better, they can opt out of the program. |
| JOY | Could I get spoiled YES...LOL Love to all of you Kay. |
| FEAR | Guess I am more scared cause this has been very speedy. |
| SURPRISE | It is so awesome to hear news like yours! |
| DISGUST<br>ANGER | I hate cancer and I sure hate what it has done to good people, like you. |
| SADNESS<br>FEAR | My cancer was very rare, non invasive Mucusom Cancer. |
| JOY<br>FEAR | Yesterday they told me they didnt see anything which brought tears of joy, but also a wave of fear. |

Table 1: Examples from our dataset.

Plutchick-8 basic emotions (Plutchik, 1980), composed of anger, fear, disgust, sadness, surprise, anticipation, trust, and joy. We use crowd-sourcing and ensure quality control measures to exclude spurious annotations.

Detecting emotions is inherently challenging, requiring a deep understanding of the writer's beliefs and reasoning, especially when dealing with health-related data. To illustrate some of these challenges, we present examples from our dataset in Table 1, and discuss a few patterns. For example, in the sentence *I just cant stand seeing her like this*, we can easily notice the writer's discontent, regardless of the absence of emotion-rich words in its content. Our data also includes a great deal of medical terminology, which adds another layer of complexity to the language used across the discussion boards. For example, in *My cancer was very rare, non invasive Mucusom Cancer*, in order to predict the perceived conveyed emotions - fear and sadness, computational models must distinguish whether *Mucusom Cancer* is a dangerous or harmless disease. In addition, a sentence may be the expression of a mixture of emotions, not just one. We further speculate that distantly supervised techniques focusing on lexical information to collect emotion-rich data (Abdul-Mageed and Ungar, 2017) are unable to capture these subtleties in a health domain, and we reinforce this idea in §3.

Our contributions in this paper are as follows: **(1)** We create CANCEREMO ♥, a novel health-related dataset for fine-grained emotion detection composed of $8,500$ sentences. We study how emotions are distributed in our dataset and how they co-occur with each other. We further analyze emotions associations with topics such as medical procedures, side effects, and drugs, and with events or activities that happen in the past, present, and future; **(2)** We experiment on the fine-grained emotion detection task and establish strong baselines based on BERT and variants; **(3)** We study different supervised and unsupervised pre-training techniques and reveal the importance of choosing the *right* pre-training domain.

## 2 Related Work

Emotion detection has been studied in computational linguistics for a long time, with researchers exploring domains ranging from music (Strapparava et al., 2012; Mihalcea and Strapparava, 2012) and classic literature (Liu et al., 2019a) to social networks (Mohammad, 2012; Islam et al., 2019; Desai et al., 2020) and online news (Bao et al., 2009). Most studies focus on two main emotion categorizations: Ekman's (Ekman, 1992) 6 basic emotions (Katz et al., 2007; Strapparava et al., 2012; Aman and Szpakowicz, 2007; Mohammad, 2012) and Plutchik's (Plutchik, 1980) 8 emotions (Abdul-Mageed and Ungar, 2017; Mohammad and Turney, 2010). Emotion detection remains a challenging task, mainly due to the limited availability of labeled data (Abdul-Mageed and Ungar, 2017). In an effort to minimize this drawback, several studies created high quality data annotated with fine-grained emotions. For example, general Twitter data was automatically annotated with emotions using corpus-specific cues (i.e., hashtags expressing emotions) (Wang et al., 2012a; Abdul-Mageed and Ungar, 2017). Other studies turned to human annotators to manually label data (Aman and Szpakowicz, 2007; Poria et al., 2018; Liu et al., 2019a).

Interestingly, despite the importance of emotion detection in the health domain, computational studies for this task are limited. Specifically, most of these studies focus mainly on identifying two types of social support from online health communities (OHCs): emotional (Eysenbach et al., 2004) or informational (Boon et al., 2007). Along the same lines, Wang et al. (2012b) used Linear Regression to predict the degree of emotional or informational support from an OHC related to breast cancer, while Biyani et al. (2014) studied the presence of such support from breast and lung cancer data using models such as Naïve Bayes, Support Vector Machines, and Logistic Regression with

part-of-speech tags and bag-of-words. Wang et al. (2014) studied social support using lexical and sentiment features, and analyzed user engagement in OHCs. Yang et al. (2019a), on the other hand, modeled social roles in OHCs. They used a Gaussian mixture model to identify coherent roles such as *emotional support provider*, *informational support provider*, *newcomer*, or *all-round expert*. The types of features they used range from linguistic behaviors or network (i.e., relationship with other users) to features regarding the context of communication (i.e., public or private). Khanpour and Caragea (2018) highlighted the need to examine emotions from health-related posts at a finer granularity and used annotators to label two datasets with the Ekman's six basic emotion set (Ekman, 1992). The authors trained a hybrid neural model composed of a word-level Convolutional Neural Network followed by a Long Short Term Memory network. However, given the limited size of the annotated datasets (~$1,000$ sentences each) and the fact that most emotions were extremely infrequent, the analysis could only be performed on the most frequent emotions: joy and sadness. In contrast to the above works, we study Plutchick-8 basic emotions and present CANCEREMO ♥, which, to our knowledge, is the first large health dataset for the fine-grained emotion detection task, being more than eight times larger than the currently available datasets of Khanpour and Caragea (2018).

CANCEREMO ♥ enables complex explorations of deep learning models including pre-trained language models, such as BERT (Devlin et al., 2018), XLNet (Yang et al., 2019b) and RoBERTa (Liu et al., 2019b), which achieve state-of-the-art performance on several NLP tasks. We use the aforementioned pre-trained language models, fine-tune the models on our dataset, then compare these approaches with baselines from Traditional and Deep Natural Language Processing.

## 3 Dataset

### 3.1 Task Structure

**Corpus** We choose an online cancer network as the basis of our data, which we will call *CancerNet*[1] throughout the paper. CancerNet was founded in 2002 and represents a platform for people suffering from cancer as well as for their caregivers, friends, and families to socialize, share experiences and emotions, and feel supported. We collected the data from the beginning until the year of 2018. The network consists of multiple discussion boards, corresponding to different types of cancer. To create our dataset, we randomly sampled sentences from the discussion boards corresponding to three frequent types of cancer: breast, lung and prostate (BLP). We model the emotion detection task at *sentence level* since longer messages usually contain multiple topics and could possibly switch between many emotions from one sentence to another (Biyani et al., 2014).

**Objective** Given a predefined set of emotions - Plutchik-8 basic emotions, the goal is to classify a sentence with all emotions contained in it, i.e., identify all emotions conveyed in a piece of text.

### 3.2 Task Construction

**Sampling Strategy** Current datasets for emotion detection usually utilize some type of sampling bias, e.g., using emotion words as a proxy for sampling. For example, Abdul-Mageed and Ungar (2017) used cues in the data (i.e., emotion hashtags) to collect and further annotate a large Twitter dataset with emotions, while making the strong assumption that a sentence can only express one emotion. We argue that a sentence can not only express emotions even in the absence of emotion words but also convey multiple emotions, as shown in Table 1 in §1. Thus, we sample at random $25,000$ sentences from the BLP boards and annotate them using crowd-sourcing. This sampling strategy also helps us analyze how many sentences convey emotions out of all sampled sentences and how many sentences that do not contain emotion words (i.e., do not have surface lexical patterns) in fact appear to convey emotions.

**Annotation** To annotate our data, we use the Amazon Mechanical Turk (AMT) crowd-sourcing platform. The emotion definitions provided to the annotators are shown in Appendix A. We ran the annotation task in several iterations in order to develop our quality control steps. Initially, we internally annotated a batch of 100 sentences using all emotions that apply from all 8 Plutchik's emotions, in a multi-class setting. Then, we explored two settings with the AMT annotators: First, we designed a form that asked annotators to select all emotions that apply for a sentence and used the same batch of 100 sentences for analysis. We noticed that the task was very difficult and resulted in a low

[1] https://csn.cancer.org/

inter-agreement. Second, we created a separate annotation form for each emotion: for an emotion *x*, a form asks the annotators to annotate a sentence with *true* or *false*, i.e., if a sentence contains *x*, the label is *true*, otherwise it is *false*. We used again the same batch of 100 sentences for analysis. We noticed that this task was much easier and resulted in a higher inter-agreement among the AMT annotators, as well as a much higher agreement with our internal annotations. Thus, for our final annotation, we chose the latter approach over creating a single annotation form for all eight emotions, in order to leverage annotation ease and prevent any implicit associations annotators might make - one might refrain from assigning both fear and joy to the same sentence, which could in fact appear together; such an example is shown in Table 1.

We use three annotators for each sentence, and the final label for a specific emotion is computed through majority vote. We avoid spamming by ruling out the annotators that are inconsistent with the majority vote in more than 25% of the cases. We compute the inter-annotator agreement using Krippendorff Alpha, and obtain an average value of $\alpha = 0.69$ on all emotions. We also studied the per-emotion inter-agreement, and observed lower inter-annotator agreement on the emotion *anticipation*, which, in line with our beliefs, was the hardest emotion to distinguish, with $\alpha = 0.5$. Emotions such as joy, sadness, and fear produced a higher agreement, with $\alpha = 0.75$.

### 3.3 Analysis

**Emotion Distribution** Table 2 shows the number of sentences annotated with no emotions and with 1-4 emotions. Interestingly, out of the 25,000 sampled sentences, 16,500 sentences (66%) do not contain any emotions at all, and only 8,500 contain at least one emotion, out of which 16% contain two or more emotions. Figure 1 shows the distribution of our 8 emotions in the 8,500 sentences. We can notice that the distribution is very unbalanced: joy, fear and sadness appear most frequently, amounting for about 75% of the data, while anticipation, anger, surprise, disgust, and trust appear rarely, a few orders of magnitude less than the frequent ones. It is interesting to see that joy is the most prevalent, despite dealing with a cancer forum.

Table 3 shows the number of sentences annotated with no emotions (EMOSENT$^-$) and with one or more emotions (EMOSENT$^+$) and for each category the number of sentences that contain at least one emotion word from EmoLex (Mohammad and Turney, 2013). EmoLex is a word-emotion lexicon composed of a list of English emotion rich words and their associations with Plutchik's eight basic emotions. As an example, the sentence *"He is always in pain .. (chest and back pain) and has trouble swallowing pills."* contains an emotion word *pain* from EmoLex, which is associated with sadness in EmoLex. The sentence is annotated with sadness by our annotators as well. In contrast, the sentence *"I just miss him so much.....we would hold hands every night"*, does not contain any emotion word from EmoLex and is annotated with sadness by our annotators. Moreover, the sentence *"So get a second opinion and don't be afraid to change doctors."* contains the emotion rich word *afraid* from EmoLex, which is associated with fear in EmoLex, whereas the sentence conveys no emotion at all (and is annotated with no emotion by our annotators). Notably, 10% of the sentences annotated with emotions do not contain EmoLex words, while 23% of sentences with EmoLex words, do not convey any emotion.

We further use EmoLex to compare sentences

| | 0E | 1E | 2E | 3E | 4E |
|---|---|---|---|---|---|
| #SENT | 16,500 | 7,098 | 1,292 | 96 | 14 |

Table 2: Number (#) of sentences with 0 to 4 emotions.

| | EMOWORD$^+$ | EMOWORD$^-$ |
|---|---|---|
| EMOSENT$^+$ | 7,659 | 841 |
| EMOSENT$^-$ | 2,220 | 14,661 |

Table 3: Number (#) of sentences with and without emotions and emotion words. $^+$ means an emotion or emotion word is present; $^-$ means otherwise.

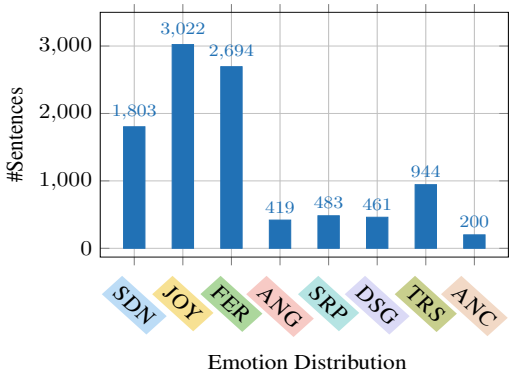


Figure 1: Emotion distribution in the dataset, including sadness (SDN), joy (JOY), fear (FER), anger (ANG), surprise (SRP), disgust (DSG), trust (TRS), and anticipation (ANC).

with and without EmoLex emotion words with respect to the difficulty to distinguish the emotions present in them. For each of the eight emotions, we separate sentences with EmoLex emotion words from those without EmoLex emotion words and calculate the AMT inter-annotator agreement. Interestingly, we find that the agreement is higher for sentences with EmoLex words only for anger, anticipation, fear, joy, and trust, and is lower on sadness, surprise, and disgust.

**Emotion co-occurrence** Since each sentence can be annotated with multiple emotions, we study what emotions tend to appear in the same context with others through a co-occurrence heatmap, shown in Figure 2. We use a logarithmic scale for a better visualization of the less frequent emotions. As expected, emotion pairs like fear-sadness or trust-joy are commonly used together. However, we observe quite a few unusual co-occurrences (of even opposing emotions) such as fear-joy or joy-sadness. For example, in the sentence *"Yesterday they told me they didnt see anything which brought tears of joy, but also a wave of fear."*, we speculate that the writer is expressing *joy* because of recent good medical analysis results, but at the same time *fear*, facing the possibility of the disease reappearing. When humans become emotional, they may indeed experience a mixture of emotions (not just one). We allow multi-labels for the same text to capture this mixture of emotions.

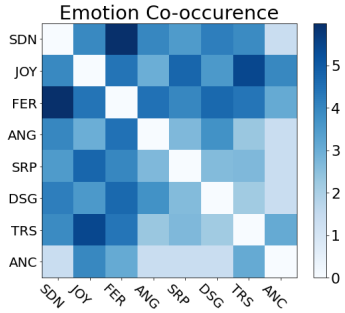


Figure 2: Emotion co-occurrence.

**Emotion Associations with Past, Present, or Future Events or Activities** We investigate whether user posts are more emotional about events or activities that happen in the past, present, or future, and how these emotions are distributed along these three dimensions. For example, in the sentence *"I just cant stand seeing her like this"*, the writer's discontent is expressed towards an event in the present, while in *"I have been through the worst fear when I started to have the pain."*, the expressed emotions are relative to an event in the past. We study this using Stanford CoreNLP Natural Language Software (Manning et al., 2014) in three steps: first, we perform a dependency parsing to extract the verb phrase in a sentence, then we take the POS tag of the verb in the verb phrase to get the sentence tense, followed by investigating the emotion conveyed in the sentence and how it relates to the identified verb tense of the sentence. Figure 3 shows the results obtained. We observe that events or activities in the present are frequently discussed across all emotions. Anticipation is, as expected, rarely discussed in the past, as well as anger and trust. Surprise, sadness, and fear on the other hand are conveyed more frequently towards past events or activities. We can also notice that emotions are associated most often with events or activities in the present.

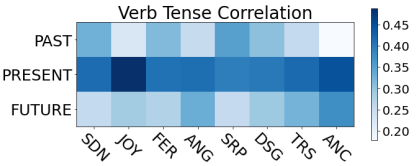


Figure 3: Emotion-Verb Tense Association. The results are normalized along the vertical axis.

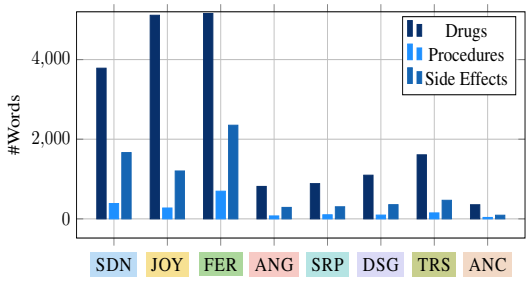


Figure 4: Emotions across Topics.

**Topics** Recognizing how patients feel about different medical topics can provide information into potential causes for the conveyed emotions. These topics are frequently discussed in OHCs and range from prescribed drugs to side effects of medication and medical procedures. We study how these medical topics relate to patient's emotions by using three medical lexicons specifically created for our cancer domain, which contain words and phrases associated with medical procedures, side effects of medication, and drugs. We collected these lexicons from online resources such as Wikipedia and WebMD.[2] These medical topics are extremely important from

[2] https://www.webmd.com/

a practical point of view, as can provide insight into how patients react to their medication, or what side-effects they may be experiencing. We match words from the three lexicons to our dataset, then study how emotions correlate with these topics. We report our findings in Figure 4. As we can see from the figure, interestingly, the topic on *Drugs* is discussed most frequently (across all emotions), while the topics on *Side Effects* and *Medical Procedures* appear more often in sentences conveying fear or sadness as compared to joy.

**Benchmark Dataset** To enable development on the fine-grained emotion detection task in health related posts, we construct a benchmark dataset. We group the positive examples (sentences conveying one or more emotions) into eight pools - one for each emotion; a sentence is part of a pool if the sentence is annotated with the respective emotion. We remind that a sentence can convey more than one emotion, so it can be part of two different pools at the same time. Next, we sample an equal amount of negative examples for each pool using the following strategy: $\frac{1}{3}$ are sampled from the sentences that convey no emotions, while the other $\frac{2}{3}$ are sampled from all the positive examples from the other pools. We followed this strategy in order to create a challenging negative set for each emotion. We sample an equal number of positives and negatives because of the imbalanced emotion distribution, which would lead to an extremely skewed ratio of positive to negative samples. Next, we randomly create an 80/10/10 split to create the train, validation and test split. We present specific details about each split in Appendix B.

To facilitate future research, we make our code available[3] along with all other resources of this project (for research purposes).

## 4 Baseline Modeling

We model the Plutchik-8 basic set of emotions in CancerEmo ♥ using the following methods:

**Statistical and Machine Learning Methods** We experiment with **(1) EmoLex** - a simple annotation scheme based on EmoLex words' emotions: we label a sentence with the union of the emotion labels of the EmoLex words (Mohammad and Turney, 2013) contained in the sentence, or no emotion if no EmoLex words appear in the sentence. **(2) Naïve Bayes** using a tf*idf weighting scheme, computed after stemming and stop-word removal; and **(3) Logistic Regression** using averaged pre-trained FastText (Bojanowski et al., 2017) word embeddings.

**Standard Neural Methods** We experiment with **(1) Bi-LSTM** (Hochreiter and Schmidhuber, 1997) **(2) CNN** (Kim, 2014) and **(3) Conv-Bi-LSTM**, a mix of the two used in prior work on the fine-grained emotion detection task (Khanpour and Caragea, 2018).

**Pre-Trained Language Models** Recently, pre-trained language models have risen in popularity, because they use transfer learning, the process of storing information learned from a task and applying it to another task. The process usually involves unsupervised pre-training on a large corpus, followed by a less computationally expensive fine-tuning, performed on the task at hand. We experiment with three models: **(1) BERT** (Devlin et al., 2018) **(2) RoBERTa** (Liu et al., 2019b), a variant of BERT, which underwent significantly more pre-training, and **(3) XLNet** (Yang et al., 2019b), which has a different language modeling objective than BERT called *Permutation Language Modeling*.

## 5 Experiments and Results

In this section, we present the set of experiments performed on the fine-grained emotion detection task on CancerEmo ♥, as well as show the results obtained using the aforementioned baselines.

**Experimental Setting** All the traditional neural network models were tested with pre-trained FastText (Bojanowski et al., 2017) word embeddings. The LSTM-based models have 300 hidden units and a dropout rate of 0.5. For the CNN, we follow the best hyper-parameters presented by Kim (2014). For the pre-trained language models, we start from the best reported hyper-parameters and perform a bi-directional linear sweep. More details on the fine-tuning techniques and the hyper-parameter values used for the best models can be found in Appendix C. The reported results represent the average of five independent runs. All experiments were carried out on an NVIDIA V100 GPU.

**Results** Table 4 shows the results in terms of F1-score, obtained using BERT-like models compared with the other weaker baselines. We can observe that EmoLex performs very poorly, reinforcing our premise that lexical level information in the form of

[3] https://github.com/tsosea2/CancerEmo.git

| METHOD | SDN | JOY | FER | ANG | SRP | DSG | TRS | ANC | AVERAGE |
|---|---|---|---|---|---|---|---|---|---|
| EMOLEX | 0.47 | 0.64 | 0.50 | 0.35 | 0.16 | 0.22 | 0.50 | 0.53 | 0.42 |
| LOGISTIC REGRESSION | 0.60 | 0.73 | 0.66 | 0.61 | 0.63 | 0.45 | 0.57 | 0.55 | 0.60 |
| NAÏVE BAYES | 0.63 | 0.71 | 0.67 | 0.60 | 0.62 | 0.54 | 0.56 | 0.60 | 0.61 |
| BI-LSTM | 0.64 | 0.74 | 0.64 | 0.67 | 0.50 | 0.57 | 0.59 | 0.53 | 0.61 |
| CNN | 0.63 | 0.73 | 0.59 | 0.58 | 0.55 | 0.59 | 0.66 | 0.54 | 0.61 |
| CONV-BI-LSTM | 0.64 | 0.73 | 0.66 | 0.65 | 0.67 | 0.54 | 0.63 | 0.72 | 0.66 |
| BERT | 0.71 | 0.81 | 0.77 | 0.68 | 0.68 | 0.59 | 0.67 | 0.70 | 0.71 |
| XLNET | 0.71 | 0.83 | 0.77 | 0.64 | 0.56 | 0.52 | 0.65 | 0.70 | 0.67 |
| ROBERTA | 0.65 | 0.83 | 0.72 | 0.65 | 0.57 | 0.54 | 0.57 | 0.78 | 0.67 |

Table 4: Binary Task F1-score on CANCEREMO ♥.

| | SDN | JOY | FER | ANG | SRP | DSG | TRS | ANC | AVERAGE |
|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.71 | 0.81 | 0.77 | 0.68 | 0.68 | 0.59 | 0.67 | 0.70 | 0.71 |
| EMONET | 0.68 | 0.78 | 0.77 | 0.64 | 0.54 | 0.54 | 0.61 | 0.67 | 0.65 |
| CNET | 0.72 | 0.83 | 0.77 | 0.66 | .68 | 0.57 | 0.66 | 0.75 | 0.71 |
| FLTR CNET | 0.74 | 0.84 | 0.79 | 0.68 | 0.69 | 0.59 | 0.67 | 0.76 | 0.72 |
| CLINICAL | 0.74 | 0.81 | 0.79 | 0.68 | 0.67 | 0.59 | 0.68 | 0.75 | 0.72 |
| CLINICAL FLTR CNET | 0.76 | 0.84 | 0.80 | 0.68 | 0.68 | 0.58 | 0.68 | 0.75 | 0.72 |
| EMONET | 0.73 | 0.83 | 0.77 | 0.67 | 0.68 | 0.56 | 0.66 | 0.73 | 0.71 |

Table 5: Intermediate task pre-training F1-score results. In order from top to bottom: **(1)** BERT, which corresponds to BERT with no intermediate pre-training (top) **(2)** Unsupervised Pre-training (middle block) **(3)** Supervised Pre-training (bottom block). An improvement over the BERT model is marked with PURPLE, while a decrease in performance is signaled using RED. The best performing model F1s are underlined.

emotion words does not necessarily reveal the emotion conveyed. Interestingly, the Conv-Bi-LSTM model manages to improve upon the other statistical and standard neural network methods by as much as 5%. The BERT base model is extremely successful across all emotions, greatly outperforming all the other baselines by 4% F1 on average.

Next, we explore intermediate task pre-training to understand if this improves the performance of our BERT models further (Pruksachatkun et al., 2020; Han and Eisenstein, 2019).

## 6 Intermediate Pre-Training

CANCEREMO ♥ is created from a health forum, i.e., a network of cancer survivors that we call CancerNet (or CNet for short). Thus, our data differs substantially from the pre-training domain of BERT (Devlin et al., 2018) (Wikipedia and Bookcorpus). As Xia and Ding (2019) noted, domain-adaptive fine-tuning (i.e., adapting the contextualized embeddings to the target domain) might implicitly incorporate inductive biases and improve the performance of the models. To investigate this, we perform an additional set of comprehensive experiments with the best performing model from the previous experiment: BERT. The experimental pipeline consists of two steps: starting from a pre-trained BERT model, we **(1)** perform an unsupervised or supervised pre-training on an *intermediate pre-training task*, followed by **(2)** fine-tuning on the *target task*, which is always the fine-grained emotion detection on CANCEREMO ♥.

**Intermediate Tasks** The unsupervised pre-training is performed using the Masked Language Modeling objective, while the supervised pre-training is carried out by adding a linear layer, followed by fine-tuning on the emotion detection task. The intermediate tasks are as follows: **(1) Unsupervised EmoNet** EmoNet (Abdul-Mageed and Ungar, 2017) is a Twitter dataset composed of tweets automatically annotated using distant supervision with Plutchik-24 emotion set. We obtained a smaller version of the dataset from the authors which contains the Plutchik-8 basic emotions. We pre-train the BERT model on all EmoNet sentences. **(2) Unsupervised CNet** We pre-train the BERT model on all CancerNet sentences, hoping to implicitly learn information specific to the health domain. **(3) Unsupervised Filtered CNet** We use lexical features to filter CancerNet. To this end, we implicitly induce both health and emotion specific biases, by only pre-training on CancerNet sen-

| | SDN | JOY | FER | ANG | SRP | DSG | TRS | ANC | AVERAGE |
|---|---|---|---|---|---|---|---|---|---|
| EMOWORDS$^+$ | 0.77 | 0.85 | 0.80 | 0.68 | 0.70 | 0.67 | 0.69 | 0.76 | 0.73 |
| EMOWORDS$^-$ | 0.63 | 0.74 | 0.75 | 0.68 | 0.67 | 0.50 | 0.65 | 0.67 | 0.66 |
| EMOWORDS$^+$FLTR CNET | 0.79 | 0.88 | 0.82 | 0.68 | 0.70 | 0.68 | 0.69 | 0.78 | 0.75 |
| EMOWORDS$^-$ FLTR CNET | 0.64 | 0.76 | 0.76 | 0.68 | 0.67 | 0.50 | 0.65 | 0.67 | 0.66 |

Table 6: F1 performance of BERT (top two lines) and FLTR CNET BERT (last two lines) F1 performance on sentences **with** (EMOWORDS$^+$) and **without** (EMOWORDS$^-$) emotion words.

tences that contain at least one emotion word from EmoLex (Mohammad and Turney, 2013). **(4) Unsupervised Clinical** We observe that sentences in our data contain terms from medical specialty vocabulary. We present some examples of this phenomenon: *"I too had Adenocarcinoma in the very top of my left lung in 1987, they removed the top half of the lung."*; *"My wife, 68, was recently diagnosed with Stage IV lung cancer with K-RAS mutation."* These examples illustrate some uses of medical terminology in the forum. We explore whether pre-training on a medical speciality corpus improves the performance of the models. To this end, we use the publicly available Clinical BERT (Alsentzer et al., 2019) medical specific contextual embeddings. **(5) Unsupervised Clinical Filtered CNet** We investigate if additional emotion guided pre-training helps Clinical BERT. Following **Unsupervised Filtered CNet** method, we pre-train Clinical BERT on CancerNet sentences that contain at least an emotion word. **(6) Supervised EmoNet** For the supervised setting, we pre-train on a multi-class emotion classification task on EmoNet (Abdul-Mageed and Ungar, 2017). We use a linear layer to perform the fine-grained emotion classification task on EmoNet, and after achieving an F1 of $0.83\%$, we drop this layer. Next, the target fine-tuning on CANCEREMO ♥ is performed using a freshly initialized linear layer.

**Results** The results in terms of F1-score obtained are compared with the BERT models in Table 5. In the unsupervised setting, we observe a few patterns. First, unsupervised pre-training on EmoNet (Abdul-Mageed and Ungar, 2017) largely hurts downstream performance. Second, approaches inducing health specific biases from **CNet** and **Clinical** perform better than BERT on sadness, joy and anticipation. Third, **Clinical Filtered CNet** consistently outperforms all the other models by as much as $5\%$ on sadness, joy, fear and anticipation, while keeping the same overall F1-score on the other $4$ emotions. We speculate that this happens because the pre-training corpus used is very close to the task domain, and we manage to implicitly induce both emotion-specific and health-specific biases. Last, interestingly, the supervised intermediate task pre-training on EmoNet improves the performance on emotions like sadness, joy, and anticipation, but performs similarly or degrades the performance on the other emotions. Still, the **Supervised EmoNet** performs much better compared with the **Unsupervised EmoNet**.

**Takeaways** One should pay close attention when dealing with very narrow domains like emotion or health, where the pre-training corpus greatly influences the performance of the models, and the *right* pre-training can improve the performance.

## 7 Emotion Word Testing

A good amount of sentences annotated with emotions by our annotators in CANCEREMO ♥ do not contain any emotion words from EmoLex (§3.3). Thus, we now investigate if the absence of emotion words affects the model performance. To this end, to depict a real scenario, we keep the train set unchanged and divide the test set in two: one set contains only sentences that have at least an emotion word, while the other contains only sentences without emotion words. As Table 6 shows, testing on sentences with emotion words provides a considerable $8\%$ average F1 increase over sentences with no emotion words. Next, we perform the same experiment using the **Unsupervised Filtered CNet** method. Surprisingly, the performance improves on both test sets (with and without emotion words) on several emotions, e.g., sadness, joy and fear.

## 8 Significance Test and Error Analysis

We investigate if our results are statistically significant. To this end, we perform paired t-tests to test significant differences between model results. We reject the null hypothesis if $p < 0.05$. Our significance tests show the following: First, the improvement of BERT over the statistical and standard neural baselines is statistically significant on

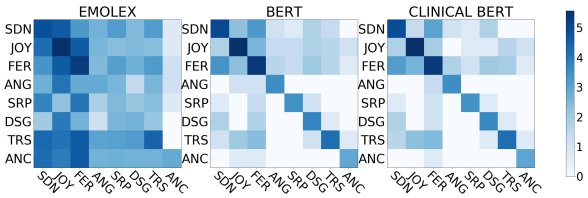

Figure 5: Predicted Lables (vertical axis) vs Actual Lables (horizontal axis).

all emotions. Second, the improvement of our best performing model (**Clinical Filtered CNet**) over the BERT model with no additional pre-training is statistically significant on sadness, joy and anticipation, but not on fear.

Next, using our best **Clinical Filtered CNet** BERT model, we manually investigate test errors to understand potential drawbacks of the model. We observe the following: First, the model often performs poorly on sentences with abbreviations or writing errors. For example, in the sentence *"As i will have alot of time, cuz i cant really sleep any significant amount of sleep."*, although the expressed emotion is sadness, the model assigns no emotion to it. Next, some errors arise from antithetic emotions in the same sentence. For example, the model assigns sadness to the following sentence: *"Still get tired but it's better every day."* Although the first part of the sentence could convey sadness, the overall emotion expressed is joy.

Finally, we construct confusion matrices to visualize commonly mislabeled classes, shown in Figure 5. We use a logarithmic scale to be able to better picture less frequent classes such as surprise, disgust, trust and anticipation. The EmoLex (Mohammad and Turney, 2013) visualization shows the poor performance of the lexicon approach, and reflects the results reported in Table 4. Next, we investigate commonly mislabeled classes by BERT and Clinical BERT, and observe a few patterns. For example, the most common mislabeling for the fear emotion is sadness and vice-versa, while quite a few sentences conveying disgust are annotated with sadness and fear.

## 9 Conclusion and Future Work

We introduced CANCEREMO ♥, a cancer-related health dataset for perceived emotion detection, which is an order of magnitude larger and more fine-grained compared with previous datasets for health-related emotion detection. Composed of $8,500$ sentences that convey at least one emotion, and $16,500$ sentences that convey no emotion at all, CANCEREMO ♥ is a challenging benchmark for fine-grained emotion detection, as shown by our results. We believe that CANCEREMO ♥ is novel and has unique characteristics: 1) covers a large spectrum of emotions - being annotated with the Plutchik-8 fine-grained emotions; 2) has a large dataset size for exploring deep learning models; and 3) provides an invaluable context - cancer - for dealing with emotions. The value of our dataset arises also from: the expressions of emotions even in the absence of emotion words and the expressions of mixtures of (sometimes opposing) emotions in the same text. We believe that these characteristics add interestingness and challenges to our dataset and we hope that our work will spur future research in emotion detection from health data, especially in the context of life-threatening diseases such as cancer. Our dataset, which is anonymized and follows ethical considerations, can be used as a benchmark for both multi-class and multi-label emotion detection.

In the future, we plan to study how contextual information (i.e., different aspects of people's interactions captured through contiguous posts in a discussion thread) affects the perceived emotions. We also plan to perform a cross-corpus analysis to investigate if emotions are expressed differently in the health domain compared to other domains. Finally, we will carry out a thorough investigation into emotion-cause pairs (Xia and Ding, 2019). Specifically, in the health domain, the cause that leads to an emotion expressed in text can be just as important as the emotion itself. A deeper understanding of emotion causes can potentially help make people feel better.

## References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, pages 699–704.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 827–836.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Heather S Boon, Folashade Olatunde, and Suzanna M Zick. 2007. Trends in complementary/alternative medicine use by breast cancer survivors: comparing survey data from 1998 and 2005. *BMC women's health*, 7(1):4.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.

Gabriel P Haas, Nicolas Delongchamps, Otis W Brawley, Ching Y Wang, and Gustavo de la Roza. 2008. The worldwide epidemiology of prostate cancer: perspectives from autopsy studies. *The Canadian journal of urology*, 15(1):3866.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4229–4239.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jumayel Islam, Robert E Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1355–1365.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Per E Kummervold, Deede Gammon, Svein Bergvik, Jan-Are K Johnsen, Toralf Hasvold, and Jan H Rosenvinge. 2002. Social support in a wired world: use of online mental health forums in norway. *Nordic journal of psychiatry*, 56(1):59–65.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019a. Dens: A dataset for multi-class emotion analysis. *arXiv preprint arXiv:1910.11769*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Klim McPherson, CaMa Steel, and JM Dixon. 2000. Breast cancer—epidemiology, risk factors, and genetics. *Bmj*, 321(7261):624–628.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2010. 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. Association for Computational Linguistics.

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

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Kathryn I Pollak, Robert M Arnold, Amy S Jeffreys, Stewart C Alexander, Maren K Olsen, Amy P Abernethy, Celette Sugg Skinner, Keri L Rodriguez, and James A Tulsky. 2007. Oncologist communication about emotion during visits with patients with advanced cancer. *Journal of Clinical Oncology*, 25(36):5748–5752.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E Greer, and Kenneth Portier. 2011. Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 274–281. IEEE.

Carlo Strapparava, Rada Mihalcea, and Alberto Battocchi. 2012. A parallel corpus of music and lyrics annotated with emotions. In *LREC*, pages 2343–2346. Citeseer.

Lindsey A Torre, Rebecca L Siegel, and Ahmedin Jemal. 2016. Lung cancer statistics. In *Lung cancer and personalized medicine*, pages 1–19. Springer.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012a. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Xi Wang, Kang Zhao, and Nick Street. 2014. Social support and user engagement in online health communities. In *International Conference on Smart Health*, pages 97–110. Springer.

Yi-Chia Wang, Robert Kraut, and John M Levine. 2012b. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 833–842. ACM.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: a new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019a. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

| | |
|---|---|
| SADNESS | The condition or quality of being sad. |
| JOY | A feeling of great pleasure and happiness. |
| FEAR | An unpleasant emotion caused by the belief that someone or something is dangerous, likely to cause pain, or a threat |
| ANGER | A strong feeling of annoyance, displeasure, or hostility. |
| SURPRISE | An unexpected or astonishing event, fact, or thing. |
| DISGUST | A feeling of revulsion or strong disapproval aroused by something unpleasant or offensive. |
| TRUST | Firm belief in the reliability, truth, ability, or strength of someone or something. |
| ANTICIPATION | The action of anticipating something; expectation or prediction. Similarly, anticipation is a feeling of excitement about something pleasant or exciting that you know is going to happen. |

Table 7: Emotion Definitions given to the annotators.

## A Emotions Definition

Table 7 shows the emotion definitions provided in the task instructions, which annotators have to read before starting to label the data.

## B Split Details

We present the emotion counts in every train/val/test split through Table 8. We color the emotion counts of the split in question. For instance, the first train/val/test line corresponds to the sadness split, as the column corresponding to sadness is colored.

## C Hyperparameters

We present the hyperparameters obtained by tuning in Table 9 and 10. The highest variance in the results is obtained by varying the learning rate, which we tune the most. For each emotion, we start from an initial value of 5e-05, then search for 5 iterations forward and backwards in steps of 1e-05. This type of tuning is performed for each emotion, and took in total 2 days on our V100 GPU. We use a batch size of 64 for the traditional baselines, while only 16 for BERT and RoBERTA and 8 for XLNet due to GPU ram restrictions.

| | SDN | JOY | FER | ANG | SRP | DSG | TRS | ANC | NOEMO | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|
| TRAIN | 1427 | 466 | 663 | 90 | 88 | 108 | 160 | 48 | 472 | 3522 |
| VAL | 196 | 47 | 93 | 15 | 14 | 11 | 14 | 3 | 62 | 455 |
| TEST | 180 | 63 | 79 | 14 | 10 | 6 | 15 | 1 | 67 | 435 |
| TRAIN | 573 | 2410 | 802 | 119 | 199 | 153 | 410 | 76 | 817 | 5559 |
| VAL | 81 | 311 | 102 | 28 | 20 | 18 | 47 | 15 | 82 | 704 |
| TEST | 59 | 301 | 114 | 14 | 33 | 20 | 49 | 6 | 108 | 704 |
| TRAIN | 682 | 758 | 2148 | 146 | 163 | 172 | 290 | 66 | 716 | 5141 |
| VAL | 80 | 94 | 260 | 22 | 17 | 21 | 40 | 12 | 103 | 649 |
| TEST | 75 | 101 | 286 | 23 | 10 | 18 | 40 | 4 | 79 | 636 |
| TRAIN | 92 | 91 | 145 | 344 | 24 | 48 | 41 | 9 | 105 | 899 |
| VAL | 11 | 11 | 22 | 41 | 4 | 3 | 6 | 0 | 15 | 133 |
| TEST | 14 | 11 | 15 | 34 | 6 | 7 | 3 | 0 | 19 | 109 |
| TRAIN | 77 | 187 | 139 | 20 | 377 | 27 | 36 | 13 | 138 | 1014 |
| VAL | 9 | 19 | 23 | 4 | 51 | 5 | 2 | 3 | 11 | 127 |
| TEST | 10 | 30 | 11 | 4 | 55 | 2 | 11 | 4 | 12 | 139 |
| TRAIN | 101 | 109 | 165 | 46 | 25 | 383 | 40 | 6 | 117 | 992 |
| VAL | 18 | 13 | 18 | 5 | 3 | 45 | 2 | 2 | 18 | 124 |
| TEST | 11 | 23 | 18 | 3 | 5 | 33 | 7 | 1 | 18 | 119 |
| TRAIN | 155 | 347 | 250 | 39 | 44 | 36 | 756 | 32 | 255 | 1914 |
| VAL | 15 | 50 | 29 | 3 | 9 | 3 | 95 | 5 | 29 | 238 |
| TEST | 24 | 46 | 33 | 4 | 5 | 5 | 93 | 4 | 30 | 244 |
| TRAIN | 23 | 81 | 47 | 8 | 16 | 12 | 31 | 162 | 55 | 435 |
| VAL | 5 | 14 | 8 | 0 | 0 | 1 | 1 | 17 | 6 | 52 |
| TEST | 3 | 8 | 6 | 1 | 0 | 1 | 6 | 21 | 5 | 51 |

Table 8: Emotion counts in each split.

| | LOGREG | BI-LSTM | CNN | CONV-BI-LSTM | BERT | ROBERTA | XLNET |
|---|---|---|---|---|---|---|---|
| EPOCHS | 8 | 10 | 10 | 10 | 4 | 4 | 4 |
| BATCH SIZE | 64 | 64 | 64 | 64 | 16 | 16 | 8 |

Table 9: Epochs and batch size used for the models.

| | SDN | JOY | FER | ANG | SRP | DSG | TRS | ANC |
|---|---|---|---|---|---|---|---|---|
| BERT | 4e-05 | 5e-05 | 5e-05 | 3e-05 | 5e-05 | 3e-05 | 5e-05 | 7e-05 |

Table 10: Best Learning Rates for BERT

.