Detection of self-labeled emotions from social media texts

Malik, Lukas

Award date:
2021

Awarding institution:
Graz University of Technology

Detection of self-labeled emotions from social media texts

Lukas Malik[1]

Abstract

Inferring the emotional state of an author from a written document has been a long-standing task in Natural Language Processing. Most datasets rely on annotations provided by readers. However, annotations provided by readers are problematic as emotions are a subjective experience. Instead, this work uses a self-labeled dataset provided by the social media platform Vent to train and compare state of the art natural language models. I conduct a comparison on three test datasets - a random sample, a sample with new users and a dataset from a later period of time. The best performing classifier is a roBERTa model, that I provide as a Huggingface model[1]. The classifiers are compared to human performance on a small sample using Amazon Mechanical Turk workers. Humans perform significantly worse than all classifiers, questioning the validity of annotation by human readers. I perform additional validation on the ISEAR and the EmoInt dataset. Performance overall is worse than that of a BERT model trained on the ISEAR dataset but outperforms the ISEAR trained model on the emotion *Anger*. On the EmoInt dataset used in the SemEval competition, performance is comparable to Median Team performance but worse than the specialized architecture used by the winning team. I discuss possible reasons for the results as well as consequences for the field of textual emotion recognition.

1. Background

Emotions are a central component of human communication and the main drivers of behaviour (Hancock et al., 2007). Despite their importance, understanding what an emotion is or what even counts as an emotion has proven difficult. Many psychologists have tried to determine a fixed set of emotions - often called basic emotions, that are independent of culture, time and the individual and therefore offer universality (Ekman, 1999). However, the universality of most if not all of the emotion frameworks has been called into question (Ortony & Turner, 1990) and what can and can not be considered a basic emotion is therefore still up for debate. In addition to the problem of constructing frameworks to understand emotion comes the fact that emotions are a subjective experience. While other people can pick up on cues indicating the presence or absence of an emotion, these cues can be missed or misunderstood, often leading to miscommunication.

Humans express emotions in a variety of mediums. In recent years there has been a steady shift in communication towards online communication, which mostly occurs through text messages (Scott et al., 2017). As more and more interactions occur in the digital space, online texts offer an additional lens through which to observe social phenomena. Understanding emotions from texts could offer important insights into social phenomena (Garcia & Rimé, 2019) and might reveal possibilities for intervention to counter phenomena like political polarization (Schweitzer et al., 2020; Garcia et al., 2015) or hate speech (Zhang & Luo, 2018).

Emotion detection from texts is a form of sentiment analysis, where the objective is to detect several emotions in the text. The number of emotions should be more than two because a set of only two principal emotions would revolve around valence. Because there are limits in expressing emotions through texts, the emotions detected by the reader might not be the same as the emotion attributed by the author. Text inherent factors that can influence the difficulty of emotion detection from the text are the valence of the underlying emotion, the medium of the text, language proficiency of the author and last but not least, how good the person is at expressing their emotions. In addition to text inherent factors, the accuracy of the reader might also be influenced by individual factors like stress and anxiety (MacCann & Roberts, 2008).

Few datasets have been curated, to tackle emotion classification tasks. The most prominent was featured in SemEval - a yearly ongoing series of natural language processing challenges that aims to push the state of the art in semantic analysis and help create high quality annotated datasets. Sentiment analysis tasks were featured as early as 2016. Here the task was to assign positive or negative sentiment and not fine-grained emotion labels (Rosenthal et al., 2017; Nakov et al., 2019). However, sentiment detection as de-

[1] https://huggingface.co/lumalik/vent-roberta-emotion

tecting positive or negative valence is considered an easier task than emotion detection (Mohammad et al., 2018). The first dataset that concerned emotion detection to my knowledge was the LiveJournal dataset that was collected in 2009 (Keshtkar & Inkpen, 2009). The dataset contains over 815 thousand blog entries, self-annotated with one of 132 moods. The first dataset in SemEval that concerned emotion detection was the EmoInt dataset (Mohammad et al., 2018). The dataset comprises randomly chosen tweets (Mohammad & Bravo-Marquez, 2017; Mohammad et al., 2018) written between 2016 and 2017. The tweets were then filtered according to four basic emotions - *anger*, *fear*, *joy* and *sadness* and for three languages English, Arabic and Spanish. There is also a multi-label version of the annotations, annotating each tweet with the presence or absence of one or more of eleven non-exclusive emotions. Crowdsource workers performed all annotations, and seven workers annotated each tweet. Each tweet was labeled with an intensity score for one of the emotions. This intensity score was calculated by asking crowdworkers to order tweets by intensity concerning one of the four emotions.

Another related task introduced in the year 2018 is emoji detection (Barbieri et al., 2018). Emoji detection can be regarded as a particular form of emotion detection because emojis are attributed by the author but are also part of the text message itself. Texts with emojis are also unique, as emojis are mainly used under specific circumstances - namely, when communication is casual and the valence of the emotion is high. Furthermore, emojis tend to be used more frequently in positive situations (Bai et al., 2019). For example, very few people would use emojis in a business context (Ćorić et al., 2018) or when writing a neutral document. Furthermore, because emojis are also part of the text, they can also change the meaning of a text.

At this point, it is important to note that all of the previously mentioned data sets and competitions with the exception of the LiveJournal dataset only consider emotion labels coming from external annotators. The labels are not self-labeled by the author. Emoji prediction datasets form an exception under the limitations provided in the previous paragraph. We summarize that data, where authors annotate their texts with emotion labels, are mostly a blind spot for research in emotion detection.

The data used to train the emotion classifier in this work comes from the social media platform Vent (Lykousas et al., 2019). Vent is a platform that requires users to tag their posts with an emotional label. Vent is built on the principles of self-expression, support, life enrichment, fun and privacy (Lykousas et al., 2019). The data set counts a total of over 33 million posts coming from over 900 thousand users collected over a period from October 2013 to October 2018 (Lykousas et al., 2019). Because users need to tag their posts, this dataset allows researchers to train a classifier to detect self-labeled emotions. There are a total of 63 emotions. Most do not fall into established basic emotions and instead are what are called seasonal emotions, indicating some event. For a detailed analysis of the emotion labels, I refer to section 2.

In addition to labeled data, emotion detection from texts also requires efficient natural language processing algorithms. Early work in emotion detection used weighted term frequency counts or word concurrences to represent texts (Sundaram et al., 2021; Winarsih et al., 2016; Mohsen et al., 2016). It has been demonstrated that Support Vector Machines are very effective with these frequency count features (Mohsen et al., 2016). One of the most widely used frequency word count techniques is term frequency inverse document frequency (TF-IDF). Here the frequency of the word is weighted inversely by how often the word occurs in documents (Ramos et al., 2003). The idea here is that words that appear in every document are less important than words that appear in only a few documents. However, frequency count methods come with some drawbacks. First, they do not incorporate a concept of word similarity (Qaiser & Ali, 2018). Second, they operate directly in word count space and might be very slow for large vocabularies (Qaiser & Ali, 2018). Third, they assume that word counts provide evidence of similarity (Qaiser & Ali, 2018). Last but not least, another problem of frequency count is that any word order is lost in this representation.

Because of these shortcomings, other forms of text representations have been developed. Another, more elaborate representation comes in the form of word embeddings, where words map to a multidimensional vector such that similarities between words correspond to the similarity in space (Mikolov et al., 2013). These embeddings derive from a single layer neural network that tries to predict the next word, given the previous word (Mikolov et al., 2013). In contrast to TF-IDF where the whole text maps to a vector, word embeddings map individual words to vectors. Bojanowski et al. (2016); Joulin et al. (2016) enrich these word embeddings by using character or subword information to enhance classification. To obtain a representation for a document, we can calculate the average of the word vectors in the document (Bojanowski et al., 2016; Joulin et al., 2016). One drawback of this approach is that the resulting representation no longer considers word order. However, Bojanowski et al. (2016); Joulin et al. (2016) demonstrated that text embeddings derived from the simple average word embeddings in a document could beat more sophisticated methods. More sophisticated methods include Long Short Term Memory models (LSTM) or Convolutional Neural Networks trained on individual word embeddings (Bojanowski et al., 2016; Joulin et al., 2016).

In the last few years, there has been a shift towards trans-

former models (Vaswani et al., 2017) for natural language processing to directly classify texts (Tenney et al., 2019; González-Carvajal & Garrido-Merchán, 2020). The most frequently used architecture is that of a bidirectional encoder representation from transformers, also known as BERT (Devlin et al., 2019). BERT builds on the concepts of masked language modelling, next sentence prediction and the usage of multi-head self-attention and subword tokenization. BERT was trained on two large corpuses - the Wikipedia corpus, with 2500 million words and the BookCorpus, with 800 million words. The model is supposed to be fine-tuned on specific tasks (Devlin et al., 2019).

Masked language modelling is a technique where individuals words in texts are masked from the network, and the network aims to correctly determine the masked word using all of the non-masked words in the sentence. This way, BERT ensures that the context of the word is sufficiently considered. This approach is bidirectional, meaning that words previous to the masked word and the words after the masked word are taken into consideration by the network. In addition to masked language modelling, BERT is also trained using next sentence prediction. In a next sentence prediction task, given two sentences, BERT has to determine whether the two sentences follow each other or whether two random sentences were concatenated. This approach is useful to enhance performance in question answering tasks. Self-Attention is a method that for each word encourages the network to look for other words that relate to that word (Alammar, 2018). Multi-Head self-attention creates multiple self-attention representations, with each representation first randomly initialized and then learned during training. Multi-Head self-attention allows the network to learn from multiple representations and enhances the models' ability to focus on different positions (Alammar, 2018). Last but not least, subword tokenization describes a method that is used when the network does not know a word. In this case, the word is divided into subwords known by the network. Subword tokenization is especially useful when it comes to grammar, as slight deviations would otherwise create unknown words (Devlin et al., 2019).

Since the first inception of the BERT architecture, much research has focused on adjusting the architecture to the various needs of natural language processing researchers. These adjustments included, increasing the size and the number of parameters (Lan et al., 2020), decreasing the number of parameters (Sanh et al., 2019; Sun et al., 2020) or using a more flexible tokenizer and improved pre-training regime (Liu et al., 2019). Especially the vocabulary present in the BERT tokenizer has proven to be limited in social media contexts (Delobelle & Berendt, 2019). One limitation is that emojis are not included as part of the tokenizer and should be added manually to improve performance on social media data (Delobelle & Berendt, 2019). A recent trend in tokenization has been to move from word piece tokenizers to byte-pair-tokenizers as they offer more flexibility (Liu et al., 2019). I aim to compare the previously mentioned classification algorithms and test their efficacy in detecting self-labeled emotions in the Vent social media dataset.

The main contributions of this work to the literature on emotion detection are:

- I train and compare popular models on emotion classification from text data, self-labeled by the authors of the text on a social media platform. I focus on five emotions namely *Anger*, *Happiness*, *Sadness*, *Fear* and *Affection*. I perform tests on three test sets: a random test set, a user set with users not contained in the train set, and a time test set with posts published after the train set. The best model is accessible via the Huggingface model hub.
- I compare the performance of the classifier to human performance. Human Performance is assessed with a random set of texts annotated by Amazon Mechanical Turk Workers. For more information on the sample, I refer to section 2.
- Additionally, I test the performance of this classifier on two unseen datasets. First, on the ISEAR dataset (Dan-Glauser & Scherer, 2012). Second, on the EmoInt dataset used on task 1 of the emotion detection challenge from SemEval 2018 (Mohammad et al., 2018). The ISEAR dataset comes from a different time epoch and describes emotional situations (Dan-Glauser & Scherer, 2012). The tweets from the EmoInt dataset are shorter due to the character restrictions and encourage the usage of hashtags. I chose these two datasets to demonstrate the robustness of the classifier.

2. Material and Methods

The Vent dataset contains a total of 33,623,415 posts and emotion labels coming from 934,095 users. English is the most frequently used language. I assessed language frequency by using the Python langdetect library (Dan-Glauser & Scherer, 2012) on a random sample of 10,000 posts. The distribution is depicted in figure 4 in the appendix. The basic emotion labels in Vent are *Feelings*, *Surprise*, *Happiness*, *Creativity*, *Sadness*, *Fear*, *Affection*, *Anger* and *Positivity*. Additionally, there are also 53 seasonal emotions that can be selected for a limited time like *LGBT+ Pride Ramadan* and many more.

Nikolas Hammerl conducted most of the preprocessing of the data. This included filtering the dataset for the basic emotions *Anger*, *Affection*, *Happiness*, *Sadness*, *Surprise* and *Fear* and splitting the data in train and test sets. The emotions were selected on the basis of Ekman's basic emotions

Anger, *Joy*, *Disgust*, *Sadness*, *Surprise* and *Fear* (Ekman, 1999). *Joy* is replaced by *Affection* and *Happiness*. *Disgust* is not available as a label. After filtering the dataset for these basic emotions, the data comprises 11,231,491 posts with their respective labels. The data was split into one train and three test sets - a random set, a user set with users not present in the train set and a time test set consisting of posts published after the train set. The user and time test set were composed in such a way that they each contain roughly 10 percent of the data. The train set comprises 7,945,618 posts; the random test set 1,093,109 posts, the user set 1,102,291 posts, and the time set 1,090,473 posts.

My team and I established a baseline for human performance using the members of our research group, see table 6 in the appendix. For the baseline, 60 samples were randomly drawn and annotated by six coworkers. Following this baseline experiment, we decided to exclude *Surprise* as there was little agreement between the annotators and the labels of the authors. Also, *Surprise* does not contain positive or negative valence and is therefore hard to categorize. Furthermore, "Surprise" was the least frequent label with only 244,753 samples. After removing *Surprise* from the dataset, I filtered for English posts using langdetect (Shuyo, 2010). I also introduced a minimum character requirement of four characters.

For the second experiment we wanted to obtain a better estimate on the accuracy of humans on this dataset. We conducted an experiment with 1000 randomly sampled posts - 200 for each of the category (*Happiness*, *Sadness*, *Fear*, *Affection* and *Anger*). In addition to the 1000 posts, we handpicked 15 posts as control questions that we considered easy - three for each emotion label. We divided the posts into 33 questionnaires. Ten questionnaires had 11 posts and three control questions and 23 questionnaires with ten posts and three control questions each. We introduced an additional requirement of a maximum of 500 characters so that workers do not spend too much time on one sample. We used Amazon Mechanical Workers to conduct the annotation. A total of five workers annotated each sample. We used the control questions to identify bad workers. If two out of the three control questions were answered wrong by a worker, I labeled the worker and their answers as "bad workers". A total of 160 unique workers answered the questionnaires. Out of these 160 workers, we marked 70 workers as "Bad Workers". We used these annotations to derive two measures of human performance. First, an "individual accuracy" was inferred from the average per item accuracy of the five workers, subtracting the number of invalid answers ("Don't know" and answers by "bad workers"). For a histogram of individual worker performance, I refer to figure 2 in the appendix. Second, we derived a "wisdom of crowd" label through a majority vote of the five workers. If the majority vote was "Don't know" or the majority of the workers were bad workers, we chose the next best emotion label. In case of a tie between two or more than two labels, we chose a random emotion from the tied emotions.

Validation for the best classifier is performed on two additional benchmarks to test generalizability. First, on the ISEAR dataset (Dan-Glauser & Scherer, 2012) because it includes texts that do not come from social media, consider emotional situations, and was collected from psychology students in the 1990s. Because students were instructed to think of situations for an emotional state and not assign an emotional state to a situation, sentence structure is different from the Vent dataset. Validation of this data should reveal whether the classifier can generalize to texts that come from a completely different time epoch. Examples for the ISEAR dataset are available in table 9 in the appendix. The ISEAR dataset only contains a total of 7666 examples with 1093 to 1096 examples per class. Second, the EmoInt dataset which was used for the emotion classification task of task 1 of the SemEval 2018 benchmark competition (Mohammad et al., 2018). I chose this dataset because SemEval is the most prominent sentiment and emotion analysis benchmark from texts focusing on social media data. The dataset contains tweets collected from Twitter between 2016 and 2017. However, the annotation task differ slightly from the annotations in Vent, such that annotators were asked to assess the intesity of an emotional state of the writer instead of the emotion the writer wanted to express. Examples for the EmoInt dataset are available in table 10 in the appendix.

I assess model performance for the three test sets using the weighted average F1-Score over all emotions. I also denote the weighted F1-Score for each emotion. I chose the F1-Score because it equally weights Precision and Recall. I use the weighted average because emotion labels are not equally distributed - see table 7 in the appendix - and detecting emotions is equally important for all classes. For classifier comparison, bootstrapping is used to derive confidence intervals. For the three Vent test sets, I randomly drew samples with a size equal to the number of samples in each respective test set for 10,000 rounds with replacement. The median of the bootstrapped results provides the average performance.

Confidence intervals are denoted by the 2.5 and 97.5 percentile of the ordered bootstrapped results. I also use bootstrapping to derive confidence intervals and median performance on the human comparison dataset, the ISEAR dataset and the SemEval dataset. Here I also use a random sample with replacement with a size equal to the number of samples in each respective dataset for 10.000 rounds. Confidence intervals are calculated in the same way as described above. Because emotion categories are equally distributed in the data used to assess human performance, I use the accuracy score to compare classifiers and human performance.

I use two McNemar tests to compare the classifier performance to human performance. One McNemar test is performed on an individual level under the exclusion of "Don't know" answers and answers from "bad workers". Another McNemar test is performed on the majority votes. The McNemar test was here performed by considering each item once.

To stay consistent with the report of Adoma et al. (2020) on the ISEAR dataset, I report F1, precision and recall. Because the emotions *Disgust*, *Shame* and *Guilt* are not present in the ISEAR dataset, I excluded the texts with these emotions from the data. I compare the classifier to a BERT model trained on 80 percent of the ISEAR dataset and testing on the remaining 20 percent by Adoma et al. (2020). This can be regarded as an upper bound for performance.

On the SemEval dataset, I report the Pearson correlation between the softmax output for each emotion and the intensity score for the label annotated by the workers. Because *Joy* is not part of the Vent dataset, I added the softmax output of the emotions *Happiness* and *Affection* and relabeled the result as *Joy*. Results are compared to the median team and the winning team. The winning team used a custom architecture that they call SeerNet (Duppada et al., 2018). It consist of multiple encoding techniques - such as DeepMoji (Felbo et al., 2017), skip-thought vectors for sentence representations (Kiros et al., 2015) and custom lexical features that the authors fed into four XGBoost models (Chen & Guestrin, 2016).

Because the emotions are not equally distributed in the train set - see table 7 in the appendix - I used random undersampling to train all the classifiers. Undersampling leads to 845,800 samples for each emotion in the train set, based on the least frequent emotion *Happiness*. I split the train set into 90 percent and 10 percent development sets for all classifiers for parameter tuning and loss monitoring. I use TF-IDF with Support Vector Machine Classifier with a linear kernel and average Fasttext supervised classifier as baseline methods. Because of the space complexity of Support Vector Machines, I used a subsample of 10.000 samples for each emotions - thus a total of 50.000 samples to train the TF-IDF Support Vector Machine model. The Fasttext supervised classifier averages the word embeddings per document and adds a multinomial logistic regression to derive the final prediction. The regression was fitted using stochastic gradient descent with a learning rate of 0.3, the default learning rate. I tested learning rates of 0.01, 0.1, 0.3 and 0.5. A learning rate of 0.3 was determined to perform best on the development set.

I also trained two transformer architectures. I trained a cased BERT (Devlin et al., 2019) architecture provided by the HuggingFace library (Wolf et al., 2020) and a roBERTa model provided on the Huggingface model hub. An uncased BERT was also trained but showed slightly lower performance, therefore I only report the results for the cased BERT. Because Vent contains many emojis and these are not recognized by the pretrained BERT architecture by default (Delobelle & Berendt, 2019), I added the 200 most frequent emojis to the tokenizer. For a depiction of the most frequent emojis in the Vent dataset, see figure 3 in the appendix. The second transformer model I used was roBERTa pretrained on a Twitter dataset for a sentence prediction task as part of the TweetEval benchmark (Barbieri et al., 2020). RoBERTa (Liu et al., 2019) was chosen because the byte-pair tokenizer could be better able to adapt to social media texts, which contain spelling mistakes, abbreviations and emojis (Liu et al., 2019). I chose a model pretrained on Twitter data because the Vent dataset shows huge similarities with Twitter in that users share posts and react to other people's posts. However, one crucial difference is that Vent posts do not have a 128 character limit. For a depiction of the character length distribution in the Vent train set, I refer to figure 2 in the appendix.

Both transformers were trained using the Pytorch Lightning framework (Falcon, 2019). I deduced the appropriate learning rate for the models from a method called the learning rate finder (Smith, 2017). The learning rate finder uses a learning rate range test. Different learning rates are applied on mini-batches of the data starting from a very small learning rate and progressively moves to a very high learning rate. The learning rate finder suggests a learning rate where the loss on the mini-batch with respect to the learning rate shows the steepest descent. Figure 5 and 6 in the appendix depict the learning rate range plot for BERT and roBERTa respectively. For the BERT architecture, a learning rate of 3.63e-4 was determined. For the roBERTa architecture, a learning rate of 8.32e-5 was determined. Training with the suggested learning rate showed a small improvement over a the learning rate suggested in the literature of 2e-5 for roBERTa. For BERT, the suggested learning rate was too high, resulting in no progress relating to the F1-Scores on either train or test set. I, therefore, used a standard learning rate of 2e-5 to train the BERT architecture.

I used an Adam Optimizer (Kingma & Ba, 2014) in conjunction with a linear learning rate scheduler. The learning rate scheduler increases the learning rate during training before then decreasing the learning rate to allow for more fine weights adjustments. One-third of the batches are warm-up steps. The remaining two-thirds are used to decrease the learning rate. I chose a token limit of 128 tokens because the focus of the emotion classifier was to detect emotions on social media datasets, and most of the posts in Vent contain fewer characters, see figure 2 in the appendix. The batch size was set to 512. Ten epochs were used as a maximum number. However, the F1 on the validation set was used as a stopping criterium. Both transformers terminated their

training after six epochs, with the highest score occurring in the fifth epoch. I conducted the training of the transformer models on the NVIDIA Cluster of the Technical University of Graz on a single NVIDIA QUADRO RTX 8000.

3. Results

Table 1. Median F1-Scores for all classifiers on the Vent random, time and user test data. Confidence intervals are derived using bootstrapping. The best performance is marked as bold.

Test type	Emotion	TF-IDF+SVM	Fasttext	BERT+Emojis	Roberta-Twitter
random	Affection	54.0 [53.8, 54.2]	63.6 [63.4, 63.8]	72.1 [72.0, 72.3]	**73.00 [72.8, 73.1]**
	Anger	55.7 [55.6, 55.9]	63.3 [63.1, 63.5]	70.9 [70.8, 71.1]	**71.6 [71.4, 71.7]**
	Fear	52.0 [51.8, 52.2]	59.0 [58.8, 59.2]	66.8 [66.7, 67.0]	**67.8 [67.6, 68.0]**
	Happiness	50.2 [50.0, 50.4]	63.6 [63.3, 63.8]	72.0 [71.8, 72.2]	**72.5 [72.3, 72.7]**
	Sadness	53.3 [53.2, 53.5]	63.3 [63.1, 63.4]	69.6 [69.5, 69.8]	**70.3 [70.1, 70.4]**
	Average	53.3 [53.2, 53.4]	62.5 [62.4, 62.6]	70.0 [69.9, 70.1]	**70.8 [70.7, 70.9]**
user	Affection	54.0 [53.8, 54.2]	60.7 [60.4, 60.9]	70.4 [70.2, 70.6]	**70.5 [70.4, 70.7]**
	Anger	55.6 [55.4, 55.8]	61.6 [61.4, 61.8]	69.6 [69.4, 69.8]	**69.8 [69.7, 70.0]**
	Fear	52.0 [51.8, 52.2]	57.4 [57.2, 57.6]	65.6 [65.4, 65.8]	**65.9 [65.8, 66.1]**
	Happiness	50.1 [49.9, 50.4]	60.4 [60.1, 60.6]	**70.1 [69.9, 70.3]**	70.0 [69.7, 70.2]
	Sadness	53.4 [53.3, 53.6]	62.2 [62.1, 62.4]	68.8 [68.7, 69.0]	**69.2 [69.0, 69.3]**
	Average	53.3 [53.2, 53.4]	60.6 [60.5, 60.7]	68.7 [68.6, 68.8]	**69.0 [68.9, 69.0]**
time	Affection	59.6 [59.4, 59.8]	65.2 [65.0, 65.4]	73.6 [73.4, 73.7]	**74.0 [73.9, 74.2]**
	Anger	54.8 [54.6, 55.0]	60.7 [60.5, 60.8]	68.9 [68.7, 69.0]	**69.3 [69.1, 69.4]**
	Fear	51.9 [51.7, 52.1]	56.8 [56.6, 57.0]	64.5 [64.3, 64.7]	**65.0 [64.8, 65.1]**
	Happiness	46.0 [45.8, 46.3]	56.8 [56.5, 57.0]	**66.5 [66.2, 66.7]**	66.1 [65.9, 66.4]
	Sadness	55.7 [55.6, 55.9]	64.3 [64.2, 64.5]	70.4 [70.3, 70.6]	**71.0 [70.9, 71.2]**
	Average	54.7 [54.6, 54.8]	61.6 [61.5, 61.7]	69.2 [69.1, 69.3]	**69.7 [69.6, 69.8]**

Table 1 depicts the results for all the classifiers on the Vent random, time and user test sets. The TF-IDF-SVM classifier provides the worst results, followed by the supervised Fasttext classifier. The two transformer models provide the best performance. When considering the average median performance, the emoji enhanced BERT model provides slightly worse performance than the roBERTa model on all three test sets. Also, concerning individual emotions, we observe that the roBERTa model provides the best performance for all test sets and all emotions except for *Happiness* on the user and time test set.

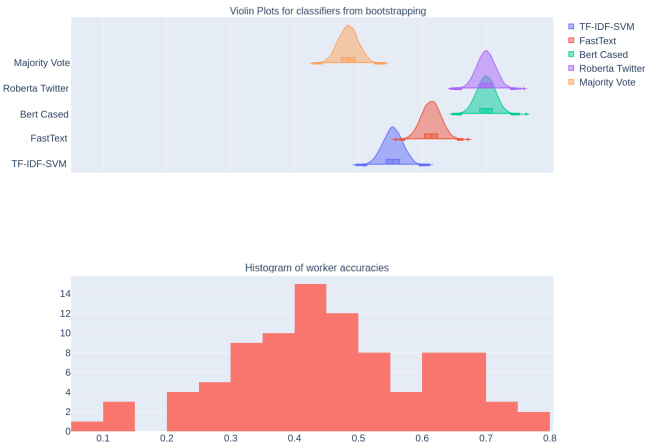
However, the confidence intervals for the two classifiers reveal a huge overlap for both transformer models. Therefore, we can conclude that the performance of the two transformer models seems comparable. With respect to the type of the test set, we observe a slight performance deterioration from random test set to the user and time test set. The performance decrease is also slightly larger for the user test set than for the time test set. Concerning the performance on individual emotions, we see that *Affection* and *Happiness* are easiest to detect on the random and time test set. Interestingly we see that performance on *Happiness* deteriorates for the time test set, indicating that there are time specific features that influence the detection of *Happiness*. BERT, roBERTa and Fasttext all show the worst performance on *Fear* among all three test sets.

For the second experiment, I compared the performance of the classifiers trained on Vent to human performance on a sample of 1000 posts randomly drawn from the random test set. The results are depicted in table 2. Again, we observe

Table 2. Comparison of the models trained on the train set with the accuracy of human annotators on a sample from the random test set. I assessed average individual worker performance per item, as well as a performance of an worker ensemble calculated from a majority vote.

Emotion	TF-IDF+SVM	Fasttext	BERT+Emojis	Roberta-Twitter	Individual Worker	Worker Majority Vote
Affection	56.0 [49.2, 62.9]	56.5 [49.5, 63.3]	67.6 [61.0, 74.0]	**70.5 [64.2, 76.8]**	41.0 [36.0, 46.2]	53.5 [46.6, 60.5]
Anger	57.5 [50.5, 64.3]	64.5 [57.8, 71.1]	71.6 [65.0, 77.7]	**68.0 [61.3, 74.4]**	40.0 [34.7, 45.3]	34.4 [28.0, 41.1]
Fear	53.0 [46.2, 59.9]	59.0 [52.1, 65.8]	**66.0 [59.2, 72.5]**	65.5 [58.7, 72.0]	33.1 [28.3, 38.0]	41.5 [34.7, 48.3]
Happiness	62.0 [55.1, 68.6]	60.5 [53.7, 67.0]	76.6 [70.4, 82.2]	**81.1 [75.3, 86.2]**	63.2 [57.7, 68.5]	65.5 [58.8, 72.1]
Sadness	48.5 [41.6, 55.3]	66.5 [59.8, 73.1]	**68.5 [62.0, 74.8]**	65.0 [58.3, 71.5]	49.0 [43.7, 54.1]	47.0 [40.0, 53.9]
Average	55.4 [52.3, 58.4]	61.4 [58.4, 64.4]	**70.0 [67.1, 72.9]**	**70.0 [67.2, 72.8]**	45.2 [42.9, 47.6]	48.4 [45.3, 51.5]

Figure 1. Violin plot of classifier accuracy gathered from bootstrapping (top). Histogram of accuracy of each worker (bottom).



that BERT and roBERTa show comparable and better performance than the baseline models, when we observe the average accuracy. Regarding the performance on individual performance, we observe that roBERTa performs slightly better than BERT on *Affection*, *Anger* and *Happiness*. Regarding the workers' performance in this annotation task, we can see that although the majority vote of the workers shows higher accuracy than the individual performance of the workers, they still perform the worst out of all the classifiers. Comparing overall roBERTa performance to human performance using a McNemar test revealed highly significant differences between roBERTa and individual human performance (t=195, $p < 0.001$) as well as for the majority vote (t=74, $p< 0.001$). Figure 7 depicts violin plots of the distribution of accuracy scores for the classifiers and the majority vote gathered from bootstrapping at the top and a histogram of the accuracy scores for each worker at the bottom. Figure 7 and 8 in the appendix depicts the confusion matrix for the majority vote and roBERTa respectively. We observe that for the majority vote, *Affection* and *Happiness* are mixed up, as well as *Fear* and *Sadness*. For the roBERTa model *Anger* and *Sadness* are more frequently mixed up than *Fear* and *Sadness*. RoBERTa is also prone to mix up

Affection and *Happiness*.

Table 3. Evaluation of roBERTa trained on Vent and only tested on the ISEAR data with a comparison with a previous BERT trained on the ISEAR data

	BERT trained on ISEAR			roberta trained on Vent		
Emotion	Precision	Recall	F1	Precision	Recall	F1
Joy	**89.0**	**94.0**	**92.0**	82.7 [77.7, 88.4]	90.8 [86.0, 94.8]	86.3 [82.0, 89.8]
Fear	**89.0**	**84.0**	**86.0**	80.3 [74.6, 85.8]	72.5 [66.4, 78.2]	76.2 [71.4, 80.5]
Anger	56.0	74.0	64.0	**74.2 [68.6, 79.6]**	**81.0 [75.6, 86.0]**	**77.4 [73.0, 81.4]**
Sadness	**85.0**	**73.0**	**78.0**	76.8 [70.7, 82.3]	69.8 [63.6, 75.8]	73.2 [68.1, 77.6]
Disgust	83.0	62.0	71.0	-	-	-
Shame	64.0	64.0	64.0	-	-	-
Guilt	54.0	60.0	57.0	-	-	-

For the third experiment, I tested the roBERTa model on the ISEAR dataset. The results are summarized in table 3. For comparison, I use the results from Adoma et al. (2020). The authors trained an uncased BERT on 80 percent of the ISEAR dataset and testing on the remaining 20 percent (Adoma et al., 2020). I calculated the output for *Joy* by summing the outputs of roBERTa for *Affection* and *Happiness*. I excluded samples labeled as *Disgust*, *Shame* or *Guilt*. The results show that the classifier trained on the ISEAR dataset performs better for emotions *Joy*, *Fear* and *Sadness*. However, the classifier trained on Vent achieves better performance on the emotion *Anger*.

For the fourth experiment, I tested the roBERTa model on the EmoInt dataset from the SemEval 2018 competition. Table 4 depicts a comparison between SeerNet - the winning model, the median team and the roBERTa model trained on the Vent data. The comparison is conducted using the metric from the competition, the Pearson correlation on the logits for each label with the labels derived from the annotation of seven workers using Best-Worst-scaling. Overall, the performance of the roBERTa model seems comparable to the performance of the median team but considerably worse than the performance of the winning team. [2]. Some selected examples are listed in table 5. Possible reasons for these differences are provided and discussed in section 4.

Table 4. Evaluation of roBERTa trained on Vent on the EmoInt dataset compared to the winning model and median team performance using the competition metric.

	SeerNet (Rank 1)	Median Team	Roberta trained on Vent
Emotion	Pearson Correlation	Pearson Correlation	Pearson Correlation
Joy	**79.2**	64.8	62.5 [61.1, 69.1]
Fear	**77.9**	67.4	60.1 [55.8, 63.9]
Anger	**82.7**	65.4	66.4 [62.4, 70.1]
Sadness	**79.8**	63.5	63.7 [58.9, 68.0]
Average	**79.9**	65.3	63.8 [61.8, 65.8]

[2]Differences in the predictions from the roBERTa model and the official labels are collected in the file `semeval_differences.txt`

Table 5. Selected missclassification on EmoInt. Intensity score was derived from best-worst scaling.

Text	EmoInt label (Intensity)	Vent roBERTa
"I gave up on the U20 Rugby bet on the Roosters! nrl "	Joy (0.58)	Anger
"Anyway I'm in a car with a furious white men and I have a really funny story to tell when I'm sober :')"	Anger (0.33)	Happiness
"#Obama #DOJ have destroyed USA!These #CharlotteProtest are acts of #terrorism dating back to Ferguson Terrorism is how it should be treated"	Fear (0.60)	Anger
"Not a great start but good comeback from the boys to earn a point. Bring on Saturday #blues"	Sadness (0.36)	Happiness

4. Discussion

First, the results in section 3 reveal that machine learning models, especially transformer architectures, can accurately infer human emotions from text, given a limited set of emotions in a multi-output classification task. The best performing classifiers are a cased BERT model, emojis added to the tokenizer, and a roBERTa model pretrained on Twitter data. The median performance of the roBERTa model is marginally higher than median performance of the BERT model. RoBERTa uses a byte-pair tokenizer, allowing it to behave more flexible in case of unknown word pieces. Unseen emojis, for example, are encoded by the tokenizer using byte-pairs, allowing for a more robust classifier.

Classifier performance is relatively stable over all three test sets. However, we observe a slight deterioration in performance concerning the time and even more pronounced for the user test set. All classifiers are affected by this deterioration. Both - user and time test set - are more restrictive than the random test set. A performance deterioration in the user set indicates that the classifier picked up on user-specific language features. Performance deterioration in the time test set suggests that the classifier picked up on language features that are very sensitive to the time of data collection. It is no surprise that language is constantly adapting and changing over time. Social media serves as an accelerator to this phenomenon.

Second, I compared the performance of the classifiers to human annotators using a small random test sample. Human annotators were sourced from Amazon Mechanical Turk and annotated the samples, which we spread over multiple questionnaires to lower processing time per worker. A total

of five workers annotated each sample. In addition to the samples, for each questionnaire, I handpicked three easy control questions. Workers that incorrectly answered two or more out of three control questions were marked as bad workers and excluded from the evaluation.

Despite this check, human annotator performance was deficient, and all the trained classifiers performed better on the samples. Introducing a majority vote slightly improved accuracy scores. Given the results obtained from the bootstrapping and referring to figure 7 we observe that the TF-IDF model with a Support Vector Machine is the only model for which accuracy score distribution overlaps with the distribution of the majority vote. No overlap between accuracy scores was observed for all other classifiers, demonstrating the superior performance in this limited experiment. Statistical tests revealed that differences in performance of roBERTa and the majority vote and individual performance are highly significant. Possible reasons for the difference in human and algorithm performance are difficulties in language comprehension, a lack of effort and motivation or a lack of knowledge about the dataset. These findings have considerable implications for the field of emotion recognition from text. The fact that emotion labels by crowd workers show little accordance with self-labeled emotions reveals possible problems with datasets that rely on crowd workers for annotations, such as the EmoInt dataset.

The main problems with crowdsourcing are difficulties in language comprehension and a lack of effort and motivation. Concerning the lack of language understanding, this might be the case if English is not the first language of the annotators. Annotators might also be unfamiliar with abbreviations or idioms used in the samples or the emotional labels. Concerning the lack of effort and motivation, we must consider that Mechanical Turk workers are incentivized to spend as little time as possible on the questionnaires to maximize their salary. In addition to these problems, familiarity with the dataset and the emotional labels should also affect performance. The confusion matrix reveals that in general mixups between the human majority vote and the roBERTa are similar. *Affection* and *Happiness* are frequently mixed up, as well as *Anger* and *Fear*.

Third, to test the generalizability of the classifiers on data that is different to Vent, I tested the classifier on the ISEAR data. I remapped the emotions from the Vent classifier such that they matched the labels in the ISEAR dataset. Although the classifier trained on the ISEAR dataset foreseeably performed better on four out of five tested emotions than the classifier trained on the Vent dataset, the performance was comparable. It even outperformed the ISEAR classifier for the emotion *Anger*. This is surprising as the ISEAR dataset is very different from the Vent dataset. Firstly, the ISEAR dataset was composed in the 1990s - it is therefore written in a completely different style than the social media data in Vent. Differences in style come from emojis, which are not used in the ISEAR data, frequency of spelling mistakes and sentence length. Secondly, the psychology students were asked to find situations for each emotion instead of matching emotions to their situation. Although this might not sound like it should make a difference, this greatly impacts the sentence's grammatical structure. See table 9 in the appendix.

Last but not least, I also tested the classifier on the EmoInt dataset from the SemEval 2018 competition. Although the classifier performance is comparable to that of the median team, it is worse than the winning team that used a custom architecture. Overall it is not surprising that this model performs worse than specialized architecture, especially since the classifier was not trained on this dataset. However, a look at the misclassifications also reveals dubious competition labels. In table 5 I collect some dubious misclassifications. It might also be the case that there is a relationship between emotion intensity scores and the softmax output of the classifier, however this relationship is not strictly linear.

Now, regarding the limitations of this work. We should consider that on the Vent platform, emotion labels are given according to the emotion that the author wants to express. However, the emotion that the author wants to express might be different from the emotional state that the author is in. Considering the example: *"Let go of resentment, it will hold you back"*. RoBERTa labels this as *Affection*, while one could reasonably assume that the emotional state of the author at the time of writing the post was most likely *Anger*. Additional limitations are due to the fact that we collected a benchmark for human performance using only Mechanical Turk workers. To get a real benchmark on human performance, a less biased selection should be considered. If we want to prove that machine learning models offer superior performance to humans in textual emotion recognition, future work should recruit people with high emotional understanding like psychotherapists for annotations. Regarding a selection bias, we should acknowledge that the Vent platform attracts a non-representative sample, namely people who want to express how they feel and want to see others who express how they feel. This selection bias is also supported by the fact that *Happiness* and *Affection* are the least frequent emotions while *Sadness* takes the top spot.

The results of this work should motivate researchers in the field of textual emotion recognition to consider self-labeled data instead of crowdsourcing annotations. Previous datasets such as the EmoInt dataset could be improved by highlighting dubious classifications and re annotating these examples by psychologists or other motivated individuals trained in emotion recognition.

Acknowledgements

I want to thank my team members from the Computational Social Science Unit at the Complexity Science Hub and the TU Graz. Special thanks to my supervisor David Garcia and Anna di Natale for conducting the Amazon Mechanical Turk experiments. I would also like to thank Max Pellert for his help with the Nvidia Cluster setup and various technical suggestions and Alina Herderich who suggested datasets for the literature review.

References

Adoma, A. F., Henry, N.-M., Chen, W., and Andre, N. R. Recognizing emotions from texts using a bert-based approach. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 62–66. IEEE, 2020.

Alammar, J. The illustrated transformer, 2018. URL http://jalammar.github.io/illustrated-transformer/.

Bai, Q., Dan, Q., Mu, Z., and Yang, M. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:2221, 2019.

Barbieri, F., Camacho-Collados, J., Ronzano, F., Anke, L. E., Ballesteros, M., Basile, V., Patti, V., and Saggion, H. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 24–33, 2018.

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., and Neves, L. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Chen, T. and Guestrin, C. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Ćorić, N., Primorac, M., and Leko, O. Usage of emojis in business communication. *Hum: časopis Filozofskog fakulteta Sveučilišta u Mostaru*, 13(19):269–270, 2018.

Dan-Glauser, E. S. and Scherer, K. R. The difficulties in emotion regulation scale (ders). *Swiss Journal of Psychology*, 2012.

Delobelle, P. and Berendt, B. Time to take emoji seriously: They vastly improve casual conversational models. *arXiv preprint arXiv:1910.13793*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Duppada, V., Jain, R., and Hiray, S. SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 18–23, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1002. URL https://aclanthology.org/S18-1002.

Ekman, P. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Falcon, WA, e. a. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 2019.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. doi: 10.18653/v1/d17-1169. URL http://dx.doi.org/10.18653/v1/D17-1169.

Garcia, D. and Rimé, B. Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological science*, 30(4):617–628, 2019.

Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., and Schweitzer, F. Ideological and temporal components of network polarization in online political participatory media. *Policy & internet*, 7(1):46–79, 2015.

González-Carvajal, S. and Garrido-Merchán, E. C. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

Hancock, J. T., Landrigan, C., and Silver, C. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 929–932, 2007.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Keshtkar, F. and Inkpen, D. Using sentiment orientation features for mood classification in blogs. In *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1–6. IEEE, 2009.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. Skip-thought vectors, 2015.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

Lykousas, N., Patsakis, C., Kaltenbrunner, A., and Gómez, V. Sharing emotions at scale: The vent dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pp. 611–619, 2019.

MacCann, C. and Roberts, R. D. New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8 (4):540, 2008.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mohammad, S. and Bravo-Marquez, F. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 65–77, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-1007. URL https://www.aclweb.org/anthology/S17-1007.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17, 2018.

Mohsen, A. M., Hassan, H. A., and Idrees, A. M. Documents emotions classification model based on tf-idf weighting measure. *International Journal of Computer and Information Engineering*, 10(1):252–258, 2016.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*, 2019.

Ortony, A. and Turner, T. J. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

Qaiser, S. and Ali, R. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29, 2018.

Ramos, J. et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pp. 29–48. Citeseer, 2003.

Rosenthal, S., Farra, N., and Nakov, P. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Schweitzer, F., Krivachy, T., and Garcia, D. An agent-based model of opinion polarization driven by emotions. *Complexity*, 2020, 2020.

Scott, C. F., Bay-Cheng, L. Y., Prince, M. A., Nochajski, T. H., and Collins, R. L. Time spent online: Latent profile analyses of emerging adults' social media use. *Computers in Human Behavior*, 75:311–319, 2017.

Shuyo, N. Language detection library for java, 2010. URL http://code.google.com/p/language-detection/.

Smith, L. N. Cyclical learning rates for training neural networks, 2017.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020.

Sundaram, V., Ahmed, S., Muqtadeer, S. A., and Reddy, R. R. Emotion analysis in text using tf-idf. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 292–297. IEEE, 2021.

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical nlp pipeline, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Winarsih, N. A. S., Supriyanto, C., et al. Evaluation of classification methods for indonesian text emotion detection. In *2016 International seminar on application for technology of information and communication (ISemantic)*, pp. 130–133. IEEE, 2016.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Zhang, Z. and Luo, L. Hate speech detection: A solved problem? the challenging case of long tail on twitter, 2018.

A. Appendix

A.1. Initial human performance benchmark pilot

Table 6. Initial human performance benchmark pilot: Samples (n=60) human performance benchmark with coworkers (n=6). Performance is measured by average F1 score over all coworkers. A Roberta pretrained on Twitter data was used and trained on the Vent training data with random undersampling. The emotional categories *Surprise* and *Fear* showed little agreement with assessment of human annotators.

EMOTION	HUMANS (F1)	ROBERTA-TWITTER (F1)
AFFECTION	59	67
ANGER	53	54
FEAR	31	44
HAPPINESS	65	80
SADNESS	54	60
SURPRISE	**30**	**33**
AVERAGE	49	56

A.2. Exploratory Analysis of Vent

Table 7. Distribution of emotion labels over the train set, the randomly drawn test set, the set with new users and the time test set respectively. We see that Happiness is the least frequent class and sadness occurs more than twice as often throughout all test sets, even more than thrice as often in the time validation set.

EMOTION	TRAIN	RANDOM	USER	TIME
AFFECTION	1,072,255	151,986	156,486	193,941
ANGER	1,607,705	218,567	218,535	199,820
FEAR	1,432,801	196,075	200,223	181,358
HAPPINESS	845,800	113,123	111,735	88,751
SADNESS	1,848,326	256,656	260,401	286,560
TOTAL	6,806,887	936,407	947,380	950,430

Figure 2. Character length distribution for the all Vent posts with a length smaller than 2000 characters. Created with an overlay and opacity set to 0.75. Median number of characters in each Vent is 70. Angry posts are longer with a median character length of 80. The maximum character length for one post is 9534.

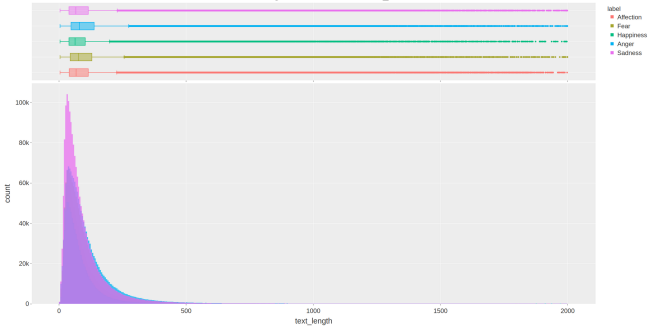


Figure 3. Emoji distribution for top 50 emojis in the train set

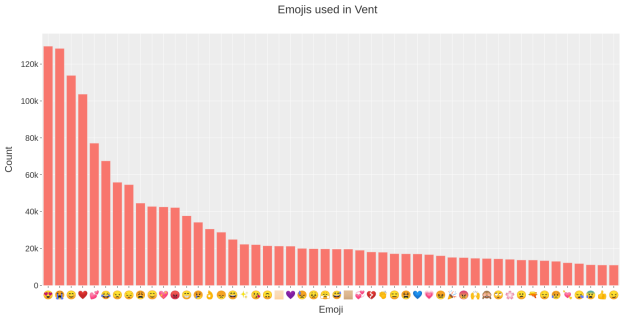
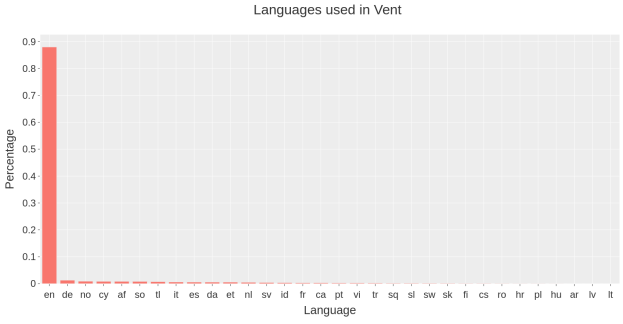


Figure 4. Languages used on a sample of 10,000 random posts. Some posts contain only smilies.



A.3. Learning rate finder plots

Figure 5. Learning rate finder results for BERT with added emojis tokens. The best learning rate is at 3.63e-4

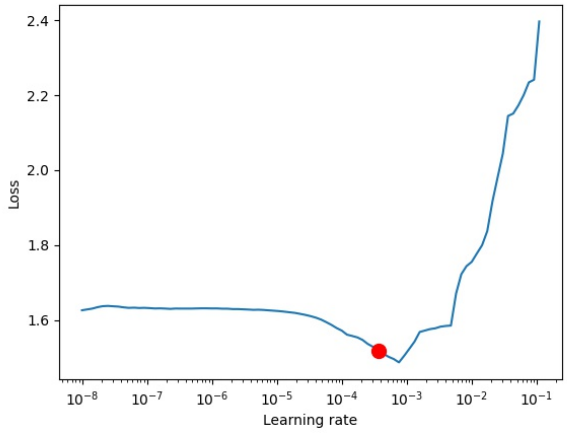
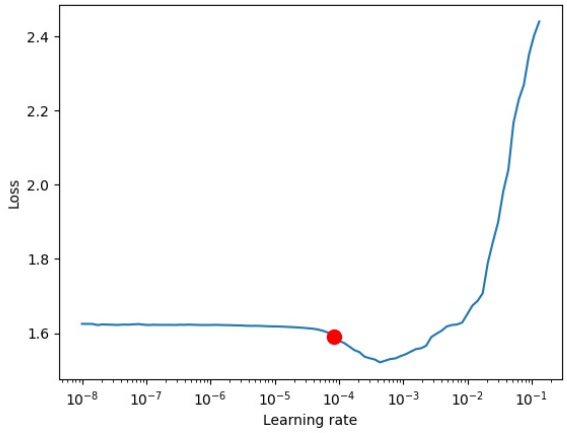


Figure 6. Learning rate finder results for Roberta pretrained on Twitter. The best learning rate is at 8.32e-5



A.4. Human benchmark

Figure 7. Confusion matrix on majority vote of workers.The columns refers to the predicted class. The row refers to the target class.

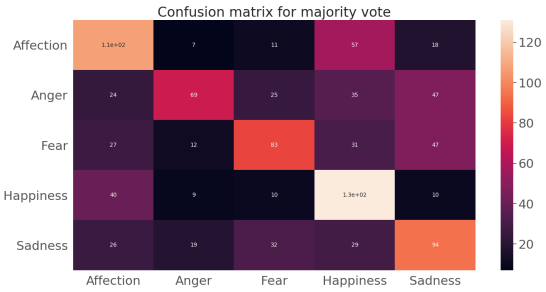
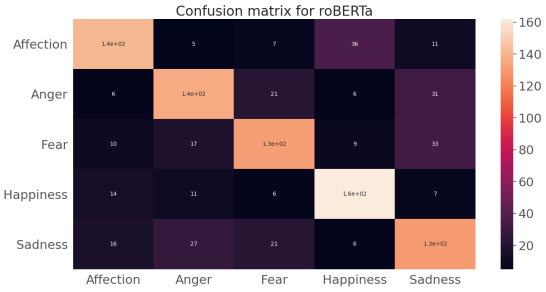


Figure 8. Confusion matrix on majority vote of roBERTa. The columns refers to the predicted class. The row refers to the target class.



A.5. Examples from the datasets

Table 8. Examples from Vent

Text	Label
”I am sooooo obsessed with George Harrison’s vocals on ’devil in her heart’ I just wanna explode. George Harrison... Ugh I can’t even...”	Affection
”My boyfriend is still mad at me and tomorrow will be our 1 month anniversary. Fucking hell I am so mad with myself.”	Anger
”Competition is this week and I am not ready at all!!!!!!!!!!!!!!!!”	Fear
”My crush told me I looked pretty...do you know how happy I am?”	Happiness
”the smiths are the best to listen to”	Sadness

Table 9. Examples from ISEAR

Text	Label
”During the period of falling in love, each time that we met and á especially when we had not met for a long time.”	Joy
”When I was driving home after several days of hard work, there á was a motorist ahead of me who was driving at 50 km/hour and á refused, despite his low speeed to let me overtake.”	Anger
”When I was involved in a traffic accident.”	Fear
”When I lost the person who meant the most to me.”	Sadness

Table 10. Examples from EmoInt

Text	Label
”Modern family never fails to cheer me up. Especially Phil.”	Joy
”At the point today where if someone says something remotely kind to me, a waterfall will burst out of my eyes”	Anger
”So nervous I could puke”	Fear
”Luckily I was helped by some good people. And they also managed to free me of my depression. Unfortunately it only lasted a little while.”	Sadness