EmoNoBa: A Dataset for Analyzing Fine-Grained Emotions on Noisy Bangla Texts

Khondoker Ittehadul Islam[†♣], Tanvir Hossain Yuvraz[†♣], Md Saiful Islam[♣♢], Enamul Hassan[♣]
[♣]Shahjalal University of Science and Technology, Bangladesh
[♢]University of Alberta, Canada
{khondoker07, tanvir54}@student.sust.edu,
mdsaifu1@ualberta.ca, enam-cse@sust.edu

Abstract

For low-resourced Bangla language, works on detecting emotions on textual data suffer from size and cross-domain adaptability. In our paper, we propose a manually annotated dataset of 22,698 Bangla public comments from social media sites covering 12 different domains such as *Personal*, *Politics*, and *Health*, labeled for 6 fine-grained emotion categories of the Junto Emotion Wheel. We invest efforts in the data preparation to 1) preserve the linguistic richness and 2) challenge any classification model. Our experiments to develop a benchmark classification system show that random baselines perform better than neural networks and pre-trained language models as hand-crafted features provide superior performance.[1]

1 Introduction

Identifying emotions has helped find solutions to numerous problems for English text, namely retrieving emotion from suicide notes (Yang et al., 2012; Desmet and Hoste, 2013), detecting insulting sentences in conversations (Allouch et al., 2018), and providing palliative care to cancer patients (Sosea and Caragea, 2020). A major reason behind such success is the amount of research and development invested towards fine-grained multi-label emotion tasks such as SemEval Affective Texts (Strapparava and Mihalcea, 2007), SemEval Affects of Tweets (Mohammad et al., 2018a) and GoEmotion (Demszky et al., 2020).

Bangla is the sixth most spoken language globally[2] and is the native language of Bangladesh.

Love	[**B**] এইরকম শো-অফ হাজার বার দেখতে চাই।
Joy	[**E**] *Want to see such show-off thousand times.*

Table 1: Example annotation from our dataset. **B** represents the original instance in Bangla, and **E** is its English translation.

With the country now graduating to a middle-income country with technologies now set to reach the rural areas (Basunia, 2022; Islam and Saeed, 2021), it is a timely need to understand the response to enhance the overall impact on social welfare and businesses.

Few datasets have been made public for detecting emotion in a low-resourced Bangla language (Rahman et al., 2019; Das et al., 2020, 2021). However, all such works are (1) small in size, including only a few thousand instances, and (2) incapable of cross-domain generalization, with coarse classification into Ekman or Plutchik emotions.

In this paper, we aim to create a multi-label emotion dataset of noisy textual data collected from social media on various topics. We use the Junto emotion wheel (Chadha, 2020) (Figure 1) as it reflects 21st century human psychology. During the data collection and annotation process, we invest efforts to improve the quality of the dataset by setting several predefined objectives. We also curate the test set such that it challenges any classification tasks. Our contributions can be summarized as follows:

- We propose EmoNoBa dataset, which comprises 22,698 multi-label **Emo**tion on **No**isy **Ba**ngla text. These texts are public comments on 12 different topics from 3 different social media platforms. Table 1 demonstrates a sample of our dataset.

[†]First and second authors contributed equally

[1]Data and code available at https://github.com/KhondokerIslam/EmoNoBa

[2]https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 128–134
November 20–23, 2022.

- We establish baselines by experimenting on linguistic features, recurrent neural networks, and pre-trained language models. We also shed light on various aspects of the problem throughout our analysis.
- We publicly release our dataset and model to foster research in this direction.

2 Development of EmoNoBa

Data Collection We set the following primary objectives before creating the dataset so that these objectives increase the generalization capabilities: Samples should contribute to making the dataset 1) domain independent and 2) less repetitive. We start by collecting user comments from YouTube, Facebook and Twitter on 12 most popular topics of Prothom Alo[3], the most circulated newspaper in Bangladesh[4]. Out of $\approx$ 50K collected comments, we keep the comments written in only Bangla alphabets. We remove duplicates and exclude instances shorter than three or longer than 50 word tokens to reduce repetitiveness and noise. Furthermore, we prioritize the instances for annotation that will increase the percentage of the unique word in the dataset (i.e., *unique word percentage* method (Islam et al., 2021)) to demand a more generalized and robust classification system.

Objective Given a predefined set of emotions - Junto-6 basic emotions, the goal is to identify all emotions conveyed in a piece of text.

Annotation We use five annotators for each instance. Emotion(s) voted by atleast three annotators were considered the final labels. Instances that could not be finalized this way were sent to authors for the final tag. We will refer to the former instances as *genInst* and the latter as *excInst*. We also kept the system fully anonymous for the authenticity of the annotations[5].

Furthermore, we evaluated the annotators with an accuracy metric. We will denote such accuracy as *AnnoAccu*. We have two different variations of equations for determining *AnnoAccu* as we have two categories of instances:

For *genInst*:

$$AnnoAccu = \frac{1}{|I|} \sum_{i \in I} \frac{T_i \cap O_i}{T_i}$$

[3] https://www.prothomalo.com
[4] https://www.topbanglanewspaper.com/
[5] The system is live at http://143.198.51.122/

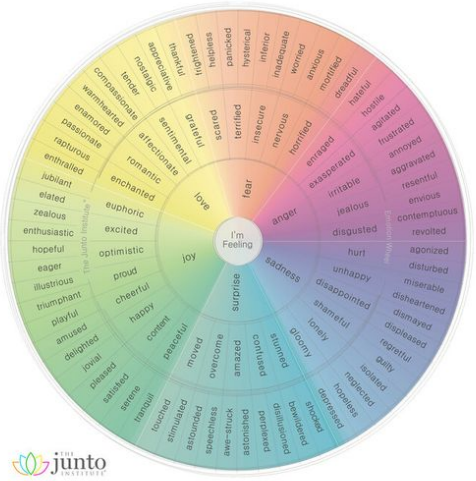

Figure 1: The Junto Emotion Wheel.

	love	*joy*	*surprise*	*anger*	*sadness*	*fear*	Avg. Score
Fleiss' κ	0.411	0.509	0.295	0.550	0.705	0.319	**0.465**

Table 2: Inter-Annotator Agreement Score for each emotion and the mean of all the scores.

For *excInst*:

$$AnnoAccu = \frac{1}{|I|} \sum_{i \in I} \frac{T_i \cap A_i}{T_i}$$

where T_i is the set of the emotions selected by this annotator for instance i, O_i is the set of the emotions selected by atleast two other annotators for instance i, A_i is the set of the emotions selected by the authors for instance i, and I is the set of instances.

We set the following criterion when choosing annotators. Annotators must be 1) well educated to understand the instances despite grammatical and spelling errors, and 2) active social media users to understand the context. Before selecting an emotion, we instructed them first to identify their child emotions from the Junto emotion wheel for better coherence. As such, 80 undergraduate students annotated 5 to 5,000 instances each, with 74 of them attaining *AnnoAccu* of 60% or more. Table 2 shows the Fleiss' κ (Fleiss, 1971) score of each emotion. One interesting finding here is that the Fleiss' κ scores are low for the less frequent emotions (see frequencies in Figure 2).

Statistics and Analysis. In total, we have 22,698 instances in the final dataset. The average length of the instance is 1.36 ± 0.82 sentences, and the

Emotion	Train					Test				
	Instances	Word Length	Sent. Length	*excInst* (%)	UW (%)	Instances	Word Length	Sent. Length	*excInst* (%)	UW (%)
Love	4,202 (20.53%)	11.66	1.32	2.09%	24.46%	390 (17.17%)	12.24	1.34	54.87%	49.87%
Joy	9,249 (45.19%)	10.56	1.27	1.32%	22.24%	857 (37.72%)	10.64	1.28	36.87%	45.89%
Surprise	939 (4.59%)	10.57	1.29	11.18%	45.66%	149 (6.56%)	10.54	1.29	71.81%	67.61%
Anger	3,905 (19.08%)	11.40	1.35	4.97%	27.01%	575 (25.31%)	11.22	1.33	54.60%	45.00%
Sadness	5,109 (24.96%)	11.93	1.36	2.00%	26.20%	572 (25.18%)	11.49	1.33	43.88%	49.16%
Fear	307 (1.50%)	11.96	1.37	20.85%	54.43%	93 (4.1%)	11.51	1.34	80.65%	65.52%
Total	20,468	11.72	1.36	1.52%	18.24%	2,272	11.52	1.35	40.18%	35.03%

Table 3: Summary statistics of our dataset with unique word (UW) percentage per emotion label.

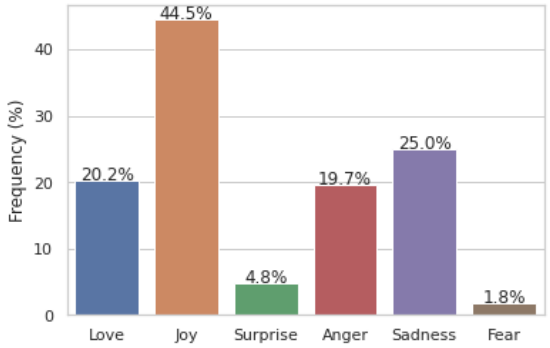

Figure 2: Percentage of instances labeled with a given emotion in our dataset.

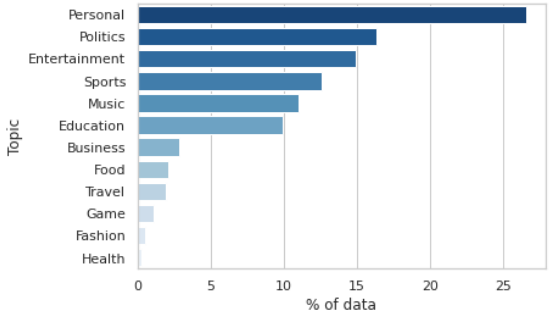

Figure 3: Topic distribution of our dataset.

average length of the sentence is 11.70 ± 10.70 words. Moreover, 77.28% of our instances source from Youtube, and 15.3% contain multiple emotions. Figure 2 shows the percentage of data labeled with a given emotion. Observe that *sadness*, *anger*, and *joy* are the most frequent emotions while *surprise* and *fear* are the two least frequent emotions. We also present the topic distribution in Figure 3. The vast majority of data are from *Personal* due to vlogging, and the least from *Health*.

We performed per-multi-label stratified split to create training (90%) and testing (10%) sets. Test set received precedence on *excInst*. In the cases of overflows, leftover instances were inserted into the training set and vice versa (Table 3). As *excInst* challenged human annotators due to noise and skeptical contextual understanding, we believe it will also challenge any classification model.

3 Methodology

In this section, we present the methods we used to develop a benchmark model for EmoNoBa.

3.1 Lexical Feature

We extract word (1-4) and character (1-5) n-grams from the instances as these lexical representations have shown strong performance in different classification tasks. Then we vectorize each instance with the TF-IDF weighted scores and train on linear SVM (Cortes and Vapnik, 1995) models.

3.2 Recurrent Neural Network

Due to the capability of capturing sequential information from both directions in texts, we use Bi-LSTM (Hochreiter and Schmidhuber, 1997). We put attention mechanism on top (Bahdanau et al., 2015) to put more weight on the words crucial for correct classification. To initialize the embedding layer, we consider 1) FastText (Grave et al., 2018) pre-trained Bangla word embeddings as it utilizes semantic information from the texts, and 2) random initialization to compare the results. FastText has coverage of 57.13% on our dataset as their training data are formal Bangla texts from Wikipedia, trained with character 5-gram.

3.3 Pre-trained Language Model

Due to the recent success of BERT (Devlin et al., 2018), we use Bangla-BERT-Base (Bangla-BERT; Sarker, 2020). This model has shown better performance against any other transformer-based models on a variety of Bangla lingual tasks. We only fine-tune its output layer with our training data.

4 Experiments & Results

4.1 Experimental Setting

We implement our experimental framework using Scikit-learn (Pedregosa et al., 2011), Pytorch (Paszke et al., 2019), and Transformers (Wolf et al.,

Method	Love	Joy	Surprise	Anger	Sadness	Fear	Macro Avg
Random	24.30	43.20	11.42	33.57	32.71	7.52	25.46
Bi-LSTM + Attn. (FastText)	0.0	52.71	0.0	0.0	22.70	0.0	12.57
Bi-LSTM + Attn. (Random)	0.0	57.79	0.0	18.49	51.97	0.0	21.38
Bangla-BERT	18.33	52.30	11.70	22.37	42.96	0.0	24.61
Word 1-gram (W1)	39.04	59.64	26.91	44.94	59.14	14.81	40.75
Word 2-gram (W2)	31.84	51.74	8.24	31.63	43.33	2.08	28.14
Word 3-gram (W3)	18.11	30.36	2.58	11.45	11.22	0.0	12.29
Word 4-gram (W4)	7.67	54.82	0.0	3.38	1.39	0.0	11.21
W1 + W2	40.93	61.68	21.87	46.79	60.35	11.76	40.56
W1 + W2 + W3	40.90	60.95	21.99	47.88	60.22	6.19	39.69
W1 + W2 + W3+ W4	41.04	61.14	22.68	**48.75**	60.56	6.19	40.06
Char 2-gram (C2)	37.30	60.88	25.75	37.21	54.74	14.75	38.44
Char 3-gram (C3)	39.14	59.15	24.80	45.85	55.35	16.07	40.06
Char 4-gram (C4)	40.28	60.39	26.47	46.38	58.40	12.00	40.65
Char 5-gram (C5)	41.42	59.07	15.91	43.79	59.28	8.25	37.96
C1 + C2 + C3	39.34	60.66	22.57	45.96	55.80	14.16	39.75
C1 + C2 + C3 + C4	41.13	61.42	24.22	46.42	59.80	**16.98**	41.66
C1 + C2 + C3 + C4 + C5	**42.96**	62.70	23.00	46.34	61.81	11.88	41.45
W1 + C1 + C2 + C3 + C4 + C5	39.55	61.82	**28.84**	48.16	62.79	11.65	42.14
W1 + W2 + W3 + C1 + C2 + C3	42.35	**63.52**	25.37	48.30	**63.57**	12.00	42.52
W1 + W2 + W3+ W4 + C1 + C2 + C3	42.22	63.09	27.45	48.63	**63.57**	11.88	**42.81**

Table 4: Binary Task F1-score of each emotion class and Macro Average F1-score of each method on EmoNoBa.

2020). We evaluate our methods using macro-averaged F1-score. As the baseline system, we compare our results with the scores obtained by randomly guessing a prediction. To reduce noise, we replace the numerical tokens with a CC token and normalize English and Bangla sentence stoppers. We randomly picked 10% instances from the training set to build the development set.

We only tune the regularizer C[6] of the SVM model. For training the BiLSTM model, we perform hyper-parameter tuning the batch size, learning rate, dropout rate, number of LSTM cells, and layers. For fine-tuning Bangla-BERT, we only tune on learning rate and batch size.

4.2 Results & Findings

Results We report our experimental results on the test set in Table 4. Results show neural network and transformer-based models have lower F1-scores than the random baseline. To breakdown, the Bi-LSTM model with FastText embedding only predicts two emotions that have the least *excInst* in the test set (Table 3). Moreover, the same model with random initialization better identifies the same emotions alongside the next least frequent *excInst*'s emotion (*anger*). The transformer-based model follows the same trend and understands the following least frequent *excInst*'s emotions (*love*, *surprise*). However, none of the models predicts the most *excInst*'s *fear* emotion. One reason for such performance across these models could be that the unique word percentage is high for the most frequent *excInst* emotions (Table 3) since Islam et al. (2021) attained similar performance on their sentiment analysis task with similar corpus and textual properties. The dip in the performance on our task is because the models had to understand more deep levels of expressions.

Among the word n-gram, unigram achieves the best result by at least 12%. Combining the word grams yields better results but fails to surpass the standalone unigram model. On the other hand, the less showing of character n-grams verdicts that the task does not rely much on the character level information as with the increase of n-grams induces better results. Integrating all word 1-4 grams with character 1-3 grams provides the best result of 42.81 F1. Similar result was achieved in Arabic and Spanish languages in SemEval 2018 E-c task (Mohammad et al., 2018b).

Findings Notice that both the negative emotions (*anger*, *sadness*, *fear*) and the positive emotions (*love*, *joy*, *surprise*) provides best results on sub-word or phrase level information.

5 Further Analysis

Dominant Features Table 5 shows some of the strong word n-grams from each emotion. We find

[6]We tested on these values: $1e^{-3}$, $1e^{-2}$, 0.1, 1, 10 (best).

Love	Joy	Surprise
বেস্ট বেস্ট (best best)	😂😂😂	মুগ্ধ (amazed)
😍😍😍	খুব সুন্দর লাকছে (looks very nice)	কেনো? (why?)
অসাধারণ 💙💙💙 (extraordinary)	আপনি বেস্ট (you are best)	. . .
খুব সুন্দর লাকছে (looks very nice)	তুমি সেরা (you are best)	আর কি (what more)
Anger	**Sadness**	**Fear**
বালের (slang)	বিচার নাই দেশে (there is no justice in the country)	ভয় থাকতো । (fear remained)
বেশি হয়ে গেছে (too much)	বাজে ভাবে উপস্থাপন (poorly introduce)	আল্লাহ হেফাজত কর (God protect us)
না কি (no what)	কান্না (cry)	ফাঁসি (execution)
তুমি খুব খারাপ (you are really bad)	শো-অফ (show-off)	বেড়ে গেলো। (increased.)

Table 5: Examples of some of the strongest word n-grams from each label with their English translations.

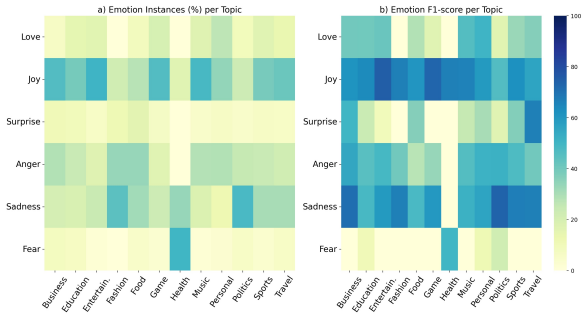

Figure 4: a) % of data of each Emotion per Topic in the test set; b) Binary Task F1-score of each Emotion per Topic from the best model.

that strong positive emoticons and compliments act as an indicator of positive emotions. On the other hand, criticism and slang fill up negative emotions. Observe that words such as বেস্ট (best) and খুব সুন্দর (very nice) occur in both *love* and *joy* emotions. The reason is these words can vary in context.

Error Analysis To investigate the test errors, we present the distribution of emotion per topic and the models' performance in Figures 4a and 4b. Notice that the model additionally predicts *sadness* in *joy* and *love* instances in *Personal* topic. The reason is negative words, such as "শো-অফ" (show-off), are the strongest words of *sad* emotion (Table 5), but they can also lie in instances containing positive emotions (refer to Table 1). Also observe that the model finds it tough to differentiate between *love* and *joy* emotions in *Business*, *Education*, *Entertainment*, *Music*, *Personal*. Reason could be phrases like "খুব ভালো লেগেছে" (looks very nice), strong word n-gram of both the emotion (Table 5), can turn from *enchanted* (child of *love* in the wheel) emotion in *Music* or *Entertainment* to *excited* (child of *joy* in the wheel) emotion in *Business* or *Education*. These two emotions also lie side-by-side in the emotion wheel (Figure 1). Hence the future work could revolve around improving transformer-based models for Bangla language. This could improve sub-word level contextual understanding and consequently help to better identify both sentimental emotions.

6 Conclusion

In this paper, we present EmoNoBa, a dataset for fine-grained emotion detection on Bangla text collected from comment sections of social media platforms on 12 different domains. We found that hand-crafted features performed comprehensively better than neural models. As the future work, we will exploit the findings identified in this work while incorporating contextual understanding.

References

Merav Allouch, Amos Azaria, Rina Azoulay, Ester Ben-Izchak, Moti Zwilling, and Ditza A Zachor. 2018. Automatic detection of insulting sentences in conversation. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–4. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sazzad Reza Basunia. 2022. E-commerce in rural bangladesh: The missing dots. *The Business Standard*.

Raman Chadha. 2020. The junto emotion wheel: Why and how we use it. *The Junto Institute*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Avishek Das, MD Asif Iqbal, Omar Sharif, and Mohammed Moshiul Hoque. 2020. Bemod: Development of bengali emotion dataset for classifying expressions of emotion in texts. In *International Conference on Intelligent Computing & Optimization*, pages 1124–1136. Springer.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. *arXiv preprint arXiv:2104.08613*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

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*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

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

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Quazi Tafsirul Islam and Nur Ibna Saeed. 2021. E-commerce in bangladesh: prospects and challenges. *New Age*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018a. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018b. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Md Rahman, Md Seddiqui, et al. 2019. Comparison of classical machine learning approaches on bangla textual emotion analysis. *arXiv preprint arXiv:1907.07826*.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Tiberiu Sosea and Cornelia Caragea. 2020. Canceremo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger,

Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hui Yang, Alistair Willis, Anne De Roeck, and Bashar Nuseibeh. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical informatics insights*, 5:BII–S8948.