# Stress Analysis in Social Media

**Sven Šćekić, Marko Kuzmić, Lovro Bučar**

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{sven.scekic, marko.kuzmic, lovro.bucar}@fer.hr

**Abstract**

Here we can write an abstract of the paper. The abstract is a paragraph of text ranging between 70 and 150 words.

## 1. Introduction

This section is the introduction to your paper.

## 2. Related Work

In scientific papers, this section usually (but not necessarily) briefly describes the related research and what makes the presented approach different from it.

## 3. Dataset

The dataset that is used in this paper was introduced in a paper which collected posts from a social media webside called Reddit (Turcan and McKeown, 2019). On that website users can create posts in communities which focus on specific topics called subreddits.

The dataset consists of posts from subreddits where stressful topics are likely to be discussed in between January 1, 2017 and November 19, 2018. These subreddits fall into one of these categories:

- **Social**: The posts from this category are from the subreddit called r/relationships. Posts from this subreddit talk about problems in a relationships, romantic or non-romantic.

- **Abuse**: posts from this category describe topics related to abuse in relationships and users share their experiences and advices. Subreddits that fall into this category are r/domesticviolence and r/survivorofabuse.

- **Anxiety**: subreddits from this category are r/anxiety and r/stress. Users in posts from this category talk about these mental illnesses, their symptoms, share advices and stories regarding these illnesses.

- **PTSD**: Same as the Anxiety category, posts from this category talk about mental illness, but focus on Post-Traumatic Stress Disorder. The only subreddit from this category is r/ptsd and same as with posts from the Anxiety category, users share advices and stories and ask questions about the illness.

- **Financial**: posts from subreddits that fall into this category talk about difficult financial situations, share stories about homelessness and other stressful financial topics. Subreddits from this category are r/almosthomeless, r/assistance, r/food_pantry and r/homeless.

In the dataset there are in total 187,444 posts from these 10 subreddits. The distribution of these posts can be seen in the Table 1.

### 3.1. Data Annotation

A portion of the data was annotated using Amazon Mechanical Turk, a crowdsourcing marketplace that allows individuals to outsource jobs.

Primary job for annotators was to determine if there was stress present in the sentence that they were presented. The definition for stress was taken from the Oxford English Dictionary states that stress is ‘a state of mental or emotional strain or tension resulting from adverse or demanding circumstance’. Each annotator first had to take a qualification test to ensure that they would label the sentences correctly. In the qualification test annotators were given instructions on how to correctly label post segments.

After that, each annotator was given 5 segments which they had to annotate with one of the following labels: “Stress”, “Not Stress” or “Can’t Tell”. In addition to the qualification test, each annotator was given one of the 50 ‘check questions’ which were labeled by creators of the dataset. If they did not label these ‘check questions’ correctly their annotations were not included in the final dataset.

Since posts can often be longer, which would prove to be complicated for the annotators to label, they were divided into five-sentence chunks. Added bonus of this technique is that the data can be used to determine the specific location of stress in the post.

In total, 3,553 labeled data points were collected out of which 39% had perfect agreement. With 52.3% of the data being labeled as stressful, the dataset is nearly perfectly balanced. Labeled data was split into two subsets: train and test subset. The train subset contains 2,838 data points and the test subset consists of 715 data points.

## 4. Models

Motivated by the results of the before mentioned paper, our approach focused on three different models which solve the given classification problem.

Table 1: Distribution of collected posts

| Topic | Subreddit Name | Total Posts | Avg Tokens/Post | Labeled Segments |
|---|---|---:|---:|---:|
| Social | r/relationships | 107,908 | 578 | 694 |
| Abuse | r/domesticviolence | 1,529 | 365 | 388 |
|  | r/survivorsofabuse | 1,372 | 444 | 315 |
| Anxiety | r/anxiety | 58,130 | 193 | 650 |
|  | r/stress | 1,078 | 107 | 78 |
| PTSD | r/ptsd | 4,910 | 265 | 711 |
| Financial | r/almosthomeless | 547 | 261 | 99 |
|  | r/assistance | 9,243 | 209 | 355 |
|  | r/food_pantry | 343 | 187 | 43 |
|  | r/homeless | 2,384 | 143 | 220 |

Firstly we will introduce the **logistic regression** model which the authors of the dataset presented as the best solution, excluding state-of-the-art solutions such as BERT. Then we are going to introduce two transformer approaches, one being **DistilBERT** and the other one, currently very popular, **Chat GPT 3.5 Turbo**.

### 4.1. Preprocessing

Before we trained our models we needed to do some form of preprocessing on our data. This step is important because it enables adequate learning from the training set. Given the dataset already contained pretty clean and structured data, we didn't need to do a lot of preprocessing to get the desired results. We started with the basic stop word removal, followed by digit and punctuation removal. Then we performed lemmatization which enabled us to properly train the word embeddings for the logistic regression model and tokenizer for the DistilBERT model.

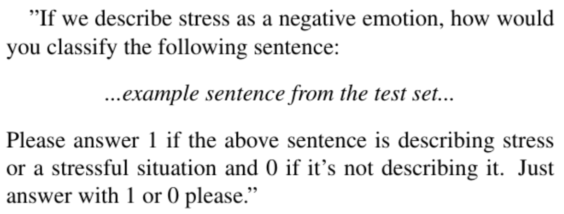
### 4.2. Logistic regression

Logistic regression is a commonly used machine learning model which can be used for binary classification problem. This is a simple model which doesn't require long and expensive training, so, inspired by the great results that this model showed in the original paper (F-score of 79.8 (Turcan and McKeown, 2019)), we have also decided to include it in our own research.

Logistic regression model expects a fixed size input but words and sentences in natural language can be of any size. Today this problem is solved by using word embeddings which represent words from our dataset as multidimensional vectors. Their advantage over some other approaches that were used in the past, such as one-hot vector representation, is that they are able to find relations between words and produce similar vectors for similar words.In our approach we used **word2Vec** (Mikolov et al., 2013) embeddings which were trained on the whole training dataset and produced 300-dimensional representations. Learned word embeddings were then used for calculating the final representation for each training example. Here we experimented with 2 approaches:

1. **Averaging of word embeddings:** summing up each representation in the input sequence and dividing it with the number of words in the sequence

2. **TF-IDF averaging:** Term Frequency - Inverse Document Frequency is an algorithm that is used to determine an importance of the given word in the document. Importance is represented as a scalar value. This scalar value for each word is then multiplied with the word's corresponding representation. The final representation is then calculated in the same way as described in 1.



"If we describe stress as a negative emotion, how would you classify the following sentence:

*...example sentence from the test set...*

Please answer 1 if the above sentence is describing stress or a stressful situation and 0 if it's not describing it. Just answer with 1 or 0 please."

Figure 1: Text input for Chat GPT 3.5 Turbo API

### 4.3. DistilBERT

The second model we used was DistilBERT (Sanh et al., 2020). This model comes from the family of deep learning models called transformers (Vaswani et al., 2017). The main advantage of these types of models is the self-attention mechanism. Self-attention allows the model to find the importance of each word in a sequence and build more effective contextual representations.

DistilBERT is a model similar to the more popular BERT (Devlin et al., 2019), but faster and more lightweight while still preserving almost all the capabilities of the larger model. Because of its features and our limited hardware we decided it would be the best option for representing current SOTA solution in the field of transformers. Model and tokenizer were trained simultaneously on the whole training set using the implementations from the *HuggingFace* library.

### 4.4. Chat GPT 3.5 Turbo

The final model that we used in our research was Chat GPT 3.5 Turbo. This model was developed by OpenAI and it aims to generate human like responses based on the provided input text. We couldn't perform direct training on this model, but instead we used the provided API to prompt the model to give us the classification result for each example in the test set. You can see how we formulated our prompt in figure 1.

Given results were then cleaned up

## 5. Results

Finally, a section which describes the acquired results.

## 6. Conclusion

Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

## Acknowledgements

Here we can write a thank you to the professors and the assistants which were involved in TAR class.

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong, November. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.