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 dataset creators. If they did not label these 'check questions' right 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 subset and test subset. The train subset contains 2,838 data points and the test subset consists of 715 data points.

4. Baseline Models

This section which will describe the models that were used to determine the

5. Results

Finally, a section which describes the acquired results.

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

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

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.