Stress Prediction with Machine Learning Models and its Analysis

Tong Guan[a]

[a] *University of Oregon, Eugene, Oregon, United States*

Abstract

This final report focuses on binary classification of stress-related contents using the Kaggle stress prediction dataset, which consists of topics extracted from various subreddits including ptsd, anxiety, relationships and homeless. The study utilizes two different methodologies, namely Bags of Words(BoW) and Word2Vec, to analyze the dataset. Logistic Regression, Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) are then employed as classification models to compare their performance in distinguishing stress-related topics from non-stress-related topics.

1. Background

Stress is a widespread issue that significantly impacts individuals' well-being. With the increasing use of online platforms, people express their thoughts and emotions through digital communication, offering an opportunity to analyze their stress levels. Accurately distinguishing stress-related topics from non-stress-related ones is vital for understanding the prevalence and nature of stress, identifying at-risk individuals, and providing timely support. This research aims to leverage advanced natural language processing techniques and machine learning models to develop an effective method for classifying stress-related topics based on individuals' online posts. The results will contribute to targeted interventions, personalized mental health support, and a deeper understanding of stress factors in various contexts.

2. Methods and Experiments

To address the problem of binary classification of stress-related topics, a multi-step approach that combines natural language processing (NLP) techniques and machine learning models has been employed. The methodology of the stress prediction consists of the following key steps:

(1). Dataset Preparation: Following an initial analysis of the sentences within the dataset and a basic data visualization, the data undergoes preprocessing. This involves removing stopwords, punctuation, and special symbols, converting the text to lowercase, and tokenizing the text into individual words.

(2). Feature Extraction: Two feature extraction methods have been explored, namely Bags of Words (BoW) and Word2Vec. The Bags of Words approach represents each post as a numerical vector, where the frequency of each word corresponds to a specific dimension. This method effectively captures the distribution of stress-related terms present in the dataset. In contrast, the Word2Vec method generates dense vector representations of words by considering their contextual usage. This technique enables us to capture the semantic meaning and contextual information associated with stress-related topics.

(3). Classification Models: Three different classification models have been applied to compare their performance in distinguishing stress-related topics. Firstly, Logistic Regression, a widely used linear model for binary classification, provides a baseline for performance evaluation. Secondly, Support Vector Machines (SVM) construct decision boundaries to separate stress-related topics from non-stress-related ones. Lastly, Convolutional Neural Networks (CNN) are employed to capture local patterns and hierarchical representations in the text. Notably, CNN is not applied to bags of words due to their disregard for word order, which is crucial for CNNs. However, CNN is suitable for Word2Vec embeddings as they preserve semantic and contextual information, enabling the network to leverage text sequencing and improve performance in natural language processing tasks.

(4). Model Evaluation: The performance of various classification models is assessed based on their accuracy. To ensure the reliability of the evaluation, the dataset is divided into training and testing sets using an 80/20 split. The allocation of samples to these sets is randomized to avoid any biases or skewed distributions.

Table 1: Comparison of Different Methods in Stress Prediction

Feature Extraction	Classification	Accuracy
Bags of Words	Logistic Regression	0.7218
Bags of Words	SVM	0.7271
Word2Vec	Logistic Regression	0.6073
Word2Vec	SVM	0.5985
Word2Vec	CNN	0.5457

(5). Interesting Insights: Two interesting questions arise regarding the performance accuracy of different method combinations. The first question is: Does the utilization of Bags of Words generally lead to improved results for classification models compared with Word2Vec, and if so, why?

One possible answer might be, the differing performance outcomes between Bags of Words (BoW) and Word2Vec can be attributed to the specific dataset characteristics and classification task. The higher performance achieved with BoW suggests a strong correlation between the stress dataset and specific keywords or word frequencies. In contrast, when the classification task requires capturing the semantic meaning and contextual information of stress-related topics, Word2Vec will demonstrates its potential for delivering superior results.

The second question is: What factors contribute to the lower performance of CNN compared to traditional methods such as logistic regression and SVM in our stress dataset?

One crucial factor to consider is the relatively small size of our stress dataset, consisting of only 2820 lines of text. This limited dataset may not provide an ample amount of information for the CNN to effectively learn the intricate patterns associated with stress. Consequently, the CNN might encounter difficulties in generalizing to unseen stress-related data points, resulting in lower performance. While there could be other factors such as overfitting and hyperparameter settings, delving into a detailed exploration of these aspects is beyond the report.

3. Conclusion

In conclusion, this report investigated the task of distinguishing stress-related contents using different feature extraction methods and classification models. The observation highlighted that both Bags of Words (BoW) and Word2Vec can be effective in capturing stress-related information, with BoW performing better in scenarios where specific keywords or word frequencies are crucial for classification. Logistic regression and SVM demonstrated competitive performance, while the CNN approach exhibited relatively lower accuracy compared to traditional methods. The lower performance of the CNN can be attributed to factors such as limited training data, potential overfitting, and the importance of hyperparameter tuning.

Looking ahead, there are several promising directions for future work. Firstly, I will try to expand the dataset size and diversify the sources of the data. What's more, it will be important to explore advanced text representation techniques such as GloVe embeddings and BERT, which have demonstrated their potential in capturing nuanced information. In addition to these technical aspects, I'm also very interested in doing more research relevant to sentiment analysis or emotion recognition.

4. References

Kreesh Rajani. "Human Stress Prediction Dataset." Kaggle, 2021, www.kaggle.com/datasets/kreeshrajani/human-stress-prediction.

Serxio. "NLP Stress Prediction with SVM." Kaggle, 2021, www.kaggle.com/code/serxio/nlp-stress-prediction-with-svm.

Shivani Malhotra. "NLP Using GloVe Embeddings." Kaggle, 2021, www.kaggle.com/code/shivanimalhotra91/nlp-using-glove-embeddings.