Stress Detection in Social Media Post: Take Reddit as an Example

Student-Nr.: S158800

Exam: Concepts in Social Data Analytics (CBUSV2201E)

Study: MSc in Business Administration and Data Science
Date: 28. January 2024
Pages: 17
Characters incl. spaces: 31415

Content

Content of Tables

Content of Figures

Introduction

According to Statista (2023), nearly 5 billion people use social media such as Facebook, Reddit, and Instagram to connect with family and friends, express their feelings and update the latest information. Social media platforms generate a large amount of social data via users' online activities, revealing individuals' preferences, views on various topics and emotional expressions (Bergemann et al., 2022). The social data not only reveal the users' preference for the products or brands, given the widespread use of social media for emotional expression but the language used on the social media can also reveal the pattern of anxiety, depression or other emotions that can indicate stress, which can disclosure the stress, making the stress more observable to the researchers and the public health organisations.

Stress, while sometimes motivating, can adversely affect an individual's well-being and behaviour. (Schneiderman et al., 2005) Excessive stress may harm the individuals' mental and physical health, causing anxiety, depression, and potentially increasing mortality risk (Pearlin et al., 2005; Slavich et al., 2010). Through the application of social data analytics techniques, this paper aims to develop a stress-detecting tool to detect the stress expressed in social media posts, enabling timely assistance for individuals suffering from stress. Although the social media contents are presented in various formats, including texts, images and videos, this paper will focus on text-based social data. Text-based posts are particularly accessible for researchers to determine if content creators are experiencing stress, as individuals often express their feelings through text when seeking to share their emotions (Jadhav et al., 2019).

The research chose Reddit posts as the data source for model training due to Reddit's text-central nature. Compared to blog-based platforms such as X and Facebook, Reddit facilitates more detailed and lengthy expressions of help-seeking behaviour, enhancing the model's ability to distinguish stress indicators in posts.

The study will begin with a literature review, followed by an explanation of the research structure (including models and pre-processing techniques), data analytics on the dataset, results and discussion, and conclude with further considerations.

Literature Review

Turcan and McKeown (2019) analyzed 2,838 Reddit post segments from ten subreddits using logistic regression with domain-specific Word2Vec embeddings, lexical, syntactic, and social media features, alongside state-of-the-art BERT, a two-layer bidirectional Gated Recurrent Neural Network (GRNN), and Convolutional Neural Network (CNN) models. The results demonstrated that although the BERT model achieved the highest performance with an F-score of 0.8065, logistic regression showed comparable effectiveness with an F-score of 0.798 while utilizing fewer computational resources. This highlights the significance of domain-specific knowledge and the value of lexical features in stress detection.

Kim et al. (2020) analysed 488,472 posts from seven mental-health-related subreddits between January 2017 and December 2018 to detect users' mental states. To address the issue of users potentially suffering from multiple mental health issues, Kim et al. (2020) created six independent binary classifier models for different mental statuses. They employed XGBoost and CNN classifiers. The result suggested that CNN models achieved higher accuracy rates across all six mental statuses. The study also noted that class imbalance significantly affected model performance.

Shen and Rudzicz (2017) analysed 22,808 Reddit posts to detect anxiety using Word2Vec, Doc2Vec, N-gram language modelling, Latent Dirichlet Allocation (LDA), and Linguistic Inquiry and Word Count (LIWC) for text vectorization. They then applied logistic regression (LR), linear kernel support vector machine (SVM), and a neural network (NN) to determine the most effective vectorisation techniques for increasing machine learning model performance. Their findings showed that Word2Vec and N-gram models alone could achieve 91% accuracy, surpassing LIWC features. However, combining Word2Vec or N-gram with LIWC increased accuracy to 98%.

Turcan and McKeown's (2019) and Shen and Rudzicz's (2017) work highlight the importance of lexical features in detecting emotion in text-based data. However, existing research often overlooks interpersonal engagement features, such as the number of comments and upvotes (likes), which are the most significant features of social media platforms. Therefore, the first part of this paper explores if the social media metrics correlate with the stress expression on social media.

Additionally, due to the context dependency in natural language processing (NLP), many research efforts, including those by Turcan and McKeown (2019), employ bidirectional models or LSTM models to capture language dependency features. This paper also examines whether feature-based models, which may not account for text sequence, can better capture the meaning of the text, achieving results comparable to models that do consider text sequence in emotion detection.

Conceptual Framework

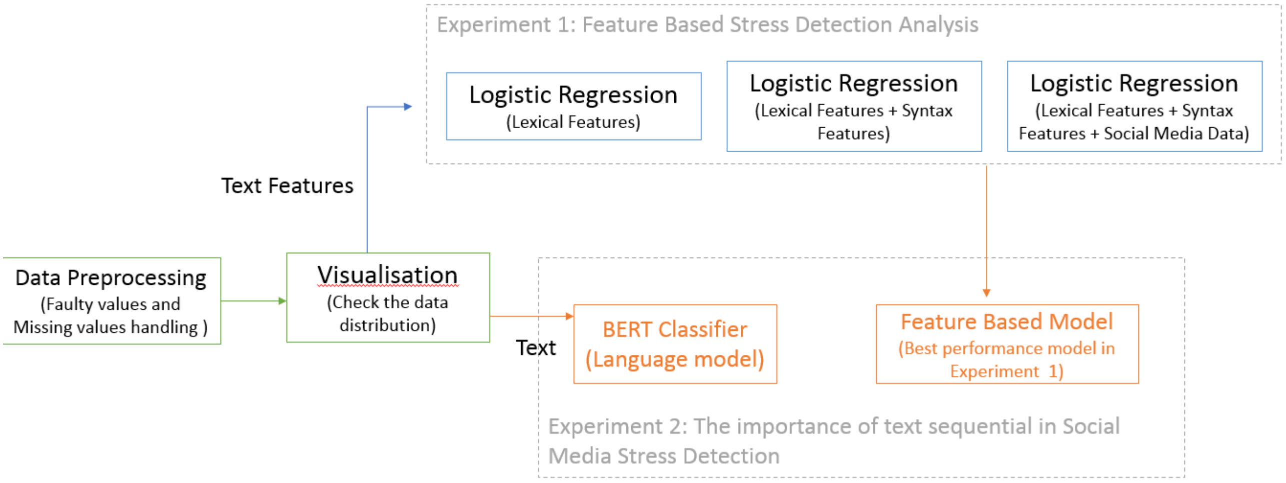
Experiment 1: Feature Based Stress Detection Analysis
Logistic Regression
(Lexical Features)
Logistic Regression
(Lexical Features + Syntax Features)
Logistic Regression
(Lexical Features + Syntax Features + Social Media Data)
Text Features
Data Preprocessing
(Faulty values and Missing values handling)
Visualisation
(Check the data distribution)
Text
BERT Classifier
(Language model)
Feature Based Model
(Best performance model in Experiment 1)
Experiment 2: The importance of text sequential in Social Media Stress Detection

Figure 1 Research Structure

The research structure of this paper includes two experiments: a feature-based stress detection analysis to understand how different types of features affect the stress detection model and an examination of text sequence importance using a comparison between the optimized feature-based logistic regression model and BERT model.

The study begins with data preprocessing for quality assurance, followed by data visualization to ensure check if the data distribution is balanced and conduct data analytics. Kim et al. (2020) note that skewed datasets can affect machine learning model performance. To mitigate this, data augmentation will be applied to address skewness. The first experiment employs a sequential addition method to examine the influence of different types of features on stress detection. Firstly, given the significance in various studies, the lexical features are input to the logistic regression model, followed by the syntactic features. Finally, the social media features are incorporated. As a part of experiment 2, the best performance feature-based model in experiment one is then compared with the BERT model to assess the impact of considering text sequence on stress detection.

Methodology

Data Description

Reddit organises user-generated content into topic-specific communities, which are known as subreddits, and its text-based structures with the comment system encourage users to create lengthy and detailed content to express the issues users encounter and seek help. These features are conducive to discussions on stress-inducing topics, such as mental health, abuse, and finance issues, making Reddit an ideal resource for analysing text-based stress patterns.

The DeReddit dataset, the data utilised in this paper, was published by Turcan and McKeown (2019). It collected 187,444 posts between January 1, 2017, and November 19, 2018, across ten different subreddits (domestic violence, a survivor of abuse, anxiety, stress, almost homeless, assistance, food pantry, homeless, PTSD and relationship) that related to stress-inducing topics. To avoid the influence of the length of the posts, Turcan and McKeown (2019) extracted 3553 five-sentence segments from these posts. Then, the segments are annotated by human assessors to determine the presence of stress or negative attitudes. Each segment is annotated as ‘stress’ (1) or ‘not stress’ (0). Additionally, Turcan and McKeown (2019) also created the 102 lexical features, 2 syntactic features and 4 social features for feature further analysing. Due to the predominance of lexical features, this research will utilize the sequential addition method to assess the impact of various feature types on the stress detection model.

Lexical Features

The lexical features in the dataset are categorized into three types. The first type involves scores from the Dictionary of Affect in Language (DAL), providing pleasantness, activation, and imagery scores for over 1,000 English words (Whissel, 2009). The dataset incorporates nine variables for the highest, lowest, and average scores in the text across these dimensions. The second type is the frequencies of words across 93 categories from the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) in the text. Lastly, sentiment scores calculated by the Pattern sentiment library (Smedt and Daelemans, 2012) range from -1 to 1, indicating negative or positive sentiments, respectively.

Syntactic Features

The syntactic features refer to the readability of the text. In this category, the dataset includes two variables, the Flesch-Kincaid Grade Level and Automated Readability Index (ARI). Flesch-Kincaid Grade Level uses the number of words per sentence and the number of syllabi per word to define the grade level needed for understanding the text. The ARI estimates the grade level based on characters per word and words per sentence, offering insights into the text's complexity.

Social Media Features

The social media features include interaction metrics from Reddit, such as the number of comments, upvote ratio (the number of upvotes divided by the total number of votes), downvote ratio (the number of downvotes divided by the total number of votes), and the submission score of the post, which is calculated as the number of upvotes minus the number of downvotes.

Data Pre-processing

In the data preprocessing, the records with the null value in the text or label are removed, while the missing values in the feature variable are assigned as 0, indicating the absence of certain features. 3 records which have the faulty data type in the text or label column are also removed. For the faulty value handling, the label value which is other than 0 and 1 are also removed.

Although natural language processing often requires text cleaning (e.g., lowercase transformation, special character removal, and tokenization), this research take two approaches, for the first experiment, the dataset provided features are used for feature analysis, while the BERT model with its own tokenisation and vectorisation layers. Therefore, explicit text cleaning and tokenisation are omitted in this research.

Data Analytics

Data Distribution

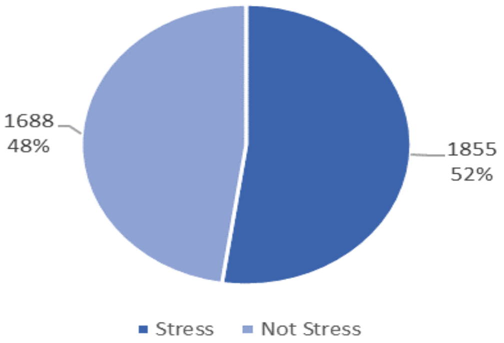

Figure 2 Stress and Not Stress Data Ratio

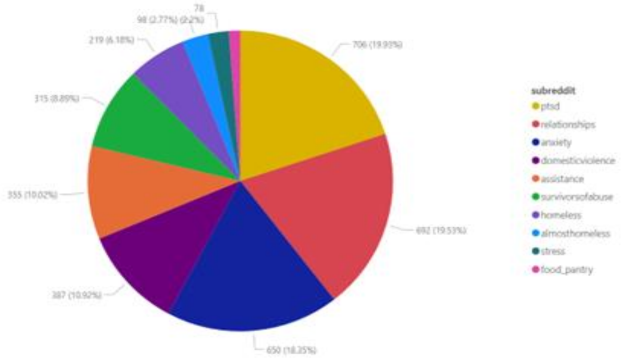

Figure 3 Data Distribution by Subreddit

In view of the total dataset, the dataset's distribution is relatively balanced, with 48% non-stress and 52% stress texts, negating the need for data augmentation to correct skewness (Figure 2).

The labelled segments, primarily sourced from PTSD, relationship, and anxiety subreddits, each contribute to around 20% of the dataset (Figure 3). However, due to the potential overlap in stress sources and word usage patterns among some subreddits, the decision was made to group similar subreddits together into five domains, enhancing the dataset's coherence and analytical utility.

The classification of label segments into domains based on stress sources includes Financial (stress from financial scarcity, covering subreddits like homeless and food_pantry), Interpersonal Interaction (stress from relationship dynamics), Abuse (experiences of violation), Mental Status (emotions and mental conditions, such as anxiety), and PTSD (unique stress expressions due to severe mental impacts). Each domain, holding approximately 700 records, ensures the model's robustness across varied social media contexts by reflecting diverse stress causes (figure 4).

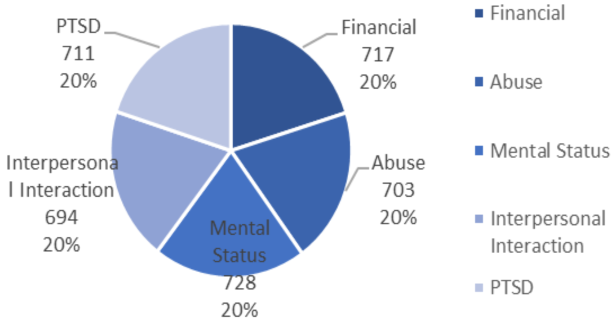

Figure 4 Data Distribution by Domain

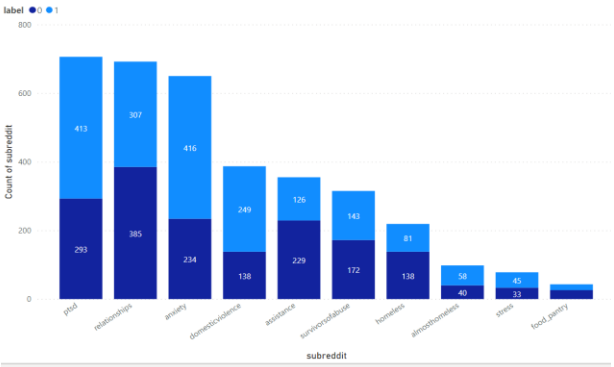

Figure 5 Stress and Not Stress Data Ratio by Sub Reddit

Subreddit Analysis

Figure 5 shows that stress and not stress data are present in different subreddits, revealing insights into which topics are more likely to be associated with stress. Topics related to mental status, like anxiety and stress, tend to have a higher proportion of stress-expressing posts, with about two-thirds of labelled segments from these subreddits classified as stress. PTSD and domestic violence topics also have a significant presence of stress-indicative posts. On the other hand, the relationship and finance-related issues, such as homelessness and assistance, show a lower ratio of the stress indicative posts.

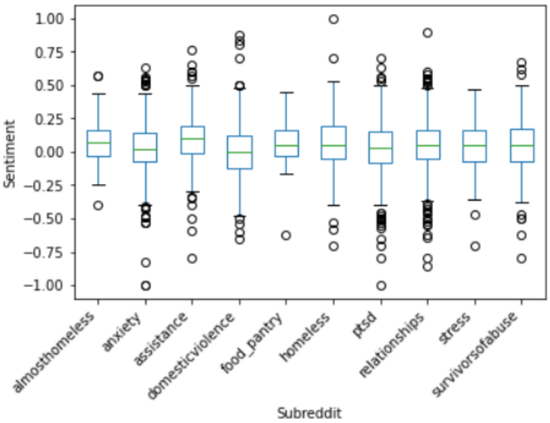

Figure 6 Sentiment Distribution by Subreddit

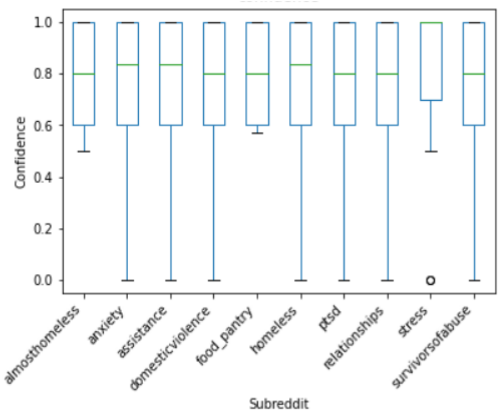

Figure 7 Confidence Distribution by Subreddit

Although some topics exhibit a higher ratio of stress-indicative posts, the sentiment scores across all topics are close to 0, suggesting a predominantly neutral tone (Figure 6). For topics with less severity, like almost homeless and food pantry, the variation in sentiment is also smaller compared to more severe topics such as PTSD and domestic violence.

Given the dataset involves manual data annotation, there's a pattern showing humans can more easily determine whether a post expresses stress if the post topic is stress. For topics other than stress, although the median confidence is less than the stress-related post, the median confidence rate is around 80%, indicating a high level of agreement among annotators on the presence of stress in posts.

Data sampling

Compared with the data obtained in the other research, the data points in this dataset are relatively less. To ensure that we have enough data to train the machine learning model, separate the dataset in an 8:2 ratio. When randomly choosing 20% of the label segments (715 segments) as the test set. The remaining 80% of the label segments (2838 segments) are used as the training set to train the machine learning model.

Modelling Framework

Logistic Regression

Logistic regression is chosen as the model to build the feature-based stress detection classifier in this research. Logistic Regression is a prominent discriminative classifier in the field of machine learning, known for its ability to directly model the posterior probability and establish a direct mapping from instances to labels. Its foundation on the sigmoid function, which maps input instances to a probability value between 0 and 1, makes it uniquely suited for binary classification tasks, including feature-based stress detection.

The logistic regression model can reveal the coefficient of each feature. Its interpretability makes the model suitable for conducting feature analyses and examining how social and syntactic features impact contribute to stress expression. Additionally, logistic regression performs well with high dimensional data, efficiently managing the 104 features to construct feature-based detection models. Lastly, due to the relatively simple structure of the logistic regression model, it can give a more explainable result for understanding the synergy of various features, providing the base to create a more advanced model.

Bidirectional Encoder Representations from Transformers (BERT)

The BERT model introduced by Vaswani et al. (2017) represents a significant advancement in natural language processing by pre-training on a large corpus of text to understand the nuances of language. The

BERT contain its own pre-processing layer to process texts by tokenizing them into words and subword units, incorporating special tokens to delineate sentences or adding padding for sequences of varying lengths. This tokenization is followed by transforming the text into vectors through word embeddings, capturing the meanings of words based on their contextual use during BERT's training phase.

BERT utilizes 12 transformer layers, leveraging self-attention mechanisms to extract relevant information from both the left and right context of the text. This bidirectional approach significantly enhances the model's capability to discern complex patterns and relationships within text sequences. As the state-of-the-art techniques, this study utilised the BERT to represent the model that takes the text sequence into consideration.

Performance Evaluation

Although the accuracy rate reveals insight into if the model correctly classified the instance in the text set, relying solely on accuracy to evaluate model performance can overlook scenarios where predictions are biased toward the majority class. In the context of stress detection, an excess number of false negatives could lead to overlooking individuals seeking help, while high false positives might result in unnecessary resource expenditure. To address this issue, the study also considers precision, which measures the ratio of true positive predictions to all positive predictions made by the model, and recall, which reflects the ratio of true positive predictions to all actual positive instances. Moreover, this study also takes the F1 score, which is the harmonic mean of precision and recall. The F1 score considers the trade-off between precision and recall. It addresses the possibility of selecting a model with high accuracy but a high number of false positives or false negatives since F1 score can be high only when both precision and recall rates are high. In addition to reporting the F1 score for each class, we also include the macro-averaged F1 score.

Additionally, the F1 score, the harmonic mean of precision and recall, is also considered in model performance evaluation to account for the trade-off between these two metrics. The F1 score is particularly useful for identifying situations that, despite having high accuracy, the model may also have a high number of false positives or negatives, as a high F1 score is achieved only when both precision and recall are high.

Results

Feature Based Model

Input Features	Accuracy	Precision	Recall	F1 Score
Lexical	0.7370	0.7384	0.7354	0.7356
Lexical + Syntactic	0.7399	0.7412	0.7382	0.7384
Lexical+ Syntactic + Social	0.7455	0.7463	0.7441	0.7443
Optimised Features	0.7455	0.7459	0.7443	0.7455

Table 1The Performance Metric of Feature-Based Logistics Regression Models

With lexical features, the logistic regression classifier achieves a 73.7% accuracy rate and a 73.56 f1 score, indicating effective feature capture. Incorporating syntactic features, the logistic regression model slightly improves each performance metric element by 0.003. This shows that post-readability is not affected by the stress expression. Including social media features like upvote and downvote rates further improves the FL score by about 0.005, indicating a weak correlation with stress expression and minimal influence on social media interaction patterns.

To minimise the noise added to the model and reduce overfitting, the absolute coefficients in the model that consider all types of features are utilised to filter the significant stress detection features. Features with coefficients above 0.2 are selected to create the optimised model, which is trained by the 18 most significant features. This model presents the accuracy rate as that of the logistic regression model while marginally improving the F1 score by 0.001. Consequently, the feature-optimised model as the best-performance feature-based model will be compared with the model considering the text sequence.

Additionally, the relatively stable performance metric shows nearly 75 % accuracy, and the f1 score might be the limit of logistic regression in this dataset.

Model Comparison

Model	Accuracy	Precision	Recall	F1 Score
Majority Baseline [1]	0.5132	0.5161	1.0000	0.6808
Optimised Feature- Based Logistic Regression	0.7455	0.7459	0.7443	0.7455
BERT	0.7944	0.7643	0.8699	0.8137

Table 2 Performance Metric

Compared with the Majority baseline, the feature-based logistic regression model significantly enhances accuracy by over 20% and F1 score by 0.1647, demonstrating the importance of lexical features in stress detection. The BERT model, which considers both word meaning and text sequence, increase 5% increase in accuracy and 0.0678 in the F1 score, achieving 79.44% of accuracy and 0.8137 F1 score in stress detection. This shows the importance of the text sequence in the stress presence text detection. Additionally, the 79% accuracy rate is close to the average confidence level of manual labelling 79.06%[2], implying the possibility of deploying the language model on stress detection in social media post. However, BERT's lower precision suggests a tendency towards false positives, labelling texts as stress-related more frequently. While false positives can waste resources, they are deemed less problematic than false negatives, which could overlook individuals needing stress support.

Discussion

Feature Analysis

Rank	Feature	Coefficient	Abs Coef.
1	lex_dal_avg_pleasantness	-1.994566	1.994566
2	lex_dal_min_pleasantness	-1.524426	1.524426
3	sentiment	-1.405589	1.405589
4	lex_dal_avg_imagery	0.884404	0.884404
5	lex_dal_min_activation	-0.586494	0.586494

[1] The Majority Baseline, as introduced by Turcan and McKeown (2019), serves as a baseline prediction method where all text segments are classified under the most frequent label present in the dataset.

[2] Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. arXiv (Cornell University). https://arxiv.org/pdf/1911.00133.pdf

6	lex_dal_max_imagery	0.497919	0.497919
7	lex_dal_max_pleasantness	-0.432344	0.432344
8	lex_liwc_feel	0.389339	0.389339
9	lex_liwc_death	0.377854	0.377854
10	lex_dal_min_imagery	-0.355434	0.355434
11	lex_dal_max_activation	0.318411	0.318411
12	lex_liwc_negemo	-0.290051	0.290051
13	lex_liwc_affect	0.276094	0.276094
14	lex_liwc_swear	0.255213	0.255213
15	lex_liwc_posemo	-0.234808	0.234808
16	lex_liwc_risk	0.233883	0.233883
17	lex_liwc_hear	0.230425	0.230425
18	lex_liwc_percept	-0.223057	0.223057
19	lex_liwc_ingest	-0.195602	0.195602
20	lex_liwc_pronoun	-0.190866	0.190866
80	syntax_fk_grade	-0.020857	0.020857
96	syntax_ari	0.005613	0.005613

Table 3 Significance of Features

Table 3 presents the coefficients and absolute coefficients for syntactic, lexical, and social media features, indicating the weight of each feature in the logistic regression model. Using absolute coefficients helps identify the significance of features by considering both negative and positive correlations with the label, offering insights into their importance in the analysis.

A significant finding is that only 18 of 104 features have absolute coefficients greater than 0.2. The finding shows that many features may not significantly contribute to stress detection. Furthermore, the table also highlights that the impactful features in stress detection are lexical since no syntactic or social media features present an absolute coefficient greater than 0.2. The minimal impact of the syntactic and social features implies the independence between the text readability and stress expression and between the readers' behaviours and stress expression.

In this study, another remarkable finding is the criticalness of DAL features, with eight out of nine such features ranking among the top 18 most significant indicators for stress detection. The fact underscores the importance of analysing the specific meanings and emotional resonances of words, such as how intense or negative the usage is, rather than counting how frequently words within certain emotional categories appear in the text. To further improve the model performance in stress detection, the model should focus on indicators that capture the intensity and emotional depth of word usage within texts to improve the sensitivity of the model to language that signals stress.

Additionally, the top three features are related to positivity in word use, the level of the least positive words, and the overall tone of the text, combined with the fact that the absolute coefficients of these top three features are nearly twice as large as that of the fourth-ranking feature. The fact underscores the critical role that positive language usage and the general emotional tone of a text play in expressing or indicating stress.

Computational Complexity Analysis al

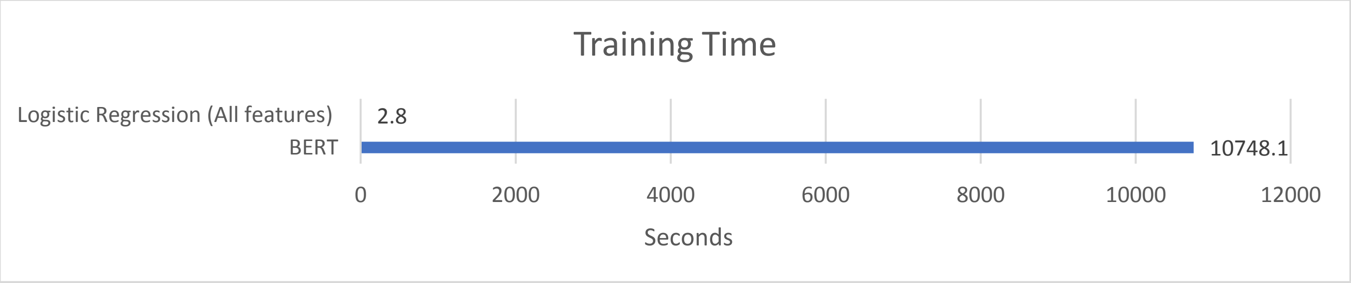

Figure 8 Training Time Comparison

In this research, model training time was analysed on an Intel Core i7-4720HQ @ 2.60GHz setup. The BERT model, with its text pre-processing layers, includes 12 transformer layers and 110 million parameters, compared with the feature-based logistic regression model which has only 18 parameters, highlighting the significant differences in training complexity. While the logistic regression model requires only 2.8 seconds for training, BERT needs almost three hours. Despite BERT's superior capabilities in text generation and classification, simpler models might be preferred for less complex tasks or datasets due to shorter training times and computational resource requirements.

Privacy and Ethic Issue

For collecting social data collection and analysing from social media platforms, if the collection only includes data set to be publicly visible, and if the data collected is not enough to identify individuals, then the data is not seen as personal data. Therefore, it does not have privacy issues. However, when data is collected on a large scale, there's still a possibility of identifying individuals through their posts if we know the posts come from the same person, even if names are not explicitly shown. In this case, to inform data subjects is necessary. This fact highlights the importance of anonymising data when collecting it from social media platforms. In this research, we only take the text and features of posts that are publicly published, excluding usernames, as a measure to avoid violating the privacy of social media users.

On the other hand, deploying stress detection tools on social media platforms to identify individuals in need and provide help should consider some privacy issues. Although the tool might only collect publicly posted content and contact the individual via the message function provided by social media, the large-scale collection can still enable data controllers to identify individuals' identities. This effectively transforms the data into personal data, thus requiring that users be informed. Furthermore, some stress-indicative posts may contain health-related information of the individual, such as mentions of PTSD experience. According to the General Data Protection Regulation (GDPR) Article 9, processing special categories of data (which includes health-related information) requires obtaining explicit consent from the data subjects, and the collected data should be used solely for the intended purpose. Therefore, to deploy the language model in big-scale social data analysis, the organizations should either contact the data subject directly or collaborate with the social media platforms to obtain consent from the users before deploying the model.

Further Research

Incorporating Topic Modelling Techniques in Stress Detection

Since this tool is designed for a Reddit-based environment, users can easily check the source of stress and what kind of help the individual might need by checking the subreddit the post belongs to. Therefore, this paper does not incorporate topic modelling as a part of the research. However, on other social media platforms such as X (Twitter), Facebook, and Thread, the posts are forced to be categorised into distinct communities; the topic of the text can only be identified by the hashtag, which is not as straightforward as Reddit.

To generalize the model's application on social media platforms, a potential research direction could include incorporating topic modelling techniques such as Latent Dirichlet Allocation (LDA). In other words, the model's output should not only indicate the presence of stress but also identify the attribution of the stress, such as financial issues or traumatic experiences. A multiclass classification model that can output the presence of stress and stress inductive source can enable users to better understand the stress resource of each case and provide relevant support.

Exploring More Complex Models on Feature-Based Prediction

Another possible research branch is to explore more complex machine learning and deep learning techniques on feature-based stress detection models. Logistic regression's simplicity, while minimizing computational resource demands, may limit its ability to capture complex patterns in the dataset. The potential underperformance of feature-based detection models compared to BERT might be attributed to the vast difference in their architectural complexity.

Incorporating models with higher computation complexity, such as Random Forest, which can capture non-linear relationships through ensemble learning, or neural network architectures, known for their ability to learn hierarchical representations of data, might improve the performance of feature-based models.

This approach would provide a more comprehensive comparison between models that consider text structure and those that do not, bridging the gap in effectiveness between simpler and more complex models.

Increasing Training Data

Since the dataset is manually annotated, compared with the size of the dataset from other studies, the dataset size is small. The insufficient amount of data might decrease the ability of the machine learning model to capture the pattern in the corpus, especially for the model, which has a relatively complex infrastructure. Although the manual annotation process in the dataset used in this research can ensure the quality of the data, it results in a dataset size that is considerably smaller compared to those used in other studies. The dataset size may impact the effectiveness of machine learning models in capturing patterns and correlations within the text. This influences the model with more complex architecture since these complex models are designed to capture intricate features within the data, a task that becomes increasingly difficult as the dataset's size decreases.

The smaller dataset also increases the risk of overfitting, where the model performs well on the training data but fails to generalize to new, unseen data. This issue shows the need to increase the dataset size in further research.

Conclusion

In the first experiment, the study demonstrated the independence of interpersonal interaction from stress expression, further suggesting that stress presence in textual data does not affect text readability. Moreover, it highlighted that the choice of word usage significantly impacts stress expression, more so than the frequency of words within specific categories.

In the second experiment, the significant difference in performance between the feature-based logistic regression model and the BERT model considering text sequence indicates the importance of text sequence in conducting stress detection within text-based data. Additionally, the BERT model, with a 79% accuracy rate, validated the potential of deploying language detection models on social media platforms to identify users who may require assistance. However, stress detection tools usage on a large scale might involve the handling of special category personal information, which necessitates obtaining user consent on social media platforms before deploying such models to aid individuals potentially suffering from stress.

Lastly, the computational analysis revealed that models with higher complexity demand significantly more computational resources and training time, necessitating a larger initial investment before deployment of the model. The facts showed that organizations with limited computational resources should consider model complexity when choosing the machine learning model for stress detection.

Remark

This research leverages Alteryx and Excel for data preprocessing, Power BI and Python for data analytics, and Python for modeling. The coding may include assistance from internet tutorials and online AI-based tools. For the full Python code and the dataset, please refer to the attached code file and dataset files.

Reference

Bergemann, D., Bonatti, A., & Gan, T. (2022). The economics of social data. *The RAND Journal of Economics*, *53*(2), 263–296. https://doi.org/10.1111/1756-2171.12407

Jadhav, S. B., Machale, A., Mharnur, P., Munot, P., & Math, S. (2019). Text Based Stress Detection Techniques Analysis Using Social Media. *International Conference on Computing Communication Control and Automation*. https://doi.org/10.1109/iccubea47591.2019.9129201

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-68764-y

Pearlin, L. I., Schieman, S., Fazio, E. M., & Meersman, S. C. (2005). Stress, Health, and the Life Course: Some conceptual perspectives. *Journal of Health and Social Behavior*, *46*(2), 205–219. https://doi.org/10.1177/002214650504600206

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. *Austin, TX: University of Texas at Austin*. https://repositories.lib.utexas.edu/bitstream/2152/31333/3/LIWC2015_LanguageManual.pdf

Schneiderman, N., Ironson, G., & Siegel, S. D. (2005). Stress and health: psychological, behavioral, and biological determinants. *Annual Review of Clinical Psychology*, *1*(1), 607–628. https://doi.org/10.1146/annurev.clinpsy.1.102803.144141

Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*. https://doi.org/10.18653/v1/w17-3107

Slavich, G. M., O’Donovan, A., Epel, E. S., & Kemeny, M. E. (2010). Black sheep get the blues: A psychobiological model of social rejection and depression. *Neuroscience & Biobehavioral Reviews*, *35*(1), 39–45. https://doi.org/10.1016/j.neubiorev.2010.01.003

Statista. (2023). Social networks: Statistics & Facts. Retrieved February 16, 2024, from https://www.statista.com/topics/1164/social-networks/#topicOverview

Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. *arXiv (Cornell University)*. https://arxiv.org/pdf/1911.00133.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, *105*(2), 509–521. https://doi.org/10.2466/pr0.105.2.509-521