Should AI Detect Social Stress? A Machine Learning Approach

Valentina Oelofse[1][0009−0009−2345−0497] and HMvE Combrink[1,2][0000−0001−7741−3418]

[1] Department of Economics and Finance, University of the Free State, Bloemfontein, South Africa

[2] Interdisciplinary Centre for Digital Futures, University of the Free State, Bloemfontein, South Africa

Abstract. Social media has become a primary channel for communication and public discourse, yet its high-frequency and emotionally charged nature contributes to the spread of misinformation and elevated levels of social stress. Existing misinformation detection models do not account for the psychological and social toll of this environment. This study proposes a machine learning-based Social Stress Indicator (SSI) to detect and quantify stress signals in social media conversations. Using a synthetic dataset of annotated microblogs labelled by social scientists for social stress levels, four machine learning models-Logistic Regression, Random Forest, Naive Bayes, and K-Nearest Neighbour-were trained using TF-IDF embeddings. Sentiment features from VADER and RoBERTa were also integrated. Results showed consistently high performance across models, with accuracy above 99.4% and macro F1-scores exceeding 0.994. These findings demonstrate that machine learning models can reliably detect social stress from text data. However, the overperformance may reflect dataset homogeneity, necessitating further testing on real-world, diverse social media data to confirm generalisability. Future work should expand across platforms and contexts to validate this approach for stress-aware infodemic response.

Keywords: Infodemic Intelligence · Machine Learning · Social Stress · Social Media · Computational Infodemiology.

1 Introduction

The use of social media platforms has become a daily necessity for interpersonal engagements between co-workers, family and friends [10]. Social media serves as a platform where people can share information. Although benefits for the use of social media platforms exist, social media platforms also pose the risk of disseminating misinformation [10]. Furthermore, with the increase in social media, the dissemination of misinformation rapidly increases, which can give rise to the social stress of a community [22]. Misinformation spreading is not a new phenomenon, as misinformation has started to spread since the development of the printing press [22]. The major difference is that people now have easier access to social media platforms to spread misinformation on [22]. With the increase in misinformation, it necessitates establishing a social stress detection model that will ultimately be able to assess a given phrase and link a social stress label (low, medium, high) to the given phrase. Within social networks, misinformation spreads faster and with a wider reach than factual information [3]. With the faster sensation of views and clicks, the parties spreading the misinformation end up benefiting the most, while the communities viewing the misinformation experience higher levels of stress [3, 22]. Some earlier studies described the knowledge gap between health information (which is seen as the evidence) and health misinformation (which is seen as what people think to be true), as an information epidemiology or as infodemiology [6]. Over time, the field of infodemiology research has grown and is recognised by the World Health Organisation as an emerging, novel scientific field [7]. This necessitates the importance of assessing how the infodemiology space within social media could affect the social stress of a community, which will ultimately help future research to understand the link between the economy and social media. This study aims to assist researchers in the field of infodemiology, with the intent to be able to use the social stress detection model in social media platforms like X, to be able to measure and detect high social stress. Four different machine learning algorithms were used, namely Random Forest, Logistic Regression, Naive Bayes and K-Nearest Neighbour. Each machine learning model was incorporated with Term Frequency-Inverse Document Frequency, which is a word frequency embedding. According to Zhau *et al.*, (2024), the addition of Term Frequency-Inverse Document Frequency improves the performance of the machine learning models [30, 23]. A secondary dataset was used, which is based on the social stress work done by [4]. The dataset contains text-based data resembling social media conversations. These conversations were generated by a Large Language model. Social scientists labelled each phrase with the level of social stress that the phrase holds. The level of social stress was depended on the following five factors: anxiety, negativity, engagement level, help-seeking behaviour, and misinformation content. Although the Social Stress Indicator was found to be successful for detecting low levels of stress, medium and high levels were underestimated [4]. This establishes the need to further explore methods for the use of social stress detection. This study used the expertly labelled social stress dataset from the study by [4] and trained various machine learning models on the dataset. Furthermore, the performance

of the machine learning models was evaluated by using a confusion matrix as well as evaluation metrics such as precision, accuracy, recall and F1-score.

2 Literature Review

2.1 Infodemiology

Misinformation is false, incomplete or inaccurate information that is shared without the intention of inflicting harm [27, 21]. Furthermore, this can be seen as information that is misleading or unintentionally spread without thinking that the information can inflict harm [27]. In contrast, disinformation is where false information is deliberately shared with the purpose of harming the public [27]. Disinformation is shared with the intent to deceive [21]. Not all types of disinformation have the same societal harm [16]. This emphasises the need to prioritise disinformation with the highest potential harm first [16]. An example is disinformation about politics, which can ultimately increase societal problems and erode norms [16]. Furthermore, disinformation is not only harmful to democratic institutions but also harmful to individuals' psychological well-being [16]. The rise in misinformation and disinformation can cause an absence of factual information, known as information voids or information vacuums [19]. When people see factual information infrequently, there is a lack of reliable information available amongst the pool of false information [19]. According to Palmer *et al.*, (2025), many people fall into the trap of the frequency bias, as people start to believe the information they see the most frequently [19]. This suggests that people will start believing misinformation just because there is an absence of credible and reliable information [19]. Infodemiology, also known as information epidemiology, has been widely used within health care systems to understand the distribution of health information and misinformation [6, 1]. Infodemiology guides health care professionals and the public on the quality of health care information on social media platforms [6, 1]. Infodemiology can be applied in fields such as public health monitoring as well as real-time disease surveillance [1]. The World Health Organisation acknowledged the term infodemic as the vast amount of accurate or false information that is spread [24]. Infodemiology has gained traction in recent years as a method to track disease outbreaks by analysing online data, which is also known as infoveillance [1]. Previous studies have demonstrated the use of infoveillance as a method to forecast Influenza outbreaks, map the sentiments of influenza vaccination, track cancer misinformation and Covid-19 misinformation [8, 19, 24]. Prior studies have also evolved disease surveillance into belief surveillance [1]. Belief surveillance is used in healthcare to measure the level of public belief based on health information posted on social media [1].

Although more and more health misinformation is being spread, not all users accept information equally. Many individuals and groups view information in different ways, which is known as information environments [19]. Information environments can largely be influenced by the amount and type of information present, how the information is arranged, and the social norms the individual or public believes in [19]. Belief systems are also based on the bandwagon effect,

where people are influenced by their social networks [19]. This means that people can endorse information just because everyone else endorses the information [19]. Belief systems can also be influenced by an individual's reference network, which is seen as people whom they care about [19]. In the next section, Social Stress will be discussed.

2.2 Social Stress

The psychological stress and harm due to societal pressures are more commonly known as social stress [26, 4, 29]. Social stress has become increasingly popular within the online interaction space, as social stress is more amplified and persistent online compared to face-to-face interactions [4]. Furthermore, social stress is also widely used in fields such as social psychology [4] and sociology [29, 26]. According to Wang *et al.*, (2019), mental stress is often a factor underestimated at an early stage, which evolves into severe health issues [28]. To address this, the use of social media platforms as a feasible platform to detect stress has been highlighted [15]. Moreover, a Factor Graph Model combined with a Convolutional Neural Network was used on Twitter (X) data to measure stress [15]. Additionally, machine learning was used for sentiment analysis to explore stress within social media [14]. Kumari *et al.*, (2022), found that the best performing machine learning models based on accuracy were Random Forest, Support Vector Machine and Logistic regression that incorporated Term Frequency-Inverse Document Frequency [14]. Social media is a platform where the public can express their opinions and retrieve and disseminate information [25]. With the increased use of social media, it necessitates a need to measure the public's perceptions and sentiment in real time [25].

Social media analytics can be used to measure the real-time social stress of societal pressures by making use of indicators. Various indicators have been used to measure stress and anxiety, namely Sentiment Analysis and Topic Modelling [4]. While information-seeking behavior is more commonly used to measure trends, information-seeking behavior can also be used to capture real-time social stress on social media platforms [4]. Sentiment analysis is a field of study that analyses people's opinions, emotions and appraisals towards various topics, in text-based content [18, 5, 12]. Sentiment analysis can also be referred to as subjectivity analysis, appraisal extraction as well as opinion mining [5]. Sentiment analysis is seen as the process of using text as the input and extracting valuable information about the sentiment as the output [5]. Examples of the use of sentiment analysis include emotion recognition applications that detect emotions through facial expressions, as well as performing credit ratings and entity reputation evaluations. Sentiment classification is a natural language processing task that is used to determine the sentiment or the emotional tone by classifying the text as either positive, negative or neutral [18, 12]. Various authors made use of the TextBlob Library in machine learning to be able to evaluate the emotions

in the text-based content [18]. Sentiment Analysis can be defined as (Eq. 1) [4]:

$$SA = -\sum_{I=0}^{N} \left(\frac{x_1 + x_2 + x_3 + \cdots + x_n}{N} \right), \tag{1}$$

where SA is seen as the sentiment and N is the sample size of the total number of microblogs measured, and $x_1, x_2, x_3, \ldots, x_n$ represents the sentiment per individual microblog for a specific time stamp [4].

Topic modelling is seen as a text mining technique to identify themes or topics present in a large corpus of text-based data [25]. Topic modelling is useful within the social media analytics space as it is able to identify and track relevant topics of concern [9]. According to Ji *et al.*, (2025), the Latent Dirichlet Allocation (LDA) unsupervised machine learning algorithm has become a popular method among researchers to perform topic modelling [12]. LDA assigns a topic to each word within the corpus, while using probabilistic methods to determine the frequency the word has appeared within the topics as well as the distribution of the topics within the corpus [17]. Topic modelling can be defined through the coherence score as (Eq. 2) [31]:

$$C(t_k) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \log \left(\frac{P(W_i, W_j) + \epsilon}{P(W_i) \cdot P(W_j)} \right), \tag{2}$$

where $P(W_i, W_j)$ denotes the probability of words W_i and W_j co-occurring and $P(W_i)$ and $P(W_j)$ represent the probabilities of W_i and W_j occurring respectively, and ϵ is seen as a smoothing term [31]. When individuals look for information online to gain knowledge in different topics, it is known as information-seeking behaviour [11]. Information-seeking behavior can provide useful psychological and socioeconomic insight [11]. Information-seeking behavior as an indicator can be seen as the number of people who searched for a specific topic over a certain period on a specific platform [4]. Google searches have been seen as a reliable platform to assess online interactions, which can be tracked with Google Trends [13]. Google Trends provides real-time data about a keyword's search volume within a selected timeframe and geography [13]. Information-seeking behaviour can also predict behavioural risks [11]. The equation for information-seeking behaviour can be defined as (Eq. 3) [4]:

$$\text{ISB} = \frac{\sum_{i=0}^{N} \beta}{\sum_{i=0}^{N} \overline{X}}, \tag{3}$$

where ISB is the information-seeking behaviour and β is the frequency of the same topic at a particular point in time, denoted as a percentage, which is specific to a series [4].

The Social Stress Indicator for social media was developed as a computational tool to quantify and measure social stress in real time, using sentiment analysis, subjectivity, and information-seeking behaviour [4]. The equation for the Social Stress Indicator (SSI) was defined as (Eq. 5) [4]:

$$SSI = \frac{-SA + (1 - SUB) + ISB}{3}, \tag{4}$$

where, SA refers to the sentiment, specifically focusing on the negative sentiment, SUB as the subjectivity and ISB as the information-seeking behaviour. Eq. 5 currently uses equal weightings for each component of the Social Stress Indicator. Moreover, the Social Stress Indicator was shown to capture real-time low levels of stress successfully, while the medium and high levels of stress were severely underestimated [4]. This study used the expertly labelled social stress dataset introduced in prior work and trained various machine learning models on the dataset [4].

3 Methods

This study followed a structured methodological framework comprising the design of the Social Stress Indicator (SSI), the creation of a synthetic dataset, the application of natural language preprocessing and feature engineering techniques, and the training and evaluation of four machine learning models. The methodology was designed to ensure that each stage of the process could be replicated and independently validated by other researchers.

The core of this approach is the Social Stress Indicator (SSI), which operationalises social stress as a composite measure of three text-derived indicators namely sentiment, subjectivity, and information-seeking behaviour. As defined in Eq. 5, the SSI is expressed as:

$$SSI = \frac{-SA + (1 - SUB) + ISB}{3}, \tag{5}$$

where SA denotes the sentiment score (focused on negative sentiment), SUB refers to the subjectivity score, and ISB represents the information-seeking behaviour measured as the normalised frequency of searches or interactions for a given topic over time. Each component was scaled to the $[0, 1]$ interval, and equal weighting was applied to reflect their combined influence on social stress. The negative sign for SA captures the effect of negative sentiment on the stress level, while $(1-SUB)$ inversely relates subjectivity to stress, acknowledging that more objective content may attenuate perceived stress. The SSI was computed for each microblog entry to produce a continuous score, which was then discretised into three classes (Low, Medium, High) based on thresholds co-developed with domain experts [4].

The dataset used to develop and evaluate the SSI was synthetically generated to emulate the structure and content of real-world social media posts. Prompt engineering with a large language model was employed to produce microblog-like texts covering diverse themes such as health misinformation, financial insecurity, family issues, and public crises. This ensured coverage of multiple stress-inducing contexts. Each generated post was manually annotated by three independent raters with expertise in social sciences, using a rubric derived from psychological and sociological stress frameworks. The rubric assessed five attributes: anxiety, negativity, engagement level, help-seeking behaviour, and misinformation content [4]. Each post received a stress classification label corresponding to Low, Medium, or High stress. Inter-rater agreement achieved a Cohen's κ score of 0.82, indicating strong reliability. The final dataset comprised 3,000 posts, equally balanced across the three stress levels.

Prior to feature extraction, the text was preprocessed to ensure consistency and to remove noise. All text was converted to lowercase, punctuation was stripped, and stopwords were removed using the Natural Language Toolkit (NLTK) [2]. Tokenisation and lemmatisation were subsequently applied to standardise word forms [20]. Each microblog was then represented by a high-dimensional feature vector constructed from a combination of lexical and engineered features. Term Frequency–Inverse Document Frequency (TF–IDF) embeddings were generated using scikit-learn's `TfidfVectorizer` with a maximum of 5,000 features, a bigram range of $(1, 2)$, and a minimum document frequency threshold of five. Sentiment features were extracted from two complementary sources: the VADER sentiment analysis tool, which provides lexicon-based compound sentiment scores in the range $[-1, 1]$, and the RoBERTa transformer model for contextual sentiment classification using the `cardiffnlp/twitter-roberta-base-sentiment` model. Subjectivity scores were derived from TextBlob and normalised to $[0, 1]$. Information-seeking behaviour (ISB) was calculated as the normalised frequency of topic-related searches over time, following the approach described in [4]. These features were concatenated to form a single feature matrix of dimension $(N, 5003)$, comprising 5,000 TF–IDF features and three engineered features (VADER, RoBERTa, and ISB).

Four supervised machine learning models were selected to evaluate the SSI, namely Logistic Regression, Random Forest, Multinomial Naïve Bayes, and K-Nearest Neighbour (KNN). Logistic Regression was implemented with an L2 penalty and the `lbfgs` solver in multinomial mode. Random Forest was configured with 200 trees and unrestricted depth using the Gini impurity criterion. Multinomial Naïve Bayes used Laplace smoothing with $\alpha = 1.0$, while KNN used five neighbours and Euclidean distance as the metric. Hyperparameters for each model were tuned using a five-fold grid search to optimise performance. The dataset was split into training and testing subsets using an 80/20 stratified split to preserve the balance of stress levels across subsets.

Model performance was evaluated using a comprehensive set of metrics, including accuracy, macro-averaged precision, recall, and F1-score, as well as weighted averages for each metric. Confusion matrices were generated for per-

class error analysis to assess misclassification patterns. All analyses were conducted in Python 3.10 using scikit-learn 1.3, pandas 2.0, NLTK 3.8, and Hugging Face Transformers 4.30 in Google Colab Pro with an NVIDIA T4 GPU and 12 GB RAM. This environment ensured reproducibility and scalability of the experiments for a reproducible baseline for stress-aware infodemic analytics.

4 Results and Discussion

A subset of annotated microblogs is presented in Table 1 to illustrate the scoring mechanism used in the social stress analysis. From the experiments, the model can distinguish between varying degrees of emotional content. For instance, positive or neutral expressions such as "Life is smooth and simple." received a low score of 0.20, indicating minimal psychosocial strain. Conversely, posts conveying acute psychological burden, such as "This is too much the pressure of expectations.", yielded scores exceeding 0.90, reflecting high-intensity stress signals. Of particular interest are the intermediate cases, such as "Staying hopeful about family issues." (0.32), where the sentiment is cautiously optimistic yet still linked to an underlying stressor. This is particularly relevant for domains such as mental health monitoring, crisis intervention, and infodemic management, where understanding the spectrum of stress expressions can inform timely and proportionate responses.

Table 1. Subset of annotated microblogs

Microblog	Social Stress Score
Life is smooth and simple.	0.20
This is too much the pressure of expectations.	0.92
Staying hopeful about family issues.	0.32
Losing sleep over family issues.	0.88

Table 2 illustrates the tokenized representation of microblog content, paired with their normalized SSS and corresponding stress classification labels. This representation highlights the transition from raw narrative text to structured analytical units suitable for computational modelling. Stress labels are assigned according to pre-defined thresholds with 0 for low or negligible stress, 1 for moderate stress, and 2 for high stress intensity. For example, the token sequence [life, smooth, simple] reflects a low-stress state (0.20, label 0), while [losing, sleep, family, issues] reveals a high-stress profile (0.88, label 2) associated with family-related anxieties. Tokens such as [staying, hopeful, family, issues], with a moderate score of 0.32 (label 1), demonstrate that models can capture nuanced emotional states where positive sentiment coexists with underlying social stressors, but this is specific to the data.

Table 2. Microblog Tokens with SSS Normalized Scores and Stress Labels

Microblog Tokens	SSS Normalized	Stress Label
`[life, smooth, simple]`	0.20	0
`[staying, hopeful, family, issues]`	0.32	1
`[much, pressure, expectations]`	0.92	2
`[losing, sleep, family, issues]`	0.88	2

Table 3 presents tokenized microblogs enriched with both stress labels and sentiment polarity scores generated from two distinct natural language processing frameworks, such as VADER and RoBERTa. The inclusion of sentiment analysis alongside stress classification enables a multidimensional interpretation of each microblog, capturing not only the presence of stress but also its affective orientation. For instance, the token sequence [life smooth simple], labelled as low stress (0), is consistently identified as predominantly positive by both models, with RoBERTa attributing a notably high positive score (0.79) compared to VADER's binary-neutral output (1.00 in the neutral category). In contrast, a high-stress sequence such as [losing sleep family issues] received overwhelmingly negative sentiment scores from both frameworks, particularly RoBERTa (0.87), highlighting its heightened sensitivity to emotionally charged language. Intermediate stress cases, such as [staying hopeful family issues] (label 1), exhibit mixed sentiment patterns, where positive sentiment persists despite the presence of stress-related terms, demonstrating the models' ability to detect nuanced emotional states.

Table 3. Microblog Tokens with Stress Labels and Grouped Sentiment Scores

Microblog Tokens	Stress Label	VADER			RoBERTa		
		Neg	Neu	Pos	Neg	Neu	Pos
`[life smooth simple]`	0	0.00	1.00	0.00	0.01	0.20	0.79
`[staying hopeful family issues]`	1	0.00	0.55	0.45	0.01	0.32	0.67
`[much pressure expectations]`	2	0.24	0.76	0.00	0.75	0.23	0.02
`[losing sleep family issues]`	2	0.39	0.61	0.00	0.87	0.12	0.01

Across metrics, Logistic Regression, Random Forest, and Naive Bayes exhibit near-identical, near-ceiling results (Accuracy ≈ 0.9965, Macro/Weighted $F_1 \in [0.9963, 0.9965]$), while KNN trails marginally (Accuracy $= 0.9947$; Macro $F_1 = 0.9947$), Table 4. The close correspondence between macro- and weighted-averaged scores suggests limited class imbalance in the evaluation set and indicates that gains are not driven solely by majority classes. The parity among

three distinct model families (linear, probabilistic, and ensemble) implies that the TF–IDF feature space already affords highly separable decision boundaries, rendering additional model complexity comparatively unimportant. While these results demonstrate excellent in-sample generalisation under the current protocol, such uniformly high scores warrant caution: they may reflect dataset homogeneity, lexical cues that trivially correlate with labels, or residual leakage between train and test partitions. We therefore recommend complementary diagnostics (per-class confusion matrices, calibrated probability assessment, and error analysis of false positives/negatives) and validation checks (stratified cross-domain validation, temporal splits, and perturbation tests) to confirm that the observed performance translates beyond the present sample and to identify any over-reliance on spurious lexical markers.

Table 4. Classification performance comparison across models using TF–IDF features.

Metric	**Logistic Regression**	**Random Forest**	**Naive Bayes**	**KNN**
Accuracy	0.9965	0.9965	0.9965	0.9947
Macro Avg Precision	0.9969	0.9969	0.9969	0.9954
Macro Avg Recall	0.9957	0.9957	0.9957	0.9940
Macro Avg F1-score	0.9963	0.9963	0.9963	0.9947
Weighted Avg Precision	0.9965	0.9965	0.9965	0.9948
Weighted Avg Recall	0.9965	0.9965	0.9965	0.9947
Weighted Avg F1-score	0.9965	0.9965	0.9965	0.9947

Table 5 presents the classification performance of four machine learning models trained with two different composite feature sets: *TF–IDF + RoBERTa* and *TF–IDF + VADER*. Across all metrics and both feature configurations, performance is uniformly high, with accuracy values exceeding 0.994 and negligible variation between models.

Table 5. Performance of ML models with different feature sets.

Metric	**TF-IDF + RoBERTa**				**TF-IDF + VADER**			
	Logistic Regression	Random Forest	Naive Bayes	KNN	Logistic Regression	Random Forest	Naive Bayes	KNN
Accuracy	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9941
Macro Avg Precision	0.9957	0.9957	0.9957	0.9947	0.9957	0.9957	0.9957	0.9947
Macro Avg Recall	0.9947	0.9947	0.9947	0.9936	0.9947	0.9947	0.9947	0.9936
Macro Avg F1-score	0.9952	0.9952	0.9952	0.9941	0.9952	0.9952	0.9952	0.9941
Weighted Avg Precision	0.9953	0.9953	0.9953	0.9942	0.9953	0.9953	0.9953	0.9942
Weighted Avg Recall	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9941
Weighted Avg F1-score	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9941

Logistic Regression, Random Forest, and Naive Bayes exhibit identical scores for each metric within a given feature set, while KNN lags marginally, particularly in accuracy and macro-averaged metrics, by approximately 0.001–0.002. No-

tably, the integration of sentiment analysis features from RoBERTa and VADER does not appear to yield differential performance improvements, as the two composite feature sets produce identical values to four decimal places for each model. This suggests that, within the constraints of the present dataset, sentiment-derived features may be redundant when combined with high-dimensional lexical representations from TF-IDF. The homogeneity of results across diverse model architectures further implies that the task is linearly separable in the transformed feature space, and that model choice exerts minimal influence on predictive accuracy. While these findings reflect strong in-sample performance, they also raise concerns regarding potential overfitting, data leakage, or limited variability in the evaluation set. To ensure external validity, future work should investigate the stability of these results under domain-shift conditions, perform stratified temporal validation, and conduct feature importance analyses to assess the relative contribution of sentiment-based features in broader, more heterogeneous corpora. By training on more data, the models can learn how to perform better with unknown samples. Given that this was a relatively small sample size, it is suggested that more data be introduced for social stress classification, as the ML approach to a SSS can be applied. While this study demonstrates promising results using machine learning models trained on a synthetically generated dataset, it is essential to acknowledge the implications of using such data on performance evaluation. The synthetic microblogs were generated using a large language model and annotated by domain experts, allowing control over topic balance and stress-level representation. However, synthetic data may not fully capture the complexity, noise, ambiguity, or adversarial phrasing typical of real-world social media content. This could result in inflated performance metrics due to reduced linguistic variability and more distinct decision boundaries between classes. The near-ceiling accuracy and F1-scores observed across all models suggest that the dataset may be overly homogeneous or artificially well-separated, which limits the generalisability of these results.

Real-world social media posts are often contextually nuanced, highly informal, and influenced by sarcasm, idioms, or multimedia context, which synthetic data might fail to emulate. Additionally, platform-specific trends, evolving language, and demographic markers embedded in real conversations are critical components for building robust social stress detection systems. Due to privacy concerns, ethical data collection restrictions, and annotation challenges related to sensitive psychological content, real-world labelled datasets for social stress remain scarce. Therefore, synthetic data was employed as an initial proof of concept to establish baseline model efficacy while mitigating topic and label bias. Therefore, future work should involve external validation on real-world datasets and explore domain adaptation methods to transition from synthetic to naturalistic data environments, ensuring that these models are ethically deployable in high-stakes settings such as mental health surveillance or infodemic early warning systems.

5 Conclusion

This study was set out to evaluate the capacity of multiple machine learning models, namely Logistic Regression, Random Forest, Naive Bayes, and KNN, to predict the SSS. Results from this study demonstrate that all evaluated machine learning models, when applied with either TF-IDF + RoBERTa or TF-IDF + VADER feature sets, achieved near-identical and exceptionally high performance metrics across accuracy, precision, recall, and F1-scores. Specifically, Logistic Regression, Random Forest, and Naive Bayes consistently recorded accuracy values of 0.9953, while KNN marginally trailed with an accuracy of 0.9941. The uniformity of macro and weighted averages across models and feature combinations indicates that the classification task posed minimal challenge to the algorithms, suggesting a potential ceiling effect caused by the dataset's structure or class distribution. While such high scores may initially appear to reflect excellent model generalisation, they also raise concerns about the underlying data complexity and diversity. In tasks with limited variance or highly separable classes, inflated metrics can obscure a model's ability to generalise to unseen, real-world scenarios. The negligible difference between the RoBERTa and VADER sentiment integration further suggests that the choice of sentiment extraction method had little influence on predictive performance within this dataset. This convergence in outcomes underscores the importance of evaluating not only statistical performance but also the representativeness and challenge level of the training data. Without a dataset that captures the variability and subtlety of the problem space, even state-of-the-art models may deliver impressive yet potentially misleading results, thereby limiting their applicability in practical, high-stakes decision-making contexts. Therefore, should we trust AI to detect social stress? In the context of this study, more experimentation is required to implement such models in a real-world scenario, but based on this experiment, if the models are trained correctly, there is promise and potential for the implementation of such models.

References

1. Al-garadi, M.A., Khan, M.S., Varathan, K.D., Mujtaba, G., Al-Kabsi, A.M.: Using online social networks to track a pandemic: A systematic review (8 2016). https://doi.org/10.1016/j.jbi.2016.05.005
2. Asad, N.A., Pranto, M.A.M., Afreen, S., Islam, M.M.: Depression detection by analyzing social media posts of user. In: 2019 IEEE International Conference on Signal Processing, Information, Communication and Systems, SPICSCON 2019. pp. 13–17. Institute of Electrical and Electronics Engineers Inc. (11 2019). https://doi.org/10.1109/SPICSCON48833.2019.9065101
3. Barnabò, G., Siciliano, F., Castillo, C., Leonardi, S., Nakov, P., Martino, G.D.S., Silvestri, F.: Deep active learning for misinformation detection using geometric deep learning. Online Social Networks and Media **33** (1 2023). https://doi.org/10.1016/j.osnem.2023.100244
4. Combrink, H.: The social stress indicator (ssi). PloS one **20**(9), e0328768 (2025)

5. Dervenis, C., Kanakis, G., Fitsilis, P.: Sentiment analysis of student feedback: A comparative study employing lexicon and machine learning techniques. Studies in Educational Evaluation **83** (12 2024). https://doi.org/10.1016/j.stueduc.2024.101406
6. Eysenbach, G.: Infodemiology: The epidemiology of (mis)information. Tech. rep. (12 2002)
7. Eysenbach, G.: How to fight an infodemic: The four pillars of infodemic management (6 2020). https://doi.org/10.2196/21820
8. Gao, J., Zhang, Y.C., Zhou, T.: Computational socioeconomics (7 2019). https://doi.org/10.1016/j.physrep.2019.05.002
9. Ilyas, B., Sharifi, A.: A systematic review of social media-based sentiment analysis in disaster risk management (6 2025). https://doi.org/10.1016/j.ijdrr.2025.105487
10. Indu, V., Thampi, S.M.: Misinformation detection in social networks using emotion analysis and user behavior analysis. Pattern Recognition Letters **182**, 60–66 (6 2024). https://doi.org/10.1016/j.patrec.2024.04.007
11. Jaidka, K., Eichstaedt, J., Giorgi, S., Schwartz, H.A., Ungar, L.H.: Information-seeking vs. sharing: Which explains regional health? an analysis of google search and twitter trends. Telematics and Informatics **59** (6 2021). https://doi.org/10.1016/j.tele.2020.101540
12. Ji, J., Xu, X., Tam, V.W., Zhang, Y.: Revealing public attitudes toward 'substituting plastic with bamboo' in china: Sentiment and topic analyses using social media data. Forest Policy and Economics **176** (7 2025). https://doi.org/10.1016/j.forpol.2025.103508
13. Kohlmann, S., Stielow, L., Löwe, B.: Did online information seeking for depression increase during covid-19 lockdown times? a google trend analysis on data from germany and the uk. Journal of Affective Disorders Reports **13** (7 2023). https://doi.org/10.1016/j.jadr.2023.100587
14. Kumari, K., Das, S.: Stress detection system using natural language processing and machine learning techniques. Tech. rep. (2022), https://iiitranchi.ac.in/
15. Lin, H., Jia, J., Qiu, J., Zhang, Y., Shen, G., Xie, L., Tang, J., Feng, L., Chua, T.S.: Detecting stress based on social interactions in social networks. IEEE Transactions on Knowledge and Data Engineering **29**, 1820–1833 (9 2017). https://doi.org/10.1109/TKDE.2017.2686382
16. Lukavská, K., Gabrhelík, R., Miovský, M., Hynek, N., Gavurová, B., Šťastná, L., Barták, M., Petruželka, B., Moravec, V.: Exploring disinformation: The interplay of exposure, trust, and sharing. Computers in Human Behavior Reports **18** (5 2025). https://doi.org/10.1016/j.chbr.2025.100686
17. M., K.R.P., D., J.S.: Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques. In: Materials Today: Proceedings. vol. 51, pp. 576–584. Elsevier Ltd (2021). https://doi.org/10.1016/j.matpr.2021.06.001
18. Madan, P.M., Madan, M.R., Thakur, D.P.: Analysing the patient sentiments in healthcare domain using machine learning. In: Procedia Computer Science. vol. 238, pp. 683–690. Elsevier B.V. (2024). https://doi.org/10.1016/j.procs.2024.06.077
19. Palmer, A., Gorman, S.: Misinformation, trust, and health: The case for information environment as a major independent social determinant of health. Social Science and Medicine **381** (9 2025). https://doi.org/10.1016/j.socscimed.2025.118272
20. Raj, S., Vishnoi, A., Srivastava, A.: Classify alzheimer genes association using naïve bayes algorithm. Human Gene **41** (9 2024). https://doi.org/10.1016/j.humgen.2024.201309

21. Roy, M., DeRoche, M.: Mapping misinformation gatekeepers in non-western contexts: A computational network analysis of fake news on x (twitter) using gephi. Telematics and Informatics Reports **18** (6 2025). https://doi.org/10.1016/j.teler.2025.100214
22. Roy, P.K., Tripathy, A.K., Weng, T.H., Li, K.C.: Securing social platform from misinformation using deep learning. Computer Standards and Interfaces **84** (3 2023). https://doi.org/10.1016/j.csi.2022.103674
23. Semary, N.A., Ahmed, W., Amin, K., Pławiak, P., Hammad, M.: Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLoS ONE **19** (2 2024). https://doi.org/10.1371/journal.pone.0294968
24. Shanker, A., Vlaev, I.: The social influence of the corrections of vaccine misinformation on social media. Vaccine **56** (5 2025). https://doi.org/10.1016/j.vaccine.2025.127177
25. Sun, S., Guo, Z., Li, L., Zheng, Z., Wang, J., Anno, S., Qian, X.: Decoding public sentiments on energy transition in japan's three major metropolitan areas: Social media analysis using machine learning. Journal of Cleaner Production **495** (3 2025). https://doi.org/10.1016/j.jclepro.2025.145038
26. Turner, R.J., Wheaton, B., Lloyd, D.A.: The epidemiology of social stress. Tech. rep. (1995)
27. Wachs, S., Kops, M., Mateos-Pérez, E., Gámez-Guadix, M.: Counteracting disinformation among young people. psychometric properties of the disinformation bystander intervention model scale, demographic differences, and associations with empathy. Computers in Human Behavior Reports **18** (5 2025). https://doi.org/10.1016/j.chbr.2025.100671
28. Wang, F., Wang, Y., Wang, J., Xiong, H., Zhao, J., Zhang, D.: Assessing mental stress based on smartphone sensing data: An empirical study. Tech. rep. (2019), https://www.researchgate.net/publication/336642084
29. Wheaton, B., Young, M., Montazer, S., Stuart-Lahman, K.: Social Stress in the Twenty-First Century, pp. 299–323. Springer Science and Business Media B.V. (2013). https://doi.org/10.1007/978-94-007-4276-5_15
30. Zhou, J., Ye, Z., Zhang, S., Geng, Z., Han, N., Yang, T.: Investigating response behavior through tf-idf and word2vec text analysis: A case study of pisa 2012 problem-solving process data. Heliyon **10** (8 2024). https://doi.org/10.1016/j.heliyon.2024.e35945
31. Çelikten, T., Onan, A.: Topic modeling through rank-based aggregation and llms: An approach for ai and human-generated scientific texts. Knowledge-Based Systems **314** (4 2025). https://doi.org/10.1016/j.knosys.2025.113219