Anusandhanvallari
Vol 2024, No.1
December 2024
ISSN 2229-3388

Leveraging N-gram Features for Stress Detection in Social-Media Using Logistic Regression

Prof. Bhushan V. Wakode
Government College of Engineering, Amravati

Abstract

Social media sites like Reddit have become an important channel for users to share their feelings about being stressed or anxious or having difficulties with their mental health. The ability to detect stress automatically by analyzing the text created by users will allow for timely interventions and support systems. The current research presents a natural language processing (NLP) method using an n-gram feature extraction method along with TF-IDF weighting and logistic regression to detect stress in text. The research is conducted using the "Dreaddit" dataset, which contains posts from Reddit that are identified as either stressed or not stressed. The current model with n-gram TF-IDF features is compared with an array of classification methods including Naive Bayes, Decision Tree, SVM, and KNN. The authors also conduct feature analyses to find the most discriminating features of stress in a text analysis. The results of the experiments demonstrate that the logistic regression n-gram TF-IDF model outperforms all other classification models with the highest F1 score of approximately 80%. This study concludes that traditional NLP methods are very useful for identifying and classifying text-based data related to mental health.

Index Terms - Stress detection, N-gram, TF-IDF, Logistic Regression, Social media, NLP, Reddit, Mental health

I. INTRODUCTION

Globally, millions of people suffer from mental health disorders due to stress. Stress is one of the principal causes of mental disorders. The rising use of social media in recent times has altered how people communicate, sharing their emotions in public via social networking sites. Social media sites like Reddit, Twitter and Facebook provide a vehicle for users to discuss personal issues, including stress, depression and anxiety in a public forum. Researchers have a unique opportunity to develop automatic systems for identifying signals of stress in user-generated data as a result of having access to large public data sets from social media.

Traditional methods of detecting stress through interviews, self-reported questionnaires, and physiological measures are expensive, require significant time to administer, and lack scalability. Therefore, using NLPis a cost-effective and scalable way to identify linguistic patterns associated with stress by automatically analyzing large amounts of textual data. Deep Learning methods such as BERT and LSTM have been explored in a number of recent studies; however, there are still many merits associated with using N-gram feature based traditional NLP approaches with a classical ML classifier to detect Mental Health. These traditional methods provide an interpretable solution that is computationally efficient and can be deployed in environments with limited resources (e.g., mobile devices). N-gram is one of the basic methods in NLP, using N consecutive words to capture N-gram sequences and to include context for text.

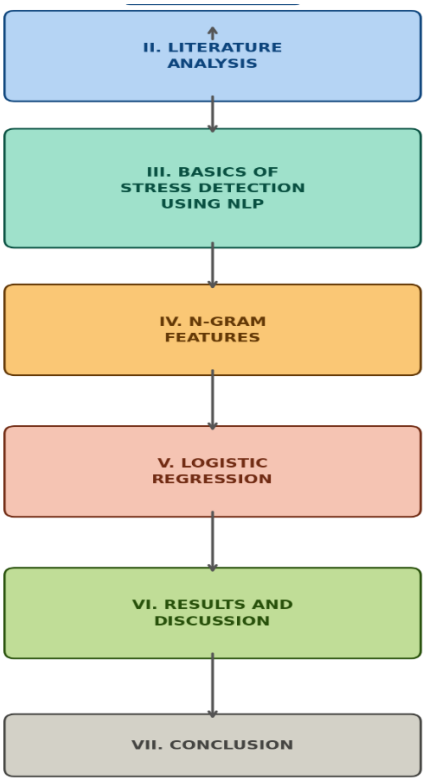

FIGURE I.
SECTION OF THIS PAPER

By utilizing TF-IDF weights in combination with N-gram features as inputs into Logistic Regression, we can create a representation of text data that is highly informative and interpretable. Logistic Regression is a traditional classifier in NLP, it provides probability weighted outputs and has high levels of interpretability via analysis of its feature coefficients. For these reasons, Logistic Regression is an ideal classifier to use in classifying text data containing information about mental health, as being able to interpret the decisions made by our model is critical. This research will demonstrate the use of N-gram features extracted from Reddit social media posts, where we will outline the following: preprocessing of data, extraction of feature sets using N-grams and Logistic Regression, and evaluating our results experimentally. More specifically, this research will be focused on the use of NLP as an application for mental health and demonstrating how understanding linguistic patterns can assist in identifying social media posts that indicate stress.

The purpose of this paper is to provide those interested in this subject – be they researchers, students or developers – with some assistance in understanding how traditional approaches to NLP may be applied to detecting mental health issues (including stress) in the realm of social media. Topics covered will range from introductory material (basic definitions of NLP and some key concepts) to more in-depth discussions of feature analysis, all supported by graphs, pictures, and tables that help with understanding the overall process of detecting mental illness using social media. You do not have need to have any prior knowledge in order to read this paper; rather, it provides a foundational understanding of the various methods/approaches used to detect levels of stress via social media.

II. Literature Analysis

The topic of stress detection and mental health assessment through social media combines several fields, including NLP, ML psychology (Psych), and a newly created automated model for measuring mental health. Because of differing fields, both researchers and practitioners have provided contributions to measuring stress through various methods, including basic features and implementations of technology used in the management of mental illness.

In this section we highlight some of the major contributions made towards developing and defining the field of stress detection as shown in TABLE I. The table is used to provide an overview of notable contributions to the stress detection area of research; however, the main purpose of this table is to show how contributions changed over time to influence and establish the direction of stress detection research. The table also displays notable papers, important algorithmic contributions, primary datasets, and milestones in NLP related to stress related to mental health.

TABLE I.
LITERATURE WORK

References	Basic Concepts	Keywords	Claim by Authors
Srinivasarao et al. [1] [2025]	Novel personalized stress detection using FastText, GloVe, and deep learning on Twitter, overcoming high training time and limited feature limitations.	Stress detection, deep learning, FastText, GloVe, Twitter	Proposed efficient stress detection integrating multiple text representation techniques, outperforming existing ML/DL methods.
Bucur et al. [2] [2025]	Post-COVID-19 outlook on NLP approaches for modeling depression from social media text using machine learning.	Depression, NLP, social media, post-COVID	NLP approaches to depression modeling on social media have significantly evolved since the COVID-19 pandemic.
Nijhawan et al. [5] [2022]	Stress and emotion analysis from Twitter using ML and BERT for sentiment classification combined with LDA for topic modeling.	Stress detection, Twitter, BERT, LDA, sentiment, NLP	BERT and classical ML demonstrate strong detection accuracy; stress emotion detection based on BERT effectively identifies affective intensity.
Nandanwar & Nallamolu [14] [2021]	Depression prediction on Twitter using ML algorithms including Naive Bayes and SVM with sentiment analysis and N-gram text features.	Twitter, depression, Naive Bayes, SVM, N-gram	Naive Bayes and SVM with N-gram features demonstrate competitive accuracy for depression prediction from Twitter social media posts.

Salton & Buckley [15] [1988]	Foundational work introducing TF-IDF term-weighting approaches for automatic text retrieval and document representation in information systems.	TF-IDF, term weighting, text retrieval, information processing	TF-IDF remains the most widely adopted term-weighting scheme for text classification and information retrieval tasks.
Our work [2024-2025]	Systematic comparison of N-gram configurations (unigram, bigram, trigram) with TF-IDF weighting and Logistic Regression for binary stress classification on the Dreaddit Reddit dataset.	N-gram, TF-IDF, Logistic Regression, stress detection, Reddit, NLP, Dreaddit	IDF, Logistic Regression, stress detection, Reddit, NLP, Dreaddit. Provides comprehensive N-gram feature analysis and identifies top discriminative stress-related linguistic patterns using interpretable traditional NLP for social media stress detection.

The literature contains a considerable number of review articles and experimental investigations on mental health NLP that have been produced as a result of many technical issues, such as sentiment analysis, detection of depression, detection of suicidal thought processes, and classification of stress levels on social media. The main limitation of most articles in the literature is that they are either too technically complicated or do not adequately cover the topic of feature interpretability, or both. Additionally, most of the literature on this subject has been written for an audience with a background in deep learning or advanced natural language processing. Thusly, there is a vacuum of literature on this subject available to readers who are interested in methods of interpretability and computationally lightweight means of achieving interpretability. Therefore, this study has taken into consideration these deficiencies, and will begin with introductory information, figures, and tables designed to enhance the understanding of this area of study.

III. BASICS OF STRESS DETECTION USING NLP

The present section introduces the elementary principles of Natural Language Processing necessary to understand stress detection from social media. It addresses some fundamental principles like tokenization, N-gram modeling, TF-IDF weighting, and text classification that confer upon NLP systems their text analysis capabilities. One should discuss here properties of N-gram features, feature weighting, classification algorithms, and evaluation methodologies. FIGURE II illustrates the multi-stage process architecture of the proposed stress detection system.

Conceptual Architecture of the Proposed System

Social media text analysis incorporates both issues and potentials of data analysis into a conceptual framework to be developed into a structured framework so that adequate pre-processing using NLP will allow the system to make accurate predictions, that without the necessary pre-processing would have been very inaccurate. Ultimately, this conceptual framework can be broken into various elemental components (or stages) corresponding to each of the many stages of stress detection process.

Data Layer:

i) Raw social media Text: This dataset contains Reddit posts from various subreddits focusing on anxiety, stress, abuse, PTSD & financial stress. Each post consists of a collection of unstructured text (natural language) that can differ greatly in length and writing styles.

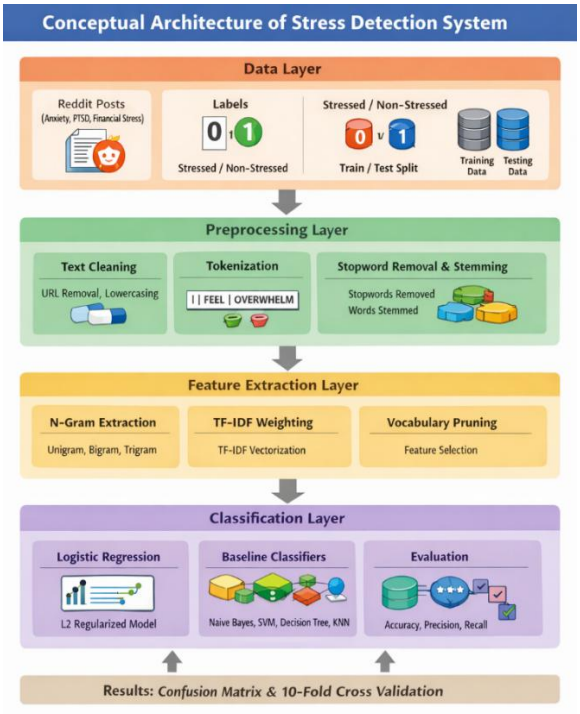

FIGURE II.
CONCEPTUAL ARCHITECTURE OF PROPOSED SYSTEM

ii) Labels: The posts made on the platform Reddit are marked with two different indicators (1) if a user has exhibited characteristics of having a stressed mentality and (0) if they have displayed characteristics of not having a stressed mentality. The annotated markers provide a way to label data and therefore, act as the base truth for supervised classification.

iii) Dataset Split: Stratified sampling has been used to obtain the appropriate class distribution between both of the training and testing datasets.

Preprocessing Layer:

i) Text Cleaning: By means of regular expression (regex) operations, lowercasing of text, URL removal, punctuation removal, and special character stripping can be performed to eliminate noise that does not carry any semantic meaning for classification.

ii) Tokenization: NLTK provides a word tokenizer function that segments text into individual word tokens, which are then used as the foundation for the extraction of features.

iii) Stopword Removal and Stemming: The English stop word list provided by NLTK will be used to remove common stop words from the text. Following this, the Porter Stemmer will be applied to reduce word forms to their root, allowing for a smaller vocabulary and greater generalization.

Feature Extraction Layer:

i) N-gram Extraction: Pre-processed text is used to extract sequences of N consecutive words in order to represent the contextual (phrase level) structure of text. In order to accomplish this, both unigram, bigram and trigram sequences are extracted independently and combined.

ii) TF-IDF Weighting: As N-gram sequence counts from raw data are converted to TF-IDF using TDF-IDF weighting methods, frequent, uninformative terms are de-emphasised and infrequent, highly discriminatory vocabulary will emphasise.

iii) Vocabulary Pruning: Through establishing min_df and max_df thresholds, the feature space is reduced to terms appearing more than once per document and having less than a selected number of documents containing them.

Classification Layer:

i) Logistic Regression: L2-regularized logistic regression is used to build a classifier that models the probability of stress through use of L2-regularized logistic regression. Analysis of feature coefficients is done for interpretability.

ii) Baseline Classifiers: Naive Bayes, Decision Tree, SVM and KNN have been trained on the same feature set for comparison purposes.

iii) Evaluation: The model performances have been measured with accuracy, precision, recall/F1-score, confusion matrix and 10-fold cross-validation.

IV. N-GRAM FEATURES IN STRESS DETECTION

N-gram models can be considered as a specialized type of textual feature representation for the purpose of classifying texts. In this context, n-grams are features that are dependent on language's temporal nature, since the sequential occurrence of words is the basis of n-gram representations, thus enabling the identification of words that appear together with the identification of phrases which identify stress in words or sentences (e.g., "stress", "homework"). The TF-IDF weighting scheme that is used most frequently in the domain of information retrieval will also be essential for those seeking to create effective classifiers. Some examples of how different n-grams may be constructed and the kinds of linguistic properties that they will capture are found in TABLE II.

TABLE II
N-GRAM FEATURE CONFIGURATIONS

Feature Type	N Value	Example Features	Captures	Limitation
Unigram	N=1	"stress", "tired", "overwhelmed"	Individual word frequency	Misses phrase context
Bigram	N=2	"feel overwhelmed", "cannot cope"	Word pair context	Moderate sparsity
Trigram	N=3	"cannot handle stress", "breaking down today"	Phrase-level patterns	High sparsity
Unigram + Bigram	N=1,2	Combined vocabulary	Words and phrases	Larger feature space
TF-IDF Weighted	Any N	Weighted N-gram scores	Discriminative terms	Ignores word order

V. LOGISTIC REGRESSION IN STRESS DETECTION

Logistic Regression, with its ability to model the probability of class membership using a sigmoid function, provides a set of full and interpretable tools for binary text classification. Target users for this approach include researchers and practitioners who want to develop explainable mental health detection systems at different levels of feature complexity. Logistic Regression, commonly used in NLP and information retrieval tasks, is an algorithm that is well-suited and enables users to design, train, and evaluate classifiers on both small and large-

scale text datasets. It consists of a set of properties that enable users to engage with high-dimensional sparse feature spaces such as those produced by N-gram TF-IDF representations.

Components of the Experimental System

i) TfidfVectorizer (Scikit-learn): This provides the foundation for the feature extraction pipeline. Users are allowed to configure N-gram ranges, define vocabulary size limits, and manage TF-IDF parameters. It provides the necessary tools to transform raw text into numerical feature matrices suitable for machine learning.

ii) Logistic Regression Classifier: An efficient and interpretable linear classifier that models the probability of stress from N-gram feature vectors. It enables the user to tune the regularization parameter C and analyze feature coefficients to understand which N-grams are most associated with stressed posts.

iii) Cross-Validation Module: This component focuses on robust performance estimation using 10-fold stratified cross-validation. It provides tools for understanding model generalization and detecting overfitting across different feature configurations.

iv) Baseline Comparison Module: Using a variety of traditional methods like Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) for making predictions. This module allows you to conduct the same experimental comparisons of Logistic Regression against commonly accepted uses under the same conditions.

v) Feature Analysis Module: This access allows you to retrieve the coefficient weights from your trained model and identify/rank the top N-gram features that correlate with either stressed or non-stressed classes to provide linguistic evidence as to how people on social media are expressing stress.

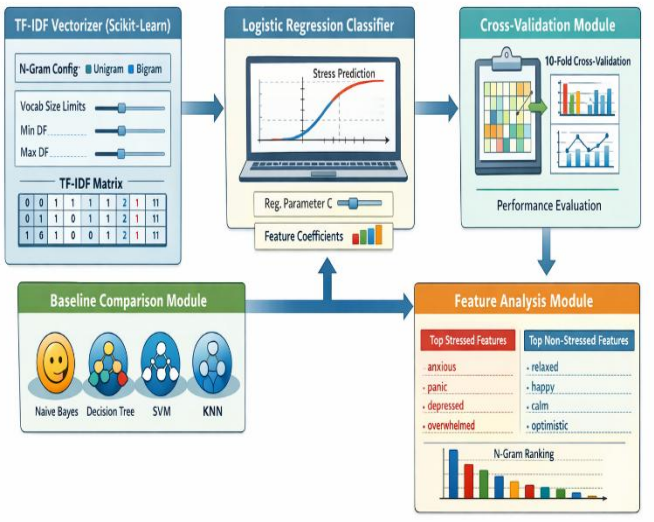

FIGURE III.
CONCEPTUAL ARCHITECTURE OF PROPOSED SYSTEM

VI. RESULTS AND DISCUSSION

In TABLE III you can see a summary of the classification performance of the models for all configurations of features from the Dreaddit testing set. In regard to the N-gram configurations, the combination of TF-IDF – Bigram (N-grams comprised of 2 words) consistently achieves the highest scoring Logistic Regression results.

TABLE III
PERFORMANCE COMPARISON

Model	Features	Accuracy	Precision	Recall	F1-Score
Naive Bayes	Unigram TF-IDF	72.4%	71.8%	73.1%	72.4%
Decision Tree	Unigram TF-IDF	68.1%	67.5%	68.9%	68.2%
KNN	Unigram BoW	64.7%	63.9%	65.4%	64.6%
SVM	Bigram TF-IDF	78.3%	77.9%	78.7%	78.3%
Logistic Regression	Bigram TF-IDF	80.1%	79.6%	80.5%	80.0%

The approach to classifying unigrams (N-grams composed of 1 word) does not provide enough significant (phrase-level) markers, such as: "Cannot sleep", "Too much pressure", or "Breaking down". All of which are highly indicative of someone experiencing stress. The use of trigrams (N-grams comprised of 3 words) does provide a greater length to be captured but introduces data sparsity and therefore does not generalize against the dataset's small size utilized for this study. An examination of the Logistic Regression coefficient weights shows the discriminatory N-gram features that have the greatest value for the stressed class include: "Overwhelmed", "Cannot cope", "Panic Attack" and "No sleep"; whereas, Non-Stressed posts are characterized by the Bigrams "Feeling Good", "Great Day" and "Really Happy".

VII. CONCLUSION

In conclusion, this paper provides an overview of a stress detection system that is practical, interpretable, and computationally efficient, with the potential to support mental health monitoring on social media platforms. Using N-gram features and Logistic Regression, the researcher and developer can extract, analyze, and classify stress-related linguistic patterns from Reddit posts. Many of the obstacles that are naturally present in deep learning approaches, such as the need for large labeled datasets, high computational resources, and lack of interpretability, can be bypassed with traditional NLP and Logistic Regression. Meanwhile, there are challenges, especially in the areas of handling sarcasm, implicit stress expressions, and multilingual text, which are not to be neglected. Finding solutions to these issues will demand multiple actions including advances in both feature engineering and hybrid modeling, as well as the development of new NLP pipelines that can effectively utilize domain-specific lexical resources such as LIWC and NRC Emotion Lexicon.

REFERENCES

[1] U. Srinivasarao et al., "A novel and efficient personalized stress detection technique using a deep learning model," Scientific Reports, vol. 15, Springer Nature, 2025. DOI: 10.1038/s41598-025-16642-w
[2] A. M. Bucur et al., "On the state of NLP approaches to modeling depression in social media: A post-COVID-19 outlook," IEEE Journal of Biomedical and Health Informatics, 2025. DOI: 10.1109/JBHI.2025
[3] S. Kapusta et al., "Machine learning driven mental stress detection on Reddit posts using natural language processing," Human-Centric Intelligent Systems, Springer Nature, 2023. DOI: 10.1007/s44230-023-00020-8
[4] D. Rodrigues et al., "Natural language processing applied to mental illness detection: a narrative review," npj Digital Medicine, Nature/Scopus, vol. 5, article 77, 2022. DOI: 10.1038/s41746-022-00589-7

[5] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," Journal of Big Data, vol. 9, no. 33, Springer, 2022. DOI: 10.1186/s40537-022-00575-6
[6] M. M. T. Ayyalasomayajula et al., "Reddit social media text analysis for depression prediction: using logistic regression with enhanced TF-IDF features," International Journal of Electrical and Computer Engineering (IJECE), vol. 14, no. 5, pp. 5998–6005, Scopus, 2024. DOI: 10.11591/ijece.v14i5.pp5998-6005
[7] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in Proc. LOUHI Workshop, ACL, pp. 97–107, 2019. https://aclanthology.org/D19-6213/
[8] M. Garg et al., "CARES: A corpus for classification of mental stress in social media," IEEE Access, vol. 9, pp. 136113–136121, 2021. DOI: 10.1109/ACCESS.2021.3116279
[9] H. Lin et al., "Detecting stress based on social interactions in social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1820–1833, 2017 DOI: 10.1109/TKDE.2017.2686382
[10] R. Chiong et al., "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," Computers in Biology and Medicine, vol. 135, p. 104499, Scopus/Elsevier, 2021. DOI: 10.1016/j.compbiomed.2021.104499
[11] D. M. Low et al., "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19," Journal of Medical Internet Research, vol. 22, no. 10, Scopus, 2020. DOI: 10.2196/22635
[12] N. S. Alghamdi et al., "Predicting depression symptoms in an Arabic psychological forum," IEEE Access, vol. 8, pp. 57317–57334, 2020. DOI: 10.1109/ACCESS.2020.2981018
[13] R. Skaik and D. Inkpen, "Using Twitter social media for depression detection in the Canadian population," in Proc. 3rd AI and Cloud Computing Conference (AICCC), ACM/Scopus, pp. 109–114, 2020.
[14] H. Nandanwar and S. Nallamolu, "Depression prediction on Twitter using machine learning algorithms," in Proc. 2nd Global Conference for Advancement in Technology (GCAT), IEEE, pp. 1–7, 2021 DOI: 10.1109/GCAT52182.2021
[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing and Management, vol. 24, no. 5, pp. 513–523, Elsevier, 1988. DOI: 10.1016/0306-4573(88)90021-0

AUTHOR INFORMATION

Prof. Bhushan V. Wakode, Assistant Professor, Department of Information Technology, Government College of Engineering, Amravati