Stress Detecting from Social Media Interaction

A Big Data Mining Approaching

Group name: The Enigma Ensemble

Tri Quan Do
(Group leader)
Computer Science
University of Illinois at Chicago
Chicago IL, USA
tdo22@uic.edu

Mosrour Tafadar
Computer Science
University of Illinois at Chicago
Chicago IL, USA
mtafad2@uic.edu

Hina Khalid
Computer Science
University of Illinois at Chicago
Chicago IL, USA
hkhali21@uic.edu

Safiya Mustafa
Computer Science
University of Illinois at Chicago
Chicago IL, USA
smust3@uic.edu

1. Problem and Goal

1.1. Project Abstract

Emotional and mental stress are serious issues that can have a significant impact on our well-being. Despite the fact that an emotional experience usually starts as a personal, internal process, it frequently results in the communal sharing of emotions with others. Emotions that are verbally expressed to others by the individual who has experienced them are referred to as being socially shared. People share their emotions with others in more than 80% of all emotional events, regardless of their age, gender, personality type, or culture (Bazarova, Choi, Sosik, Cosley, Whitlock 1). Due to social media's widespread use, people are accustomed to posting about their everyday activities and connecting with acquaintances on these platforms, making it possible to use information from online social networks to identify stress.

In fact, studies have found that more than 69% of Facebook and Twitter users express their feelings there, with negative feelings being expressed more frequently than positive ones. In a similar vein, an analysis of Instagram posts revealed that 40% of users said their use of the platform caused them stress. These results demonstrate the potential of social media as the sources of information for tracking and comprehending mental health and human emotions (Dixon 1). Social media represents a promising source of data for investigating individuals' mental states. By analyzing the textual content of social media posts, analysts can gain valuable insights into individuals' stress levels. The present project aims to examine the contextual features of social media posts and utilize machine learning techniques to predict individuals' stress scores. This research endeavor is expected to contribute to the existing literature on the relationship between social media use and mental health, with potential implications for the development of effective interventions and support systems for individuals experiencing stress.

1.2. Project Introduction

The initial step of this research project involves identifying a set of words that are commonly associated with emotional stress. Using this set of words, the models aim to compute an overall stress score for each individual under investigation. However, it is critical to acknowledge that some words may carry a higher intensity than others. Hence, the project purpose will segregate the identified set of words into distinct categories based on their intensity levels, namely high, moderate, and low to parallelly conduct a word frequency analysis to identify words or phrases that occur frequently, specifically those that pertain to emotions or stress. This research approach is expected to provide valuable insights into the underlying patterns and associations between language use and emotional stress, thereby contributing to the existing knowledge base on the topic.

Robust technologies for processing and analyzing massive amounts of social media data include Support Vector Machines (SVM) and MapReduce, which can be used to forecast stress levels based on social media posts. SVM is a machine learning algorithm that divides the data into classes before identifying the hyperplane that best distinguishes the classes. Large datasets can be processed concurrently on a distributed computing system using the model and software framework known as MapReduce.

1.3. Project Limitations

Although many study was conducted on examining social media posts to gauge stress levels, it is crucial to recognize that these posts frequently fail to accurately reflect a person's emotional condition. Individuals often only share the positive aspects of their lives on social media because of the stigma associated with doing so. Furthermore, various people have diverse preferences for social networking sites. Our study focuses primarily on Twitter because of its character limit of 240, which can prevent complicated emotion from being expressed. Despite conventionally conveying the personal emotion using photographs or brief videos, the research is restricting text-based contributions. In spite of these restrictions, looking at social media posts might nevertheless offer insightful information about people's stress levels.

2. Formalization

2.1. General Approach

The emotional distress word by using $S = \frac{m}{n}$ where n is the total number of emotional distress words exists in dataset and m is the total number of words that shown in the social media for an individual user. S be the continuous interval representing the stress of the social media post from [0, 1]. The formula presented above provides a fundamental approach to analyzing social media posts for stress levels, as it does not account for the varying intensities of individual words. To better capture the nuances of social anxiety and its expression on social media, the research proposes the following formula.

$$S = \left(\frac{l}{n}\right) . \sum(f_i \, . \, c_i) \text{ where}$$

n : total number of emotional distress words in group.
m: number of distress words by the user in message.
S : stress level of the message, continuously on [0,1].
i : the i-th distress word in your group.
f_i : the frequency of i-th in the user's message.
c_i : the intensity of words, continuously on [0,1].

2.2. MapReduce Approach

For processing and analyzing huge datasets in parallel, the programming model MapReduce is utilized by dividing the data into smaller bits and processing them concurrently across a cluster of computers, enhancing the processing speed and effectiveness. To divide the social media posts and evaluate the context in order to distinguish between different stress levels, MapReduce will be employed. For the purpose of identifying trends in the data and various trends that might be found in the dataset, various word combinations will be compared to an existing key. The posts made by the individuals can be examined during data querying in order to ascertain their level of stress and classify them appropriately. Additionally, MapReduce will be used in sorting the data into meaningful groups of stress levels and then classifying them accordingly to return a calculated stress score.

2.3. SVM Approach

In order to generate a decision boundary that maximizes the margin between the two classes, SVM first determines the boundary between the two classes of data points, in this case, stressed and non-stressed people. SVM is highly suited for assessing social media data because it has been demonstrated to be effective in analyzing huge datasets and can handle high-dimensional feature spaces.

2.4. L2 Regularization

L2 regularization can be employed in the SVM model to avoid overfitting the training data and to enhance its generalization capability. L2 regularization functions by including a penalty term in the cost function, which compels the model to minimize the weights of the features that are insignificant in predicting the target variable. Consequently, the model becomes less responsive to noise in the data and can generalize more accurately to new, unobserved data.

3. Data

Our approach can be applied to any social media platform where users share their daily activities. We can consider the following datasets (A direct link to the dataset is provided in the reference section):

1. *Facebook:* On Facebook, users share their activities in a status update. Any information that is shared can be collected and analyzed using a similar approach as outlined in our project.

2. *Twitter:* Twitter users also share their daily activities through their tweets. For this project, we will be using Twitter data as our primary dataset.

3. *Reddit:* Similar to Facebook and Twitter, Reddit data can also be collected and analyzed using our approach. However, it is important to note that Reddit data is community-based, so if we collect data from stress or depression communities, it is highly likely that the data will contain high levels of stress-related content.

In the scope of the research, data is restricted users to utilize solely tweet data, Facebook status data, or Reddit main post data. However, a more exhaustive examination of social media and stress would entail obtaining not only textual data but also multimedia such as images, videos, and information regarding users' interaction with posts, such as sharing and commenting. By incorporating these supplementary sources of data, the project can attain a more profound comprehension of the influence of social media on stress levels.

4. Schedules

The schedule is tentative, in which it could be updated based on the team progress and the success in training models and data.

Date plan	Content
Feb. 25th	Data planning, including team to verify and generate approach to clean data
Mar 01-03rd	Project proposal due, team will generate document for proposal idea and the leader last verify
Mar 05-10th	Data cleaning, including eliminating outliners, invalid datatypes followed by IQL models. Data checking!
Mar 15th	Applying MapReduce to have word groupby
Mar 25th	Applying SVM models for classification
Apr 02nd	Mapping the world with slang/negative word set
Apr 03rd	Analyze the similarity and occurrence
Apr 07th	Midterm check, preliminary report to Gradescope
Apr 11th	Generate L2 Regularization to check over/underfitting
Apr 26th	Team final review checking code, result consistence
Apr 27-30th	Team generate reports and leader will last verify
Apr 30th	Final Report via Gradescope at 11:59 PM

REFERENCES (APA Style)

[1] Brandwatch. (2021, January 5). Social media demographics to drive your brand's online presence in 2021. https://www.brandwatch.com/blog/social-media-demographics/

[2] H. Lin et al., "Detecting Stress Based on Social Interactions in Social Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1820-1833, 1 Sept. 2017, doi: 10.1109/TKDE.2017.2686382. https://ieeexplore.ieee.org/document/7885098

[3] Hootsuite. (2021, February 10). Social media use statistics to inform your 2021 strategy. https://www.hootsuite.com/resources/social-media-use-statistics

[4] Kazanova, A. (n.d.). Sentiment140 dataset with 1.6 million tweets. Kaggle. https://www.kaggle.com/kazanova/sentiment140

[5] Ng, A. (n.d.). L2 regularization. Retrieved February 25, 2023, from https://www.coursera.org/lecture/machine-learning/l2-regularization-OmmMx.

[6] Perrin, A. (2021, April 7). Social media fact sheet. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/social-media/

[7] Perrin, A. (2015, January 15). Psychological stress and social media use. Pew Research Center. https://www.pewresearch.org/internet/2015/01/15/psychological-stress-and-social-media-use-2/

[8] Statista. (n.d.). Social networks - statistics & facts. https://www.statista.com/topics/1164/social-networks/

[9] The Data Society. (n.d.). Twitter user data. data.world. https://data.world/data-society/twitter-user-data

[10] Towards Data Science. (2020, February 6). Support vector machine - introduction to machine learning algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47