Summary of the First Paper

Source:
https://www.researchgate.net/publication/327251557_Depression_detection_from_social_network_data_using_machine_learning_techniques

The objective of this paper is to analyze Facebook data and detect Facebook's user who might be depressed. But first we had to know these factors such as-

What depression is and what are the factors that lead to depression, what factor should we look for in Facebook comments and posts? How to extract these comments and what the relation between these comments leading to depression? What machine learning technique should we use to detect depression in these comments?

As we all know that Facebook users share their feelings by post or comments, their post and comments often contain emotional feelings like 'joy', 'sadness', 'fear', 'anger', or 'surprise'. So, in this study they collect data from different Facebook pages such as (from bipolar, depression and anxiety Facebook page). First, they synthesized the literature on various emotion detection techniques to detect depression. Second, they designated four features for their specific research problem and elaborate on the lesson learned from using each type. Third, their experiments are carried out on datasets of Facebook user's comments. Fourth, they suggest machine learning techniques to utilize all factors and maintain robustness. They also identify that a Decision Tree classifier outperforms other classifiers (a SVM, KNN and Ensemble) for their dataset.

They used NCapture tool to identify if the comment has any depression related content. It gives a place to arrange and deal with material to discover knowledge in a more proficient way. After collecting the raw data then it was analyzed by LIWC. It calculates the degree to which various categories of words are used in a text, and can process texts ranging from e-mails to speeches, poems and transcribed natural language in either plain text or Word formats. Our dataset consists of five emotional variables (positive, negative, sad, anger, anxiety), three temporal categories (present focus, past focus and future focus), and 9 standard linguistic dimensions (e.g., articles, prepositions, auxiliary verb, adverbs, conjunctions, pronoun, verbs and negations).

They use psycholinguistic dimensions for considering five features of the emotion state manifested in the comments: positive affect (PA), negative affect (NA), sadness affect (SA), anger affect (AA), and anxiety affect (AnA). Generally, temporal process word provides information about past focus category, present focus category and future focus category of how people are referencing each other and their degree of emotionality. In their study they use nine specific linguistics features (articles, prepositions, auxiliary verbs, adverbs, conjunctions, personal pronoun, impersonal pronouns, verbs, and negations) to characterize user comments for their experimental analysis.

The analysis is conducted using MATLAB 2016b. We applied four major classifiers: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision trees (DT), and Ensemble. Each classifier has sub-classifiers such as Decision trees—Simple DT, Medium DT, and Complex DT; SVM—Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, and Coarse Gaussian; KNN—Fine, Medium, Coarse, Cosine, Cubic and Weighted, Ensemble— Boosted tree, Bagged tree, Subspace discriminant, Subspace KNN, RUSBoosted Tree. The evaluation matrices

parameters (precision, recall and F-measure) have been used to execute these classifiers. It has conducted based on four different ways. True Positive (TP) = the depression cases that are positive and anticipated as positive True Negative (TN) = the depression cases that are negative and anticipated as negative. False Negative (FN) = the depression cases that are positive but anticipated to be negative. False Positive (FP) = the depression cases that are actually negative but anticipated to be positive.

For a better understanding of the general intuition behind depression, in this paper, they applied Decision Tree, KNN, SVM and Ensemble classifier techniques for depression detection of emotional terms. They showed that all of these classification techniques based on linguistic style, emotional process, temporal process and all (Linguistic, emotional and temporal) features are able to successfully extract the depressive emotional result. It can be observed that Decision Tree gives the better outcome.

All of the classifiers results are almost between 60 and 80%. Though it gives a good outcome but it's not perfect. So they need to make it more efficient. They need to collect more data and train they program to do better result. Close to 90% would be more efficient. There is no mention about time complexity about the whole process to be done. If it needs a lot of time then it will not be a better process.

This study gave a good idea how to collect data from Facebook comments and posts and what pages to look in for the data. How to analyze data by LIWC? I didn't know about this tool so it has made easy for me to analyze the data. The study also confirms that decision tree gives optimal result so there is no need to try all the classifier. There is also a good idea about the whole project such as how to differentiate the emotional words from the joyful word. And how to do the linguistic, emotional and temporal features successfully.

Summary of 2nd paper

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6431661/?fbclid=IwAR3Jbm1Fkpqv-rVeYkBNiMGIoBYbzKCWxnjG-4dZQY_3RoWExTdtaq6wnNI

In this study they used DASS-21. The DASS-21 is a set of three self-report scales designed to measure the emotional states of depression, anxiety and stress. In this study, we work with the subscale of depression, which assesses dysphoria, hopelessness, devaluation of life, self-deprecation, and lack of interest/involvement, anhedonia, and inertia. In our study, we work with either the total score (0–21 points achieved) or with the cut-off score (non-depressive ≤ 6, depressive > 6).

Four fictive letters were written on a computer in a pre-defined electronic interface. All four letters were written by each participant. The content of the text could be entirely fictional. The analyses were conducted on 688 texts that create a corpus of 99,481 words. In all texts, quantitative linguistic variables on various levels of classification (e.g., number of all adjectives, number of superlative forms of adjectives, number of words in singular, etc.) were automatically detected in the process of lemmatization with morphological tagging. Quantitative linguistic variables are included in the analyses in the form of relativized isolated features (ratios) and compound indicators (special metrics) as described in the following lists.

Data was collected by participants who were recruited using leaflets and advertisements on social networks. Quota selection was used to sample participants. The decisive criterion for determination of quotas was age, gender, and education. The model they are using is regression. One of the difficult questions was the choice of linguistic variables to include in the models. We have decided for a statistics-based procedure. In the first step, we have excluded variables with low variability. Sufficient variability has been proven for 6 of 16 selected single morpho-syntactic variables: the number of words per sentence, number of finite verbs per sentence, number of punctuation marks per sentence, proportional variables of relative occurrence of singular, possessive singular, negativity, and for all 8 indexes consisting of combinations and ratios of more morpho-syntactic characteristics: index of coherence, pronominalisation, formality, trager, readiness to action, aggressiveness and activity. This means that, in our study, only a limited amount of selected single morpho-synaptic characteristics was found to be suitable for use in distinguishing between non-depressive and non-depressive texts because of low variability; while all indexes showed sufficient variability.

The present study was conducted on a quota-selected sample of Czech native speakers. I think there are flaws about choosing the model too. Because regression model can only work with number so they had to improve the collected data first to work with the model. That's why I think if they choose any classifier like decision tree, it would work better.

Working with this study is not a waste, because it gave me a good idea about how to work with regression model. I also learn about how to collect data without any social media. How to collect these data from participant? I also know about the DASS-21 system. So if we ever need to use it in our project in future I have idea how it works.