AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

FACULTY OF SCIENCE AND TECHNOLOGY

Prediction of suicidal behavior in twitter data using machine learning

A Thesis Presented to the
DEPARTMENT OF COMPUTER SCIENCE
In Partial Fulfillment of the Requirements for the Degree
BSc. in CSE

Supervised By

Md. Tohedul Islam

Assistant Professor

Department of Computer Science

Faculty of Science and Technology

Submitted By

Bushra, Mahfuzatul – 16-31040-1

Noman, Abdullah Al – 16-32102-2

Fahmida, Maisha – 16-32192-2

Tareq, Md. Ikram – 16-32231-2

Declaration

We would like to declare that this is our first thesis book and has not been awarded any degree or diploma at a university or other tertiary institution. Information from published and non-published publication is permitted in the text and a reference book is provided.

_____

Bushra, Mahfuzatul

_____

Noman, Abdullah Al

_____

Fahmida, Maisha

_____

Tareq, Md. Ikram

Approval

The thesis titled "Prediction of suicidal behavior in twitter data using machine learning" has been submitted to the following respected members of the board of examiners of the department of Computer Science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science on July 2020 by Bushra, Mahfuzatul (16-31040-1), Noman, Abdullah Al (16-32102-2), Fahmida, Maisha (16-32192-2) and Tareq, Md. Ikram (16-32231-2) has been accepted as satisfactory.

Md. Tohedul Islam
Assistant Professor and Supervisor
Department of Computer Science
American International University
Bangladesh

DR. M. M. MAHBUBUL SYEED
Associate Professor and Department Head [FST]
Department of Computer Science
American International University-
Bangladesh

Prof. DR. Tafazzal Hossain
Pro Vice Chancellor, Dean
Faculty of Science and Technology
American International University-
Bangladesh

DR. Carmen Z. Lamagna
Vice Chancellor
American International University-
Bangladesh

Acknowledgements

At first we would like to express countless gratitude to our honorable supervisor Md. Tohedul Islam sir for introducing to this interesting topic and guiding us. His profound knowledge in this field, keen interest, patience and continuous support lead to the completion of our work. His instructions have contributed comprehensively in every aspect of the thesis.

Finally, appreciations is placed for respected parents, honorable teachers, fellow classmates and supportive friends for sharing their knowledge and ideas that contributed in accomplishing this thesis and to all who participated in many ways during this thesis.

Abstract

Nowadays, people try to express their thoughts, emotions, ideas, and activities through their status using social media such as Twitter, Facebook, Instagram, and so on. We hypothesized that social media would be linked to intellectual disabilities. To test this hypothesis, we used all pre-post prospective longitudinal design with a sample of twitter users (n = 132). In this study, we report an advanced suicide prediction model using social media data such as twitter along with observed suicide data. We have used live twitter data for our databases as well as previously labeled twitter and look for conclusions about data to gain insight into the potential of users who are going through suicidal thoughts. We have used some classification using Weka tools to get the accuracy of our dataset. We have noted that people who are well-known and often work in the media may be depressed and try to attempt suicide. We believe that our work has helped to gain some insight behind the reason for suicide.

Keywords: Artificial Intelligence, Suicide, Twitter, LIWC, WEKA.

A. Table of Contents

Chapter 1

1. Introduction

In today's world, social media is one of the ways to express ideas and our character. Especially for adults, social workers are some kind of catharsis. Through these experiences, we can absorb the feelings of the persons and the things that they are going through. People who do wrong do not show their feelings through the media in the near future [23]. Compare to users who spend more time on social media such as Twitter, Instagram, Facebook and other platforms with a higher level of stress than those who spend less time, according to recent studies [24]. There are a number of causes and symptoms online that can mentally make a person suicidal or not. Our project focuses on identifying people who are committing suicide based on what they post. Despite being aware of the potential risk of suicide in an accident, the data is still not included in a reliable estimate. At the same time, the current suicide risk assessment program is still very active, with risks viewed as part of a system that relies heavily on clinicians. Again, traditional diagnostic methods, which were based on data analysis, failed to predict suicide patterns over the age of probability. In view of this impossibility, a high-resolution machine learning (ML) method, as part of artificial intelligence (AI), is being developed by increasing the frequency of improving suicide prevention. The purpose of AI is to develop systems that enhance aspects of a person's intelligence such as planning, thinking, accepting values and problem-solving. The method involves machine (i.e. non-human) 'learning' to identify patterns and strategies using mathematical or numerical algorithms [25]. AI can now be considered 'narrow or fragile' because it requires a clear strategy to perform a specific function [26]. The evolution of AI technologies is completely dependent on ML. Taken together, ML and AI have been developed to develop algorithms that can determine the consequences of accidents (and safety) and suicide outcomes, predict suicide, and identify those at risk. Beyond the prediction, ML algorithms can also facilitate the development of 'smart' technologies that, by enhancing human intelligence, can detect and respond to suicidal behavior immediately. These AI applications can be accessed immediately in crisis or deployed at the hospital level to support risk assessment, behavioral management, or acceptance. The purpose of this study was to analyze predictions about suicide using information sources from social media. We have chosen to explore the Twitter community. Twitter is the eighth most popular site in the United States and the third most popular social media site, after Facebook and Instagram [27].One in four adults on Twitter (about 700 million users) sends more than 350,000 tweets per minute [28].The highest percentage of Twitter users are Americans; up to 40% of children aged 18 to 29 use social media websites [29]. During this year, Twitter was more popular among college students (with 47% using Twitter) than non-college students [29]. Twitter's easy-to-find timeline as well as users 'ability to publish unlimited content makes it one of the best social media sites to test our concept.

1.1. Problem Statement

AI-based suicide predictions are developed on two separate tracks. In “Medical Error Prediction,” AI analyzes data from a patient’s medical records. In "Social Voice Prediction," AI analyzes consumer behavior derived from social media, smartphone applications, and the Internet of Things. Medical homicide prediction tools are developed by methods based on some regulations and by the way developers are published in the reviewed academic journals. In contrast, predictions of social suicide are usually outside the health care system. Corporations hold their suicide prediction methods as proprietary trading secrets. Despite the lack of transparency, predictions of social suicide are used around the world to affect people's lives every day. But little is known about their safety or efficiency. It would be better if there is a way or a system to know someone's predicted results by using social media's data such as twitter in advance and this may help other people to take the necessary steps to make a decision or get to a conclusion. Furthermore, people also can use this system to predict suicidal activity before it is too late to take action.

1.2. Objectives

The primary objective of this study is to predict suicidal activity through twitter status so that needful actions can be taken for the betterment of people. Besides based on the collected data some observations can make to help people for taking a wise decision.

1.3. Research Questions

After visualizing numerous problems in problem statement portion, the purposes of this paper are ascertained and this research represents the answer of the following questions:

1. How can we predict suicidal case activity by tweets?
2. What attribute should consider and to detect suicidal case?

Chapter 2

2. Literature Review

AI means Artificial intelligence, it is also known as machine intelligence. Intelligence that is demonstrated by machine is called Artificial intelligence. It is the ability of computer program or a machine to think and learn by using this. A machine can do any type of action. AI has many different fields such as mathematics, linguistic, psychology, neuroscience and philosophy. It makes computer smarter than before. A perfect machine with intelligence is a flexible intermediary which can understand its environment and do action to reach a goal.

Assessing Suicide Risk [1] …, they used A Web-based survey of Weibo users to assess the respondents' suicide risk and emotional distress (ie, depression, anxiety, and stress). They also used Weibo API for downloading survey (12 months before survey). SC-LIWC Categories as Markers and Automatic Machine Classifiers were used in their research. For SC-LIWC Categories, P<.05 was selected as the cut-off for statistical significance. For Automatic Machine Classifiers, they applied SVM classification and they have found the significantly identified those with high suicide probability or severe anxiety.

Anxious Depression [2]…, for the dataset, they used Twitter API to scrap the data with past one month tweets of 100 users. The data consists of user name, tweet count, date of account creation, account verification status (verified or not), language and description about tweets. Multinomial Naïve Bayes, Gradient Boosting and Random Forest these three machine learning classifiers were used in their research. For the final prediction they used an ensemble vote classifier with majority voting mechanism. For Multinomial Naïve Bayes, accuracy 77.89 %;Random Forest, accuracy 81.04%; Gradient Boosting , accuracy 79.12%; Ensemble Vote Classifier, accuracy 85.09 %(Highest).

An Overview of Suicide Research [3] ….. , World Health Statistics Annuals are for 1987–1990, 1992, and 1994. For these six years they got their data based on suicides per 100,000 Population are reported. They got 151 variables and analyzed the data with the SPSS program. After all of analyzing, they had found that married people commit more suicides than the non-married in China. Married Chinese 51.6% commit suicide than singles 43.3% and 2.3% divorced; 2.8% widowed in China.

Advanced Daily Prediction [4]….., they obtained data from social media using Daumsoft, in South Korea which is one of the leading consulting firms and social media analysis. They used a serial prediction procedure to capture secular trends in the data for the 7-year time period between 1 January 2008 to 31 December 2014. A stepwise Akaike Information Criterion (AIC) used for selecting the best prediction model. 38.83 (standard deviation=8.51) with a range from 16 to 72 was the mean of absolute daily suicide number during this study.

Depression Detection using [5]…., Twitter API used to collect the dataset that comprises of tweets. A total of 10,000 Tweets were collected for generating the training and test dataset.

Multinomial Naive Bayes has been used as classifier. They used SVM for analyzing the dataset. SVM classifies them as one or the other of two categories. The results of Multinomial Naïve Bayes performed the best with the F1 score of 83.29 whereas SVM achieved a lower F1 score of 79.73.

Descriptive Analysis [6]…., after successfully acquiring an access key for developers account, they gathered their dataset live from twitter using Tweepy package in python. They used K MEANS algorithm for their appropriate result and obtained 6 clusters giving an insight to their usage of jargon's and thought processes. After all they finally found between people who are depressed and usage of social media.

Detection of suicide [7]…., they used the Twitter streaming API to collect tweets for dataset. They used the Fleiss kappa to measure agreement for the 55 tweets with three annotators that gives 78.3%.multinomial Naıve Bayes, Sequential Minimal Optimization (SMO) with a poly kernel, C4.5 decision tree (J48), nearest neighbor classifier (IB1), multinomial logistic regression, rule induction (Jrip), Random Forest, and SMO with a Pearson VII universal kernel function (PUK) all of these 8 classifiers used in their research by Weka. The Highest precision was 71% and 80.7% for J48 and SMO.

Ethics and Artificial Intelligence [8]…., they used Machine Learning because it provided the most nuance in deter-mining the suicidal intent of a post. They created ML model, there is an evaluation phase to assess the performance quality of the model, that is, to assess if the classifier is good enough to put it to use. Finally their result was that social support can help prevent suicide. Facebook gives us the opportunity to reduce our distress.

Prediction by data mining [9]…., they utilized the data from the survey of mental health in 2011 from the National Youth Policy Institute of Korea for their study. Conduct a decision tree analysis of data mining they used the Answer Tree 3.0 program. Data mining is efficient for analyzing big data because there is an interaction between a great numbers of variables. The model gives us the knowledge of depression show distinctive pathways with different reasons to suicide attempts. At the end of all, they found that depression is the main reason for suicide.

Machine Classification and Analysis [10]…, Twitter Streaming Application Programming Interface (API) used to collect data from Twitter. They classified data into six suicide related categories. Support Vector Machine (SVM), Rule Based (we used Decision Trees (DT)), and Naive Bayes (NB) were the classification that they used in their research by Weka. The result of the Rotation Forest (RF) was the highest accuracy.

Suicide Prediction in Twitter [11]…, data was collected from ten different websites that contains effective and identifiable material related to suicidal contemplation.TF-IDF method used in this research. This method has created a list of 62 keywords and expressions that related to suicide. In this research they found the best classifier is NB method produced score of 0.82 as precision.

Machine Learning and Semantic [12]…, for their Dataset Twitter4J API is used to collect tweets .Twitter4J is a Java library for programming applications related to Twitter. It is an integrated Java library. They used some Algorithms like IB1 J48 CART SMO Naive Bayes. In their research they showed us correctly classified instances and incorrectly classified instances.

Multi-class machine classification [13]…, they collected their data from known suicide Web forums, blogs and micro-blogs, and asking human annotators. Term Frequency or Inverse Document frequency was applied by this paper[13] of TF-IDF method to the corpus of annotated documents. They used some Classifier such as NB DT SVM RF. This research gives us the idea about the interrelation between the rate of tweet and the rate of suicide.

Recognizing Depression from [14]…., they gathered their information on depression levels of Twitter users and their tweets. For this, they published a website to administer a questionnaire and disseminated information about the website over Twitter1 through the Twitter application programming interface (API)2. MeCab was used to for morphological stemming and categorization. The tweets of each user, as estimated by using a representative topic model LDA. SVM classifier was used for estimating the presence of active depression and the accuracy. This study explained how useful the various features extracted from Twitter user Tweets or history for recognizing depression, and the degree of accuracy with which the presence of active depression could be detected by using these features.

Sentiment Analysis [15]…., First, they had managed a number of profiles from the social network Twitter through available data. Using R code they eliminate all noisy words from the original text. It was provided a high quality of suicide risk. LIWC text analysis software (Pennebaker, 2001) was used to get more relevant features associated with emotions detection. NLP3 as a machine learning based toolkit for processing natural language, which is used to understand the language used in a text and uncovers the sentiment behind it. They had selected five classifiers including BayesNet, Adaboost, J48, SMO and Random Forest by using Weka. The best performing classifier in terms of precision is Random Forest yielding a value of 83%.

Based on these Research works, many researchers were used different tools and methods of data mining techniques which is related to AI to predict and analysis. For analyzing suicide attempts most of the researchers try to suggest the main reason behind suicide using analyzing methods. These papers have their different data from different site .Most of the cases, depression have been cited as the major cause for suicide. Some have tried to give some solution for reducing depression.

In previous works most of the data were collected from different web sites but few data were collected using API. API gives live data or current data. For this reason, this paper collected data from twitter directly. Python is used for scraping data by using API. Here, this paper have collected data of some famous people who have committed suicide and also active on twitter . These previous papers have some lacking of data cleaning. Data cleaning is mainly used for improving the quality of data. These papers used their data directly. That can't give specific

prediction . Firstly, this paper have applied LIWC on scraping data for text analysis which gives meaningful categories psychologically. Then R language have used in these meaningful categories for again data cleaning .This time this paper have applied regsubsets with nvmax 4 to 10 which gives some asterisk variables or features . This specific variables help us to predict whether any person commit suicide or not.

Chapter 3

3. Methodology

The first thing we have done is searching several websites where we can find out the peoples with their basic information who committed suicide past several years. Then analysis twitter profiles through available information. We try to find out profiles of popular person who committed suicide. At that time we do not pay attention to the tweets .We confirm identity by other features like User's name ,profile photo, country, language, profile creation date, profile description, followers, likes, re-tweets and no tweets after death.

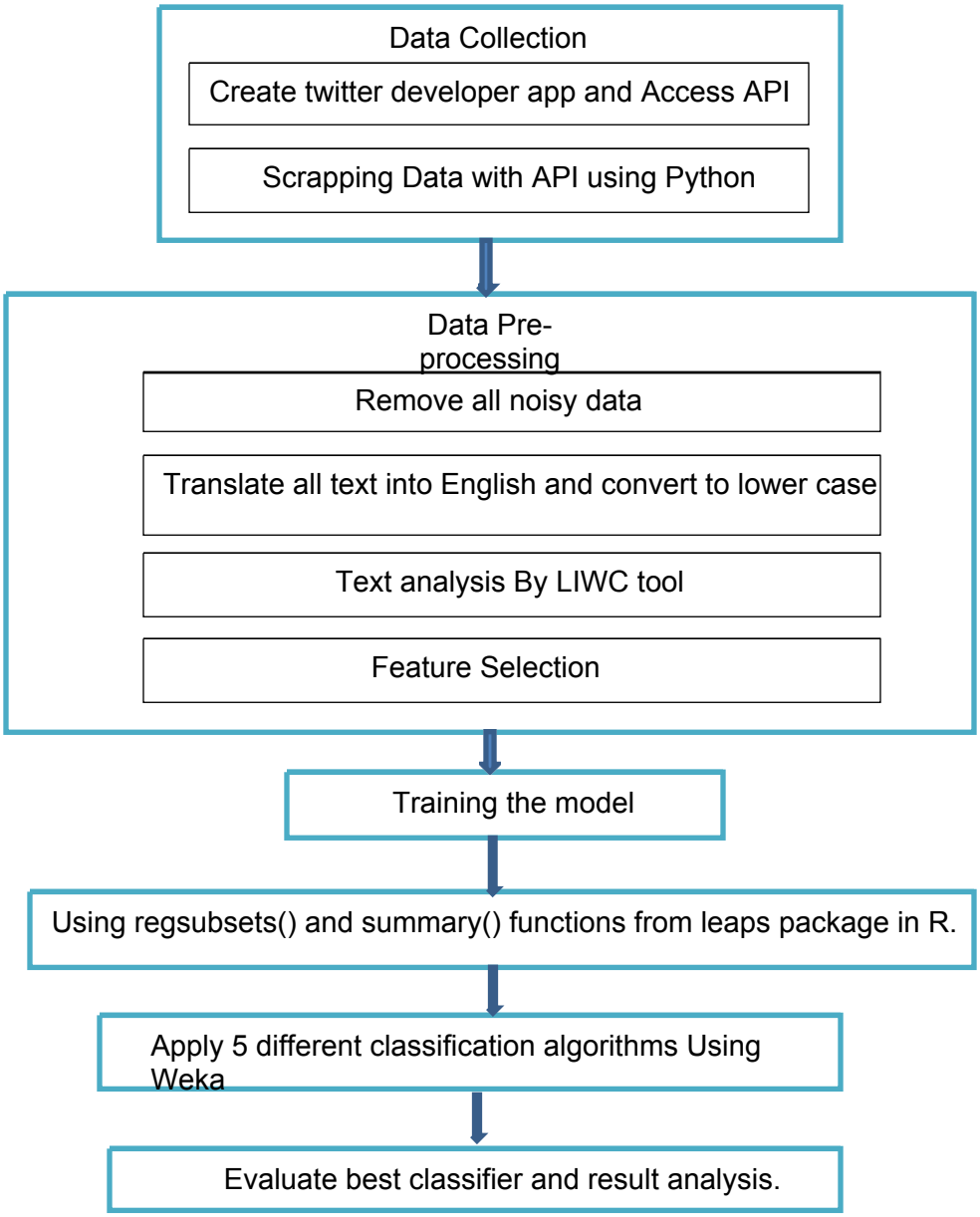


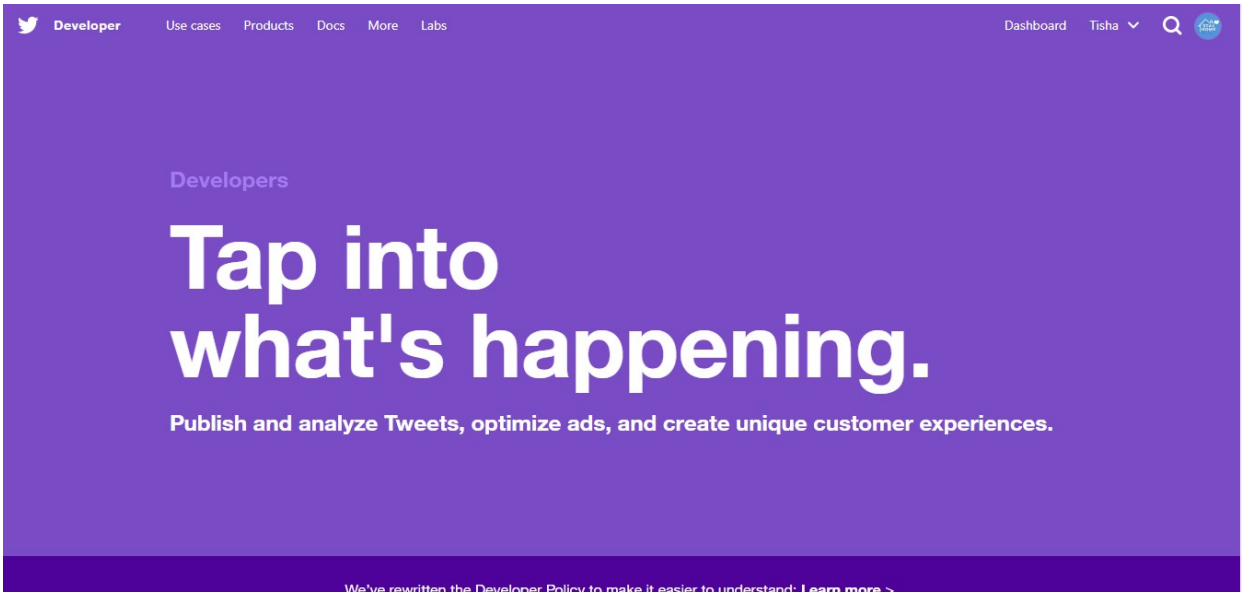
Figure 1: Proposed Framework

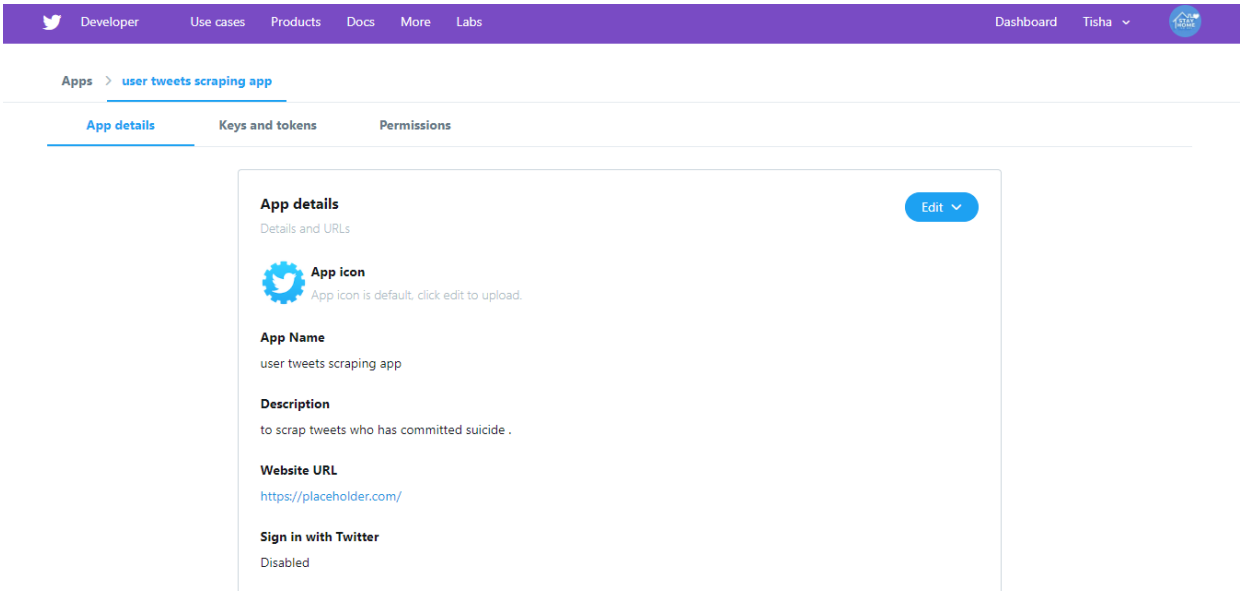
3.1. Data Collection:

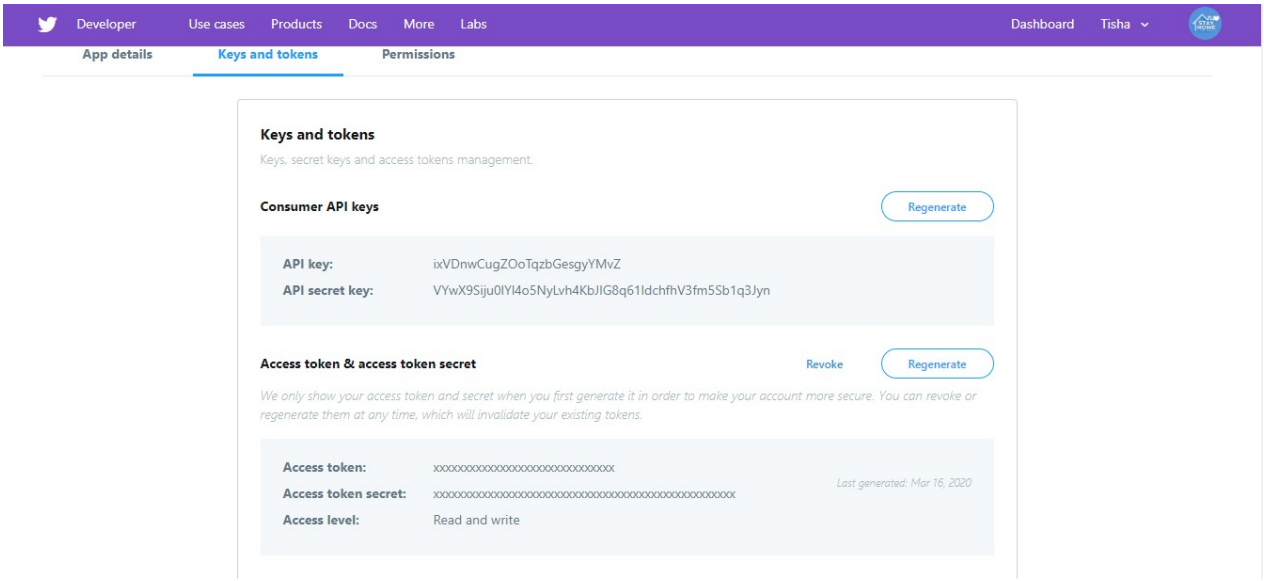
For collecting data from Twitter, we get started our work by Twitter API access. So we easily can download data and do some interesting analysis using python And R programming.

3.2. Scrapping Twitter data:

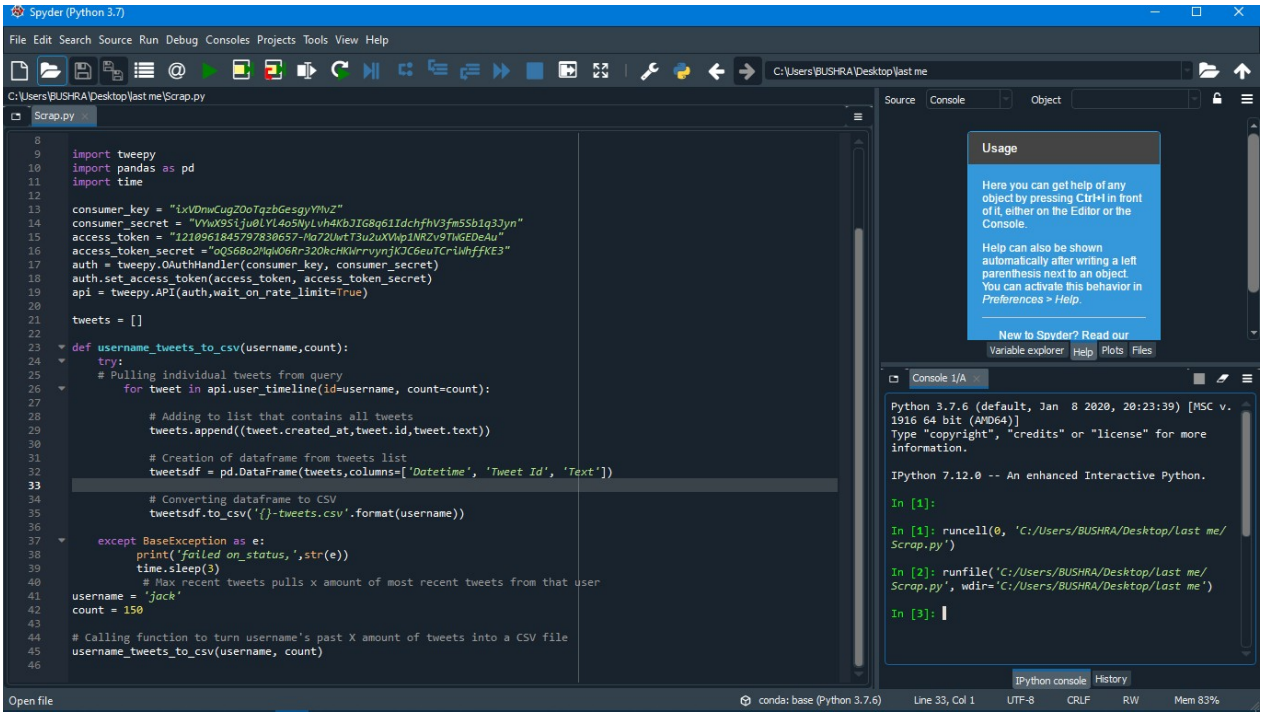
An official API is offered by Twitter which we use to scrape public twitter data. It is free to use. Firstly we apply for a developer account. This is generally used for both academic research or building an app. We fill out the form and create an app named ‘user tweets scrapping app’. Within 24 hours twitter granted our access. Then our twitter account is approved to use their API, after that we generate a bearer token to access their Standard Data API.







Then with the API keys and Access token we scrap data by python. With the following code we scrape tweets with User name .The data automatically save as csv file to the folder where we have defined to save. In each csv file there is maximum 150 tweets.



By this code we collect 68 csv files of twitter users who did not commit suicide and 67 csv files who committed suicide.

3.3. Data preprocessing:

For further use of data we need data preprocessing. Some datasets are publically available for Twitter those are hash tweets, emoticons, ISIEVE tweets, patient dataset, Stanford movie reviews, spam dataset, sarcasm and nasty review etc. Also some tweets with informal language with Unicode characters, acronyms, spelling mistakes, URL's (e.g.www.diggy.com), abbreviations, unknown symbols, contractions. So it is necessary to tidy out and normalize data. As the tweets we collect the users are from different countries.

1. Firstly we translate all data from different languages to English.

2. For translating we have used Google translator.

3. The tweets were stubby and turbulent. We correct them and convert all text to the lower case.

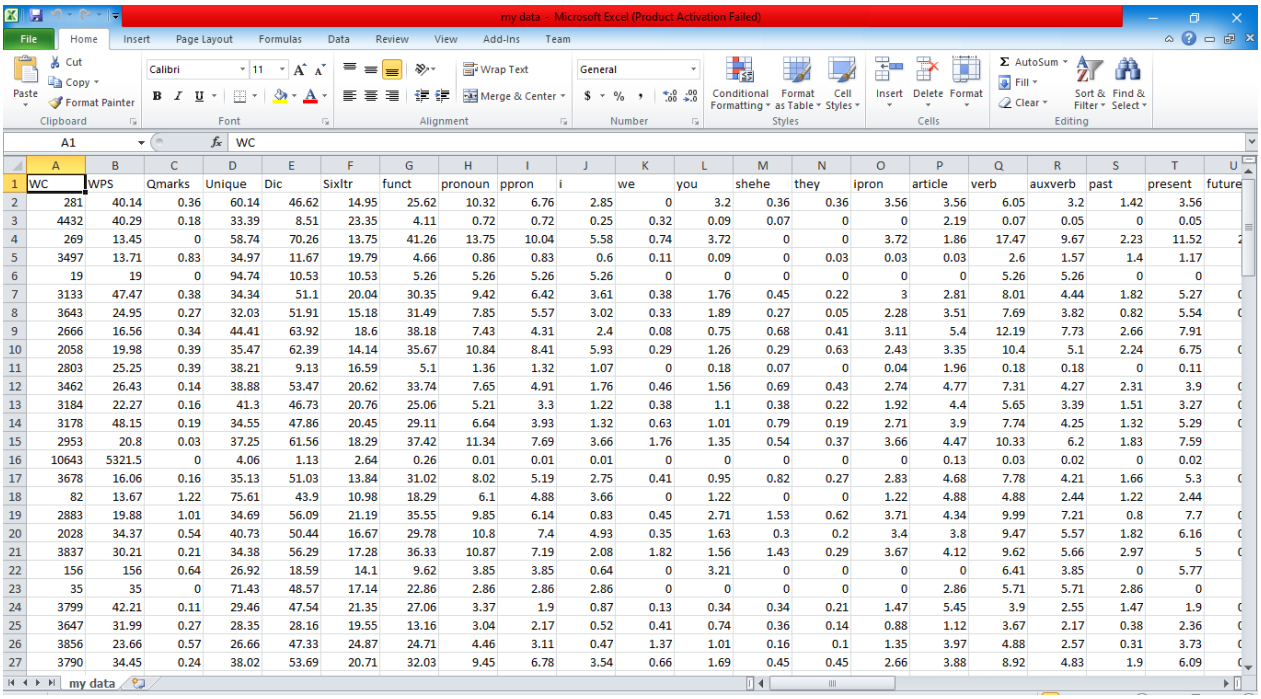
4. We eliminate all noisy words from the csv files.

3.4. Features Selection and Dataset Categories:

For delivering the exact probability of suicide, collecting the wealth of information about user is so important. Several types of features can detect the risk of suicide more efficiently. So we use data mining tool LIWC (Linguistic Inquiry and Word Count) which analysis the text word by word analogous with emotions and the percentage of words in a text which consist of psychological and linguistic features. For lifting details of user's emotional side LIWC help us a tot. Emotional features are essential for sentiment analysis also.

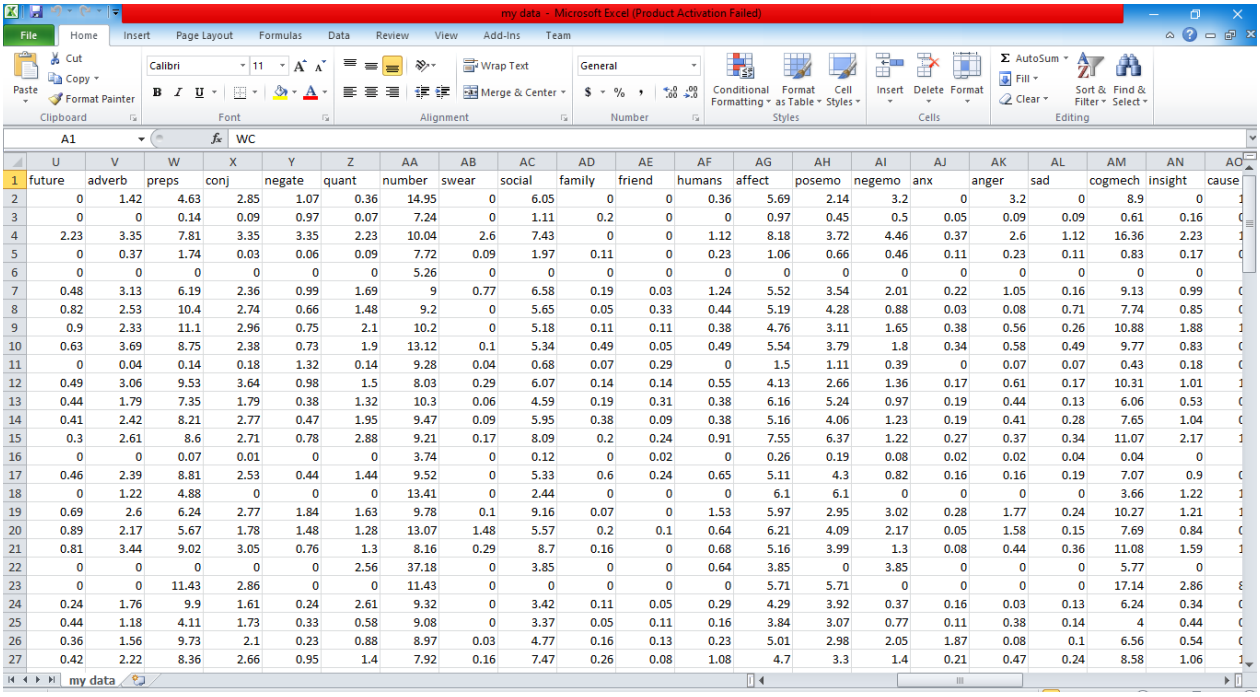
As we know people have different written style. Linguistic features are used mostly for preventing suicidal risk. LIWC tool is used for extract linguistic features included LIWC features (such as adjective, pronoun, adverb, health, death etc). It also calculates percentage of used hash tags, special words, frequent words, N-grams, elongations, number of used languages, average length of used words, average length of used sentences etc. But emotional features, emoji's, temporal features, sentiment analysis are also important for detect suicide risk through tweets. LIWC variables are included here 70 variables with important linguistic, psychological and social processes.

In LIWC some steps are followed to show the output. First select the text file .Then LIWC process the words and count the total words. After that compared text with those in its internal dictionary .Finally calculated the percentage and build output. We process data with LIWC then get one dataset for users who committed suicide and other one who did not commit suicide. Then merge both and add an extra column named class for our further work. Who committed suicide used binary number 1 and who did not used binary number 0 for comparing with other column. After all the editing the final dataset contains 132 instances and 71 variables. In this dataset all are numeric values.



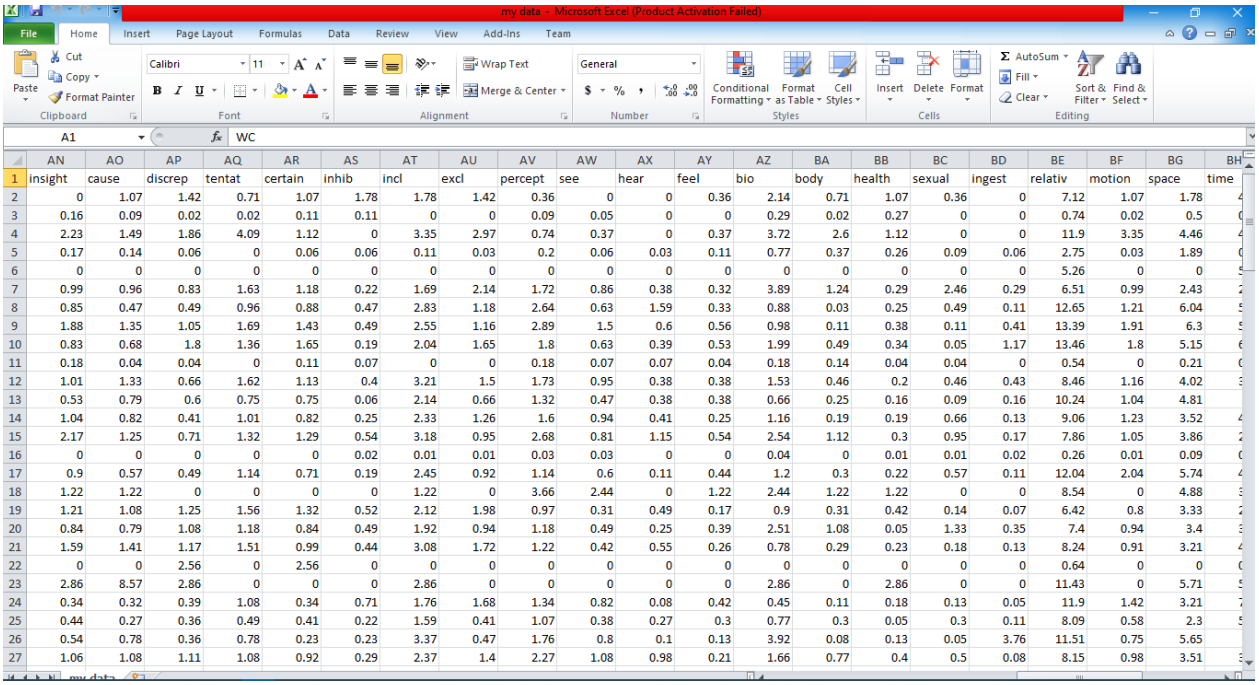
WC	WPS	Qmarks	Unique	Dic	Sixltr	funct	pronoun	ppron	i	we	you	shehe	they	ipron	article	verb	auxverb	past	present	future
281	40.14	0.36	60.14	46.62	14.95	25.62	10.32	6.76	2.85	0	3.2	0.36	0.36	3.56	3.56	6.05	3.2	1.42	3.56	
4432	40.29	0.18	33.39	8.51	23.35	4.11	0.72	0.72	0.25	0.32	0.09	0.07	0	0	2.19	0.07	0.05	0	0.05	
269	13.45	0	58.74	70.26	13.75	41.26	13.75	10.04	5.58	0.74	3.72	0	0	3.72	1.86	17.47	9.67	2.23	11.52	2
3497	13.71	0.83	34.97	11.67	19.79	4.66	0.86	0.83	0.6	0.11	0.09	0	0.03	0.03	0.03	2.6	1.57	1.4	1.17	
19	19	0	94.74	10.53	10.53	5.26	5.26	5.26	5.26	0	0	0	0	0	0	5.26	5.26	0	0	
3133	47.47	0.38	34.34	51.1	20.04	30.35	9.42	6.42	3.61	0.38	1.76	0.45	0.22	3	2.81	8.01	4.44	1.82	5.27	0
3643	24.95	0.27	32.03	51.91	15.18	31.49	7.85	5.57	3.02	0.33	1.89	0.27	0.05	2.28	3.51	7.69	3.82	0.82	5.54	0
2666	16.56	0.34	44.41	63.92	18.6	38.18	7.43	4.31	2.4	0.08	0.75	0.68	0.41	3.11	5.4	12.19	7.73	2.66	7.91	
2058	19.98	0.39	35.47	62.39	14.14	35.67	10.84	8.41	5.93	0.29	1.26	0.29	0.63	2.43	3.35	10.4	5.1	2.24	6.75	0
2803	25.25	0.39	38.21	9.13	16.59	5.1	1.36	1.32	1.07	0	0.18	0.07	0	0.04	1.96	0.18	0.18	0	0.11	
3462	26.43	0.14	38.88	53.47	20.62	33.74	7.65	4.91	1.76	0.46	1.56	0.69	0.43	2.74	4.77	7.31	4.27	2.31	3.9	0
3184	22.27	0.16	41.3	46.73	20.76	25.06	5.21	3.3	1.22	0.38	1.1	0.38	0.22	1.92	4.4	5.65	3.39	1.51	3.27	0
3178	48.15	0.19	34.55	47.86	20.45	29.11	6.64	3.93	1.32	0.63	1.01	0.79	0.19	2.71	3.9	7.74	4.25	1.32	5.29	0
2953	20.8	0.03	37.25	61.56	18.29	37.42	11.34	7.69	3.66	1.76	1.35	0.54	0.37	3.66	4.47	10.33	6.2	1.83	7.59	
10643	5321.5	0	4.06	1.13	2.64	0.26	0.01	0.01	0.01	0	0	0	0	0	0.13	0.03	0.02	0	0.02	
3678	16.06	0.16	35.13	51.03	13.84	31.02	8.02	5.19	2.75	0.41	0.95	0.82	0.27	2.83	4.68	7.78	4.21	1.66	5.3	0
82	13.67	1.22	75.61	43.9	10.98	18.29	6.1	4.88	3.66	0	1.22	0	0	1.22	4.88	4.88	2.44	1.22	2.44	
2883	19.88	1.01	34.69	56.09	21.19	35.55	9.85	6.14	0.83	0.45	2.71	1.53	0.62	3.71	4.34	9.99	7.21	0.8	7.7	0
2028	34.37	0.54	40.73	50.44	16.67	29.78	10.8	7.4	4.93	0.35	1.63	0.3	0.2	3.4	3.8	9.47	5.57	1.82	6.16	0
3837	30.21	0.21	34.38	56.29	17.28	36.33	10.87	7.19	2.08	1.82	1.56	1.43	0.29	3.67	4.12	9.62	5.66	2.97	5	0
156	156	0.64	26.92	18.59	14.1	9.62	3.85	3.85	0.64	0	3.21	0	0	0	0	6.41	3.85	0	5.77	
35	35	0	71.43	48.57	17.14	22.86	2.86	2.86	2.86	0	0	0	0	0	2.86	5.71	5.71	2.86	0	
3799	42.21	0.11	29.46	47.54	21.35	27.06	3.37	1.9	0.87	0.13	0.34	0.34	0.21	1.47	5.45	3.9	2.55	1.47	1.9	0
3647	31.99	0.27	28.35	28.16	19.55	13.16	3.04	2.17	0.52	0.41	0.74	0.36	0.14	0.88	1.12	3.67	2.17	0.38	2.36	0
3856	23.66	0.57	26.66	47.33	24.87	24.71	4.46	3.11	0.47	1.37	1.01	0.16	0.1	1.35	3.97	4.88	2.57	0.31	3.73	0
3790	34.45	0.24	38.02	53.69	20.71	32.03	9.45	6.78	3.54	0.66	1.69	0.45	0.45	2.66	3.88	8.92	4.83	1.9	6.09	0

Figure 2: The Training Dataset.



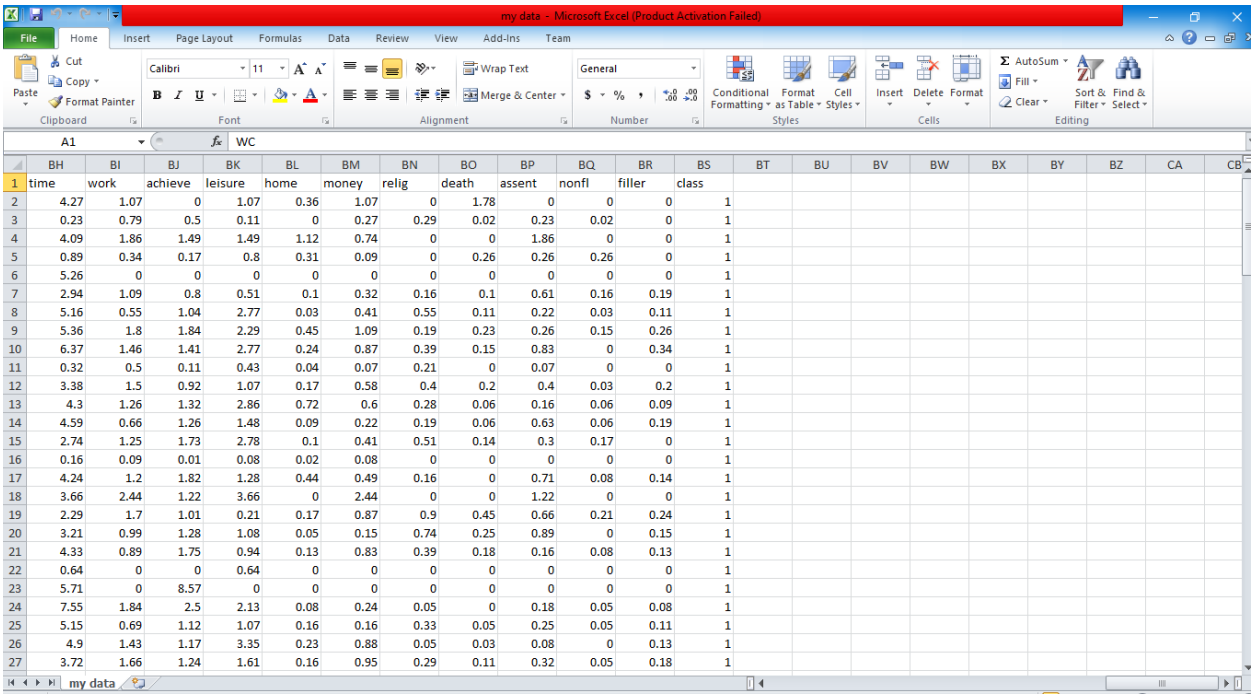
future	adverb	preps	conj	negate	quant	number	swear	social	family	friend	humans	affect	posemo	negemo	anx	anger	sad	cogmech	insight	cause
0	1.42	4.63	2.85	1.07	0.36	14.95	0	6.05	0	0	0.36	5.69	2.14	3.2	0	3.2	0	8.9	0	1
0	0	0.14	0.09	0.97	0.07	7.24	0	1.11	0.2	0	0	0.97	0.45	0.5	0.05	0.09	0.09	0.61	0.16	0
2.23	3.35	7.81	3.35	3.35	2.23	10.04	2.6	7.43	0	0	1.12	8.18	3.72	4.46	0.37	2.6	1.12	16.36	2.23	1
0	0.37	1.74	0.03	0.06	0.09	7.72	0.09	1.97	0.11	0	0.23	1.06	0.66	0.46	0.11	0.23	0.11	0.83	0.17	0
0	0	0	0	0	0	5.26	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.48	3.13	6.19	2.36	0.99	1.69	9	0.77	6.58	0.19	0.03	1.24	5.52	3.54	2.01	0.22	1.05	0.16	9.13	0.99	0
0.82	2.53	10.4	2.74	0.66	1.48	9.2	0	5.65	0.05	0.33	0.44	5.19	4.28	0.88	0.03	0.08	0.71	7.74	0.85	0
0.9	2.33	11.1	2.96	0.75	2.1	10.2	0	5.18	0.11	0.11	0.38	4.76	3.11	1.65	0.38	0.56	0.26	10.88	1.88	1
0.63	3.69	8.75	2.38	0.73	1.9	13.12	0.1	5.34	0.49	0.05	0.49	5.54	3.79	1.8	0.34	0.58	0.49	9.77	0.83	0
0	0.04	0.14	0.18	1.32	0.14	9.28	0.04	0.68	0.07	0.29	0	1.5	1.11	0.39	0	0.07	0.07	0.43	0.18	0
0.49	3.06	9.53	3.64	0.98	1.5	8.03	0.29	6.07	0.14	0.14	0.55	4.13	2.66	1.36	0.17	0.61	0.17	10.31	1.01	1
0.44	1.79	7.35	1.79	0.38	1.32	10.3	0.06	4.59	0.19	0.31	0.38	6.16	5.24	0.97	0.19	0.44	0.13	6.06	0.53	0
0.41	2.42	8.21	2.77	0.47	1.95	9.47	0.09	5.95	0.38	0.09	0.38	5.16	4.06	1.23	0.19	0.41	0.28	7.65	1.04	0
0.3	2.61	8.6	2.71	0.78	2.88	9.21	0.17	8.09	0.2	0.24	0.91	7.55	6.37	1.22	0.27	0.37	0.34	11.07	2.17	1
0	0	0.07	0.01	0	0	3.74	0	0.12	0	0.02	0	0.26	0.19	0.08	0.02	0.02	0.04	0.04	0	
0.46	2.39	8.81	2.53	0.44	1.44	9.52	0	5.33	0.6	0.24	0.65	5.11	4.3	0.82	0.16	0.16	0.19	7.07	0.9	0
0	1.22	4.88	0	0	0	13.41	0	2.44	0	0	0	6.1	6.1	0	0	0	0	3.66	1.22	1
0.69	2.6	6.24	2.77	1.84	1.63	9.78	0.1	9.16	0.07	0	1.53	5.97	2.95	3.02	0.28	1.77	0.24	10.27	1.21	1
0.89	2.17	5.67	1.78	1.48	1.28	13.07	1.48	5.57	0.2	0.1	0.64	6.21	4.09	2.17	0.05	1.58	0.15	7.69	0.84	0
0.81	3.44	9.02	3.05	0.76	1.3	8.16	0.29	8.7	0.16	0	0.68	5.16	3.99	1.3	0.08	0.44	0.36	11.08	1.59	1
0	0	0	0	0	2.56	37.18	0	3.85	0	0	0.64	3.85	0	3.85	0	0	0	5.77	0	
0	0	11.43	2.86	0	0	11.43	0	0	0	0	0	5.71	5.71	0	0	0	0	17.14	2.86	8
0.24	1.76	9.9	1.61	0.24	2.61	9.32	0	3.42	0.11	0.05	0.29	4.29	3.92	0.37	0.16	0.03	0.13	6.24	0.34	0
0.44	1.18	4.11	1.73	0.33	0.58	9.08	0	3.37	0.05	0.11	0.16	3.84	3.07	0.77	0.11	0.38	0.14	4	0.44	0
0.36	1.56	9.73	2.1	0.23	0.88	8.97	0.03	4.77	0.16	0.13	0.23	5.01	2.98	2.05	1.87	0.08	0.1	6.56	0.54	0
0.42	2.22	8.36	2.66	0.95	1.4	7.92	0.16	7.47	0.26	0.08	1.08	4.7	3.3	1.4	0.21	0.47	0.24	8.58	1.06	1

Figure 3: The Training dataset.



	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH
1	insight	cause	discrep	tentat	certain	inhib	incl	excl	percept	see	hear	feel	bio	body	health	sexual	ingest	relativ	motion	space	time
2	0	1.07	1.42	0.71	1.07	1.78	1.78	1.42	0.36	0	0	0.36	2.14	0.71	1.07	0.36	0	7.12	1.07	1.78	4
3	0.16	0.09	0.02	0.02	0.11	0.11	0	0	0.09	0.05	0	0	0.29	0.02	0.27	0	0	0.74	0.02	0.5	0
4	2.23	1.49	1.86	4.09	1.12	0	3.35	2.97	0.74	0.37	0	0.37	3.72	2.6	1.12	0	0	11.9	3.35	4.46	4
5	0.17	0.14	0.06	0	0.06	0.06	0.11	0.03	0.2	0.06	0.03	0.11	0.77	0.37	0.26	0.09	0.06	2.75	0.03	1.89	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.26	0	0	5
7	0.99	0.96	0.83	1.63	1.18	0.22	1.69	2.14	1.72	0.86	0.38	0.32	3.89	1.24	0.29	2.46	0.29	6.51	0.99	2.43	2
8	0.85	0.47	0.49	0.96	0.88	0.47	2.83	1.18	2.64	0.63	1.59	0.33	0.88	0.03	0.25	0.49	0.11	12.65	1.21	6.04	5
9	1.88	1.35	1.05	1.69	1.43	0.49	2.55	1.16	2.89	1.5	0.6	0.56	0.98	0.11	0.38	0.11	0.41	13.39	1.91	6.3	5
10	0.83	0.68	1.8	1.36	1.65	0.19	2.04	1.65	1.8	0.63	0.39	0.53	1.99	0.49	0.34	0.05	1.17	13.46	1.8	5.15	6
11	0.18	0.04	0.04	0	0.11	0.07	0	0	0.18	0.07	0.07	0.04	0.18	0.14	0.04	0.04	0	0.54	0	0.21	0
12	1.01	1.33	0.66	1.62	1.13	0.4	3.21	1.5	1.73	0.95	0.38	0.38	1.53	0.46	0.2	0.46	0.43	8.46	1.16	4.02	3
13	0.53	0.79	0.6	0.75	0.75	0.06	2.14	0.66	1.32	0.47	0.38	0.38	0.66	0.25	0.16	0.09	0.16	10.24	1.04	4.81	
14	1.04	0.82	0.41	1.01	0.82	0.25	2.33	1.26	1.6	0.94	0.41	0.25	1.16	0.19	0.19	0.66	0.13	9.06	1.23	3.52	4
15	2.17	1.25	0.71	1.32	1.29	0.54	3.18	0.95	2.68	0.81	1.15	0.54	2.54	1.12	0.3	0.95	0.17	7.86	1.05	3.86	2
16	0	0	0	0	0	0.02	0.01	0.01	0.03	0.03	0	0	0.04	0	0.01	0.01	0.02	0.26	0.01	0.09	0
17	0.9	0.57	0.49	1.14	0.71	0.19	2.45	0.92	1.14	0.6	0.11	0.44	1.2	0.3	0.22	0.57	0.11	12.04	2.04	5.74	4
18	1.22	1.22	0	0	0	0	1.22	0	3.66	2.44	0	1.22	2.44	1.22	1.22	0	0	8.54	0	4.88	3
19	1.21	1.08	1.25	1.56	1.32	0.52	2.12	1.98	0.97	0.31	0.49	0.17	0.9	0.31	0.42	0.14	0.07	6.42	0.8	3.33	2
20	0.84	0.79	1.08	1.18	0.84	0.49	1.92	0.94	1.18	0.49	0.25	0.39	2.51	1.08	0.05	1.33	0.35	7.4	0.94	3.4	3
21	1.59	1.41	1.17	1.51	0.99	0.44	3.08	1.72	1.22	0.42	0.55	0.26	0.78	0.29	0.23	0.18	0.13	8.24	0.91	3.21	4
22	0	0	2.56	0	2.56	0	0	0	0	0	0	0	0	0	0	0	0	0.64	0	0	0
23	2.86	8.57	2.86	0	0	0	2.86	0	0	0	0	0	2.86	0	2.86	0	0	11.43	0	5.71	5
24	0.34	0.32	0.39	1.08	0.34	0.71	1.76	1.68	1.34	0.82	0.08	0.42	0.45	0.11	0.18	0.13	0.05	11.9	1.42	3.21	7
25	0.44	0.27	0.36	0.49	0.41	0.22	1.59	0.41	1.07	0.38	0.27	0.3	0.77	0.3	0.05	0.3	0.11	8.09	0.58	2.3	5
26	0.54	0.78	0.36	0.78	0.23	0.23	3.37	0.47	1.76	0.8	0.1	0.13	3.92	0.08	0.13	0.05	3.76	11.51	0.75	5.65	
27	1.06	1.08	1.11	1.08	0.92	0.29	2.37	1.4	2.27	1.08	0.98	0.21	1.66	0.77	0.4	0.5	0.08	8.15	0.98	3.51	5

Figure 4: The Training dataset.



	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB
1	time	work	achieve	leisure	home	money	relig	death	assent	nonfl	filler	class									
2	4.27	1.07	0	1.07	0.36	1.07	0	1.78	0	0	0	1									
3	0.23	0.79	0.5	0.11	0	0.27	0.29	0.02	0.23	0.02	0	1									
4	4.09	1.86	1.49	1.49	1.12	0.74	0	0	1.86	0	0	1									
5	0.89	0.34	0.17	0.8	0.31	0.09	0	0.26	0.26	0.26	0	1									
6	5.26	0	0	0	0	0	0	0	0	0	0	1									
7	2.94	1.09	0.8	0.51	0.1	0.32	0.16	0.1	0.61	0.16	0.19	1									
8	5.16	0.55	1.04	2.77	0.03	0.41	0.55	0.11	0.22	0.03	0.11	1									
9	5.36	1.8	1.84	2.29	0.45	1.09	0.19	0.23	0.26	0.15	0.26	1									
10	6.37	1.46	1.41	2.77	0.24	0.87	0.39	0.15	0.83	0	0.34	1									
11	0.32	0.5	0.11	0.43	0.04	0.07	0.21	0	0.07	0	0	1									
12	3.38	1.5	0.92	1.07	0.17	0.58	0.4	0.2	0.4	0.03	0.2	1									
13	4.3	1.26	1.32	2.86	0.72	0.6	0.28	0.06	0.16	0.06	0.09	1									
14	4.59	0.66	1.26	1.48	0.09	0.22	0.19	0.06	0.63	0.06	0.19	1									
15	2.74	1.25	1.73	2.78	0.1	0.41	0.51	0.14	0.3	0.17	0	1									
16	0.16	0.09	0.01	0.08	0.02	0.08	0	0	0	0	0	1									
17	4.24	1.2	1.82	1.28	0.44	0.49	0.16	0	0.71	0.08	0.14	1									
18	3.66	2.44	1.22	3.66	0	2.44	0	0	1.22	0	0	1									
19	2.29	1.7	1.01	0.21	0.17	0.87	0.9	0.45	0.66	0.21	0.24	1									
20	3.21	0.99	1.28	1.08	0.05	0.15	0.74	0.25	0.89	0	0.15	1									
21	4.33	0.89	1.75	0.94	0.13	0.83	0.39	0.18	0.16	0.08	0.13	1									
22	0.64	0	0	0.64	0	0	0	0	0	0	0	1									
23	5.71	0	8.57	0	0	0	0	0	0	0	0	1									
24	7.55	1.84	2.5	2.13	0.08	0.24	0.05	0	0.18	0.05	0.08	1									
25	5.15	0.69	1.12	1.07	0.16	0.16	0.33	0.05	0.25	0.05	0.11	1									
26	4.9	1.43	1.17	3.35	0.23	0.88	0.05	0.03	0.08	0	0.13	1									
27	3.72	1.66	1.24	1.61	0.16	0.95	0.29	0.11	0.32	0.05	0.18	1									

Figure 5: The Training Dataset

Then we use leaps package. By regsubsets() function we can select best sub-set by identifying the best model which contains a given number of predictors, where **best** is quantified using RSS. And summary() gives the best set of variables for each model size.

```
> library(leaps)
> library(MASS)
> #for nvmax=4
>  reg1<-regsubsets(class~WC+WPS+Qmarks+Unique+Dic+Sixltr+funct+pronoun+ppron+i+we
+you+shehe+they+ipron+article+verb+auxverb+past+present+future+adverb+preps+conj+negat
e+quant+number+swear+social+family+friend+humans+affect+posemo+negemo+anx+anger+sa
d+cogmech+insight+cause+discrep+tentat+certain+inhib+incl+excl+percept+see+hear+feel+bio
+body+health+sexual+ingest+relativ+motion+space+time+work+achieve+leisure+home+money
+relig+death+assent+nonfl+filler,data=data1,nbest=1,nvmax=4,really.big=TRUE)
> summary(reg1)
```

Here **data** is our data frame, **nvmax** is the maximum size of subsets to examine, **nbest** is the number of subsets of each size to record, **really.big = T** to perform exhaustive search because in our dataset there are more than 50 variables. For describing our data we choose linear regression model where we follow the best subsets regression approach. We have run for nvmax 4, 5,6,7,8,9,10. The more we increase the nvmax value, the more features we get.

Nvmax	Count	Features
4	7	they,work,funct,I,leisure, ppron,past
5	7	they,work, funct,I,leisure, ppron,past
6	8	they,work, funct,I,leisure, ppron,past, relig
7	8	they,work, funct,I,leisure, ppron,past,relig
8	11	they,work, funct,I,leisure,ppron,past,relig,conj,quant,incl.
9	19	they,work,funct,I,leisure,ppron,past,relig,conj,quant,incl,Dic,we,verb,social,body,time,achieve,home
10	20	they,work,funct,I,leisure,ppron,past,relig,conj,quant,incl,Dic,we,verb,social,body,time,achieve,home,friend

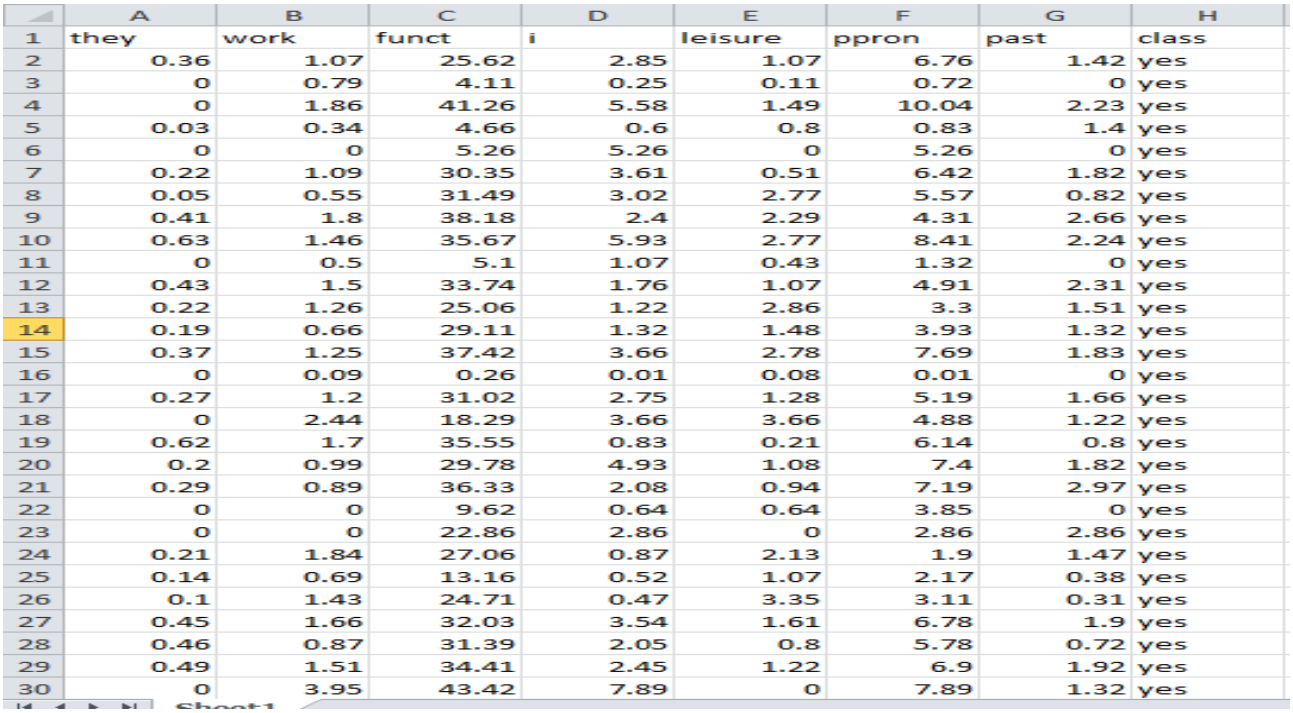
Table 1: Features count and list for different nvmax value.

One of the Outputs is shown below, **for nvmax=10,**

1 subsets of each size up to 10

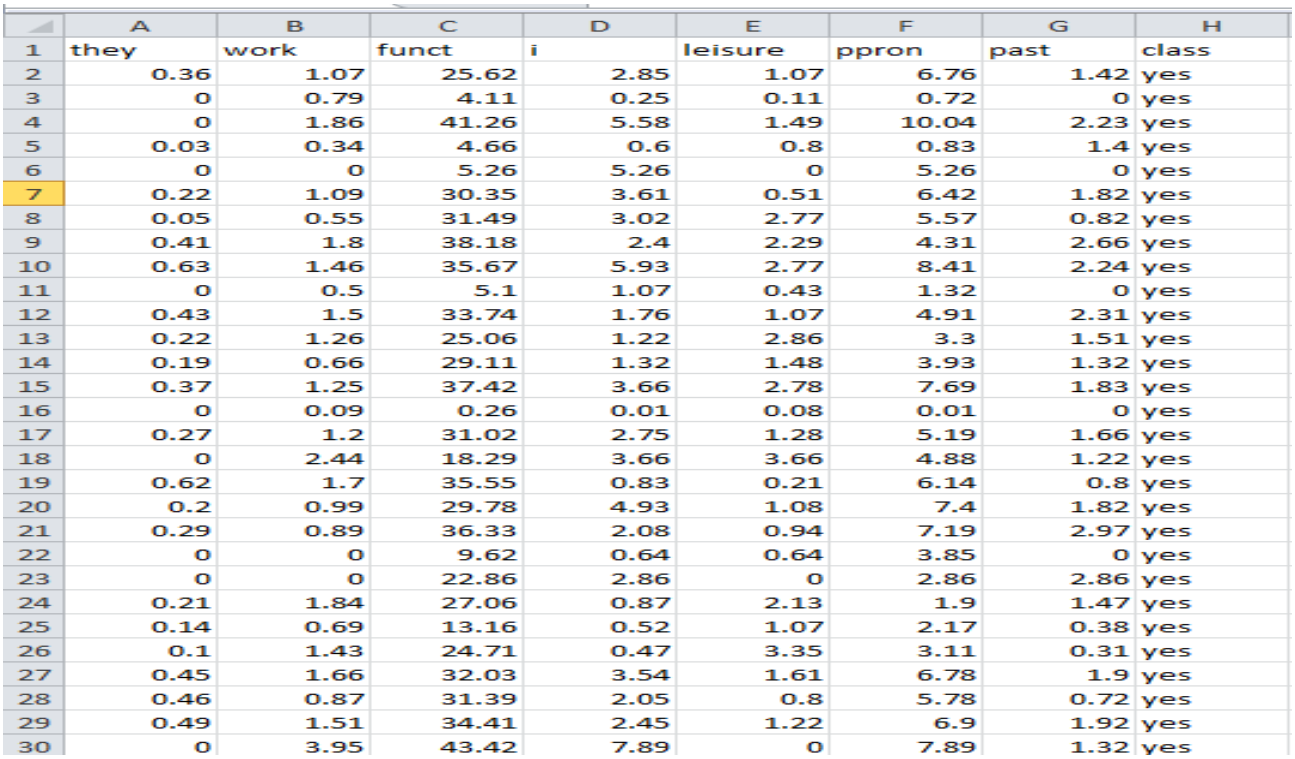
```
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          WC  WPS Qmarks Unique Dic Sixltr funct pronoun ppron i   we  you shehe they ipron article verb auxverb
1  ( 1 )  " " " " " " "     " "    " " " " "     " "    " "    " " " " " " " " " " " "     " "  " "     " "      " "  " "
2  ( 1 )  " " " " " " "     " "    " " " " "     " "    " "    " " " " " " " " " " " "     "*"  " "     " "      " "  " "
3  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    " " "*" " " " " " " " " "     " "  " "     " "      " "  " "
4  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    "*" " " " " " " " " " " "     " "  " "     " "      " "  " "
5  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    "*" " " " " " " " " " " "     "*"  " "     " "      " "  " "
6  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    "*" " " " " " " " " " " "     "*"  " "     " "      " "  " "
7  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    "*" " " " " " " " " " " "     " "  " "     " "      " "  " "
8  ( 1 )  " " " " " " "     " "    " " " " "     "*"    " "    "*" " " " " " " " " " " "     " "  " "     " "      " "  " "
9  ( 1 )  " " " " " " "     " "    "*" " " "     " "    " "    " " " " "*" " " " " " " "     " "  " "     " "      "*"  " "
10  ( 1 ) " " " " " " "     " "    "*" " " "     " "    " "    " " " " "*" " " " " " " "     " "  " "     " "      "*"  " "
          past present future adverb preps conj negate quant number swear social family friend humans affect posemo
1  ( 1 )  " "  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
2  ( 1 )  " "  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
3  ( 1 )  " "  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
4  ( 1 )  "*"  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
5  ( 1 )  "*"  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
6  ( 1 )  "*"  " "      " "    " "    " "   " "  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
7  ( 1 )  "*"  " "      " "    " "    " "   "*"  " "    " "   " "    " "   " "    " "    " "    " "    " "    " "    " "
8  ( 1 )  "*"  " "      " "    " "    " "   "*"  " "    "*"   " "    " "   " "    " "    " "    " "    " "    " "    " "
9  ( 1 )  " "  " "      " "    " "    " "   " "  " "    " "   " "    " "   "*"    " "    " "    "*"    " "    " "    " "
10  ( 1 ) " "  " "      " "    " "    " "   " "  " "    " "   " "    " "   "*"    " "    " "    "*"    " "    " "    " "

          negemo anx anger sad cogmech insight cause discrep tentat certain inhib incl excl percept see hear feel
1  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
2  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
3  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
4  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
5  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
6  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
7  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  "*"  " "  " "     " " " " " "  " "
8  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  "*"  " "  " "     " " " " " "  " "
9  ( 1 )  " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
10  ( 1 ) " "    " " " "   " " " "     " "     " "   " "     " "     " "    " "  " "  " "  " "     " " " " " "  " "
          bio body health sexual ingest relativ motion space time work achieve leisure home money relig death
1  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  "*"  " "     " "     " "  " "   " "   " "
2  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  "*"  " "     " "     " "  " "   " "   " "
3  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   " "   " "
4  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   " "   " "
5  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   " "   " "
6  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   "*"   " "
7  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   "*"   " "
8  ( 1 )  " " " "  " "    " "    " "    " "    " "    " "  " "  " "     "*"     " "  " "   "*"   " "
9  ( 1 )  " " "*"  " "    " "    " "    " "    " "    "*"  " "  "*"     "*"     "*"  " "   " "   " "
10  ( 1 ) " " "*"  " "    " "    " "    " "    " "    "*"  " "  "*"     "*"     "*"  " "   " "   " "
          assent nonfl filler
1  ( 1 )  " "    " "   " "
2  ( 1 )  " "    " "   " "
3  ( 1 )  " "    " "   " "
4  ( 1 )  " "    " "   " "
5  ( 1 )  " "    " "   " "
6  ( 1 )  " "    " "   " "
7  ( 1 )  " "    " "   " "
8  ( 1 )  " "    " "   " "
9  ( 1 )  " "    " "   " "
10  ( 1 ) " "    " "   " "
```



	A	B	C	D	E	F	G	H
1	they	work	funct	i	leisure	ppron	past	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	yes
6	0	0	5.26	5.26	0	5.26	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	yes
22	0	0	9.62	0.64	0.64	3.85	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	yes

Figure 6: NVMAX=4 output CSV.



	A	B	C	D	E	F	G	H
1	they	work	funct	i	leisure	ppron	past	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	yes
6	0	0	5.26	5.26	0	5.26	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	yes
22	0	0	9.62	0.64	0.64	3.85	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	yes

Figure 9: NVMAX=7 output CSV.

	A	B	C	D	E	F	G	H	I
1	they	work	funct	i	leisure	ppron	past	relig	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	0	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	0.29	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	0	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	0	yes
6	0	0	5.26	5.26	0	5.26	0	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	0.16	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	0.55	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	0.19	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	0.39	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	0.21	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	0.4	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	0.28	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	0.19	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	0.51	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	0	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	0.16	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	0	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	0.9	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	0.74	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	0.39	yes
22	0	0	9.62	0.64	0.64	3.85	0	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	0	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	0.05	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	0.33	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	0.05	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	0.29	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	2.2	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	0.44	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	0	yes

Sheet1

Figure 8: NVMAX=6 output CSV.

	A	B	C	D	E	F	G	H	I
1	they	work	funct	i	leisure	ppron	past	relig	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	0	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	0.29	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	0	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	0	yes
6	0	0	5.26	5.26	0	5.26	0	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	0.16	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	0.55	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	0.19	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	0.39	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	0.21	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	0.4	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	0.28	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	0.19	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	0.51	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	0	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	0.16	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	0	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	0.9	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	0.74	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	0.39	yes
22	0	0	9.62	0.64	0.64	3.85	0	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	0	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	0.05	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	0.33	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	0.05	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	0.29	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	2.2	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	0.44	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	0	yes

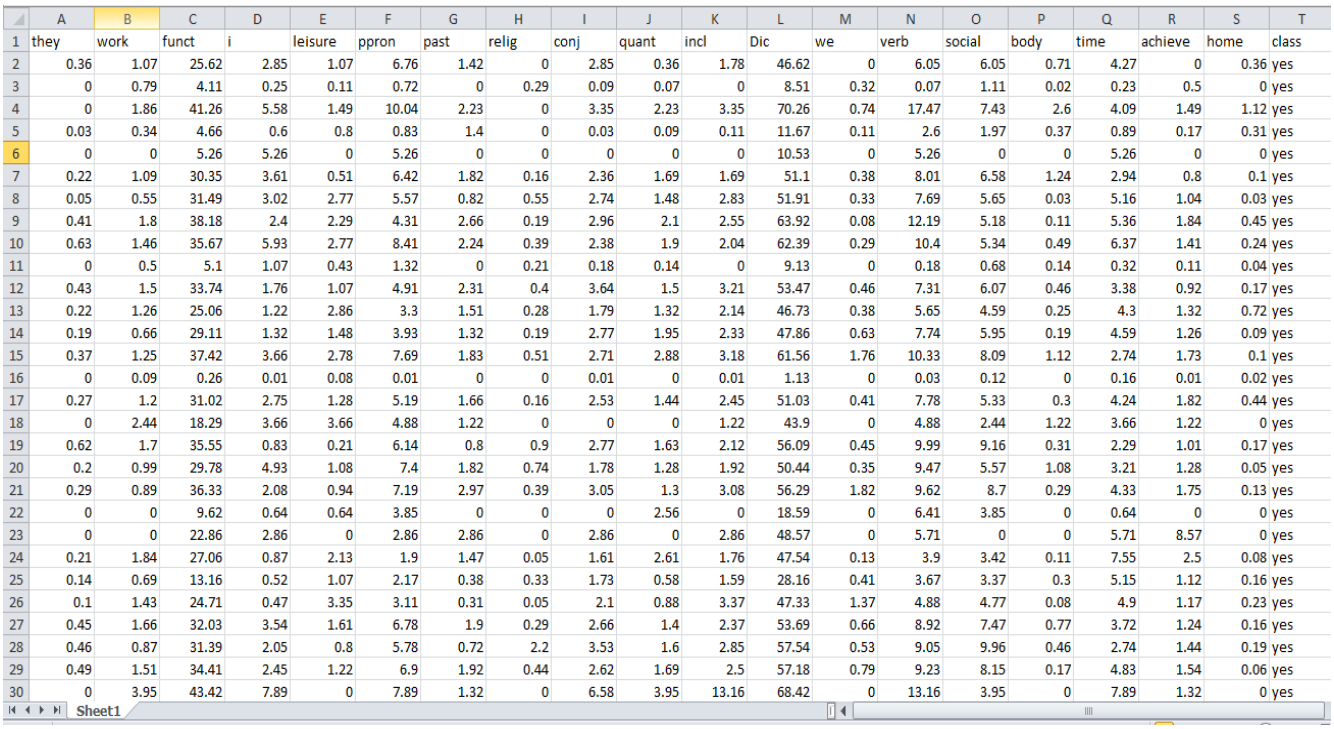
Sheet1

Figure 9: NVMAX=7 output CSV.

	A	B	C	D	E	F	G	H	I	J	K	L
1	they	work	funct	i	leisure	ppron	past	relig	conj	quant	incl	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	0	2.85	0.36	1.78	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	0.29	0.09	0.07	0	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	0	3.35	2.23	3.35	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	0	0.03	0.09	0.11	yes
6	0	0	5.26	5.26	0	5.26	0	0	0	0	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	0.16	2.36	1.69	1.69	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	0.55	2.74	1.48	2.83	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	0.19	2.96	2.1	2.55	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	0.39	2.38	1.9	2.04	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	0.21	0.18	0.14	0	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	0.4	3.64	1.5	3.21	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	0.28	1.79	1.32	2.14	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	0.19	2.77	1.95	2.33	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	0.51	2.71	2.88	3.18	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	0	0.01	0	0.01	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	0.16	2.53	1.44	2.45	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	0	0	0	1.22	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	0.9	2.77	1.63	2.12	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	0.74	1.78	1.28	1.92	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	0.39	3.05	1.3	3.08	yes
22	0	0	9.62	0.64	0.64	3.85	0	0	0	2.56	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	0	2.86	0	2.86	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	0.05	1.61	2.61	1.76	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	0.33	1.73	0.58	1.59	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	0.05	2.1	0.88	3.37	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	0.29	2.66	1.4	2.37	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	2.2	3.53	1.6	2.85	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	0.44	2.62	1.69	2.5	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	0	6.58	3.95	13.16	yes

Sheet1

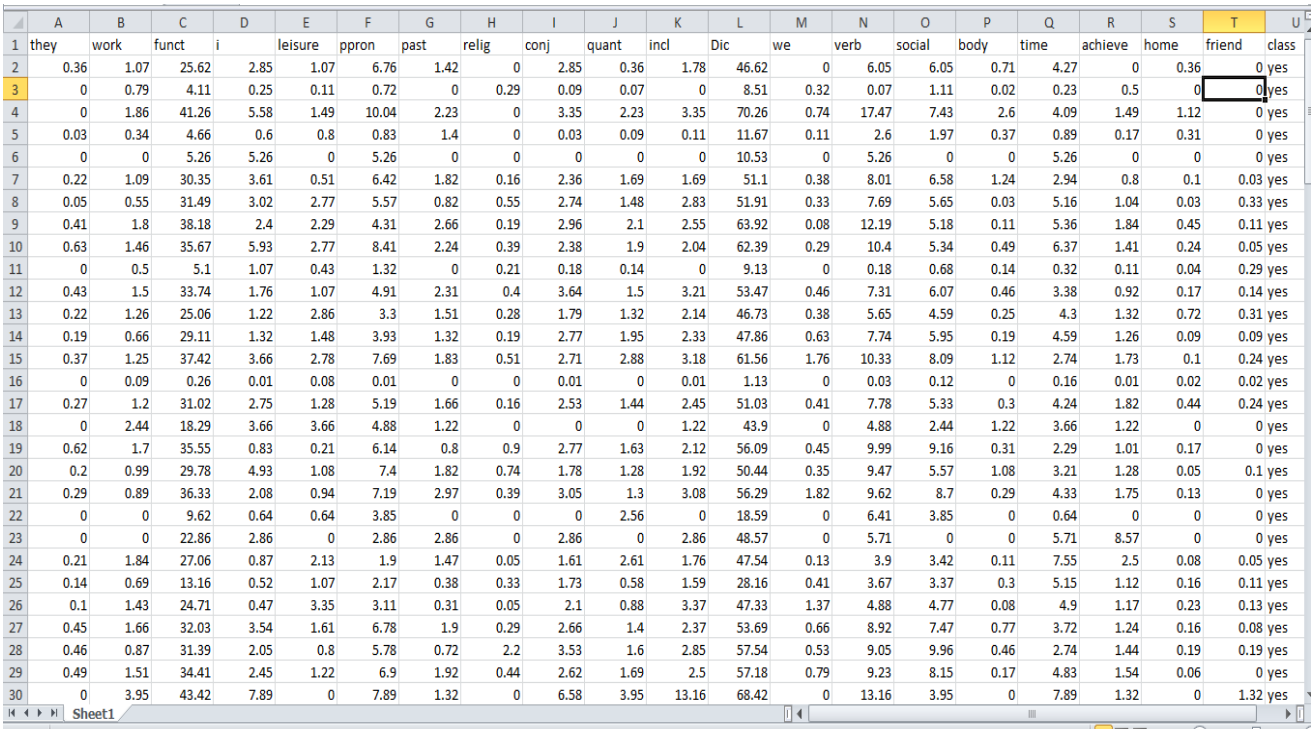
Figure 10: NVMAX=8 output CSV.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	they	work	funct	i	leisure	ppron	past	relig	conj	quant	incl	Dic	we	verb	social	body	time	achieve	home	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	0	2.85	0.36	1.78	46.62	0	6.05	6.05	0.71	4.27	0	0.36	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	0.29	0.09	0.07	0	8.51	0.32	0.07	1.11	0.02	0.23	0.5	0	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	0	3.35	2.23	3.35	70.26	0.74	17.47	7.43	2.6	4.09	1.49	1.12	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	0	0.03	0.09	0.11	11.67	0.11	2.6	1.97	0.37	0.89	0.17	0.31	yes
6	0	0	5.26	5.26	0	5.26	0	0	0	0	0	10.53	0	5.26	0	0	5.26	0	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	0.16	2.36	1.69	1.69	51.1	0.38	8.01	6.58	1.24	2.94	0.8	0.1	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	0.55	2.74	1.48	2.83	51.91	0.33	7.69	5.65	0.03	5.16	1.04	0.03	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	0.19	2.96	2.1	2.55	63.92	0.08	12.19	5.18	0.11	5.36	1.84	0.45	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	0.39	2.38	1.9	2.04	62.39	0.29	10.4	5.34	0.49	6.37	1.41	0.24	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	0.21	0.18	0.14	0	9.13	0	0.18	0.68	0.14	0.32	0.11	0.04	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	0.4	3.64	1.5	3.21	53.47	0.46	7.31	6.07	0.46	3.38	0.92	0.17	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	0.28	1.79	1.32	2.14	46.73	0.38	5.65	4.59	0.25	4.3	1.32	0.72	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	0.19	2.77	1.95	2.33	47.86	0.63	7.74	5.95	0.19	4.59	1.26	0.09	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	0.51	2.71	2.88	3.18	61.56	1.76	10.33	8.09	1.12	2.74	1.73	0.1	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	0	0.01	0	0.01	1.13	0	0.03	0.12	0	0.16	0.01	0.02	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	0.16	2.53	1.44	2.45	51.03	0.41	7.78	5.33	0.3	4.24	1.82	0.44	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	0	0	0	1.22	43.9	0	4.88	2.44	1.22	3.66	1.22	0	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	0.9	2.77	1.63	2.12	56.09	0.45	9.99	9.16	0.31	2.29	1.01	0.17	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	0.74	1.78	1.28	1.92	50.44	0.35	9.47	5.57	1.08	3.21	1.28	0.05	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	0.39	3.05	1.3	3.08	56.29	1.82	9.62	8.7	0.29	4.33	1.75	0.13	yes
22	0	0	9.62	0.64	0.64	3.85	0	0	0	2.56	0	18.59	0	6.41	3.85	0	0.64	0	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	0	2.86	0	2.86	48.57	0	5.71	0	0	5.71	8.57	0	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	0.05	1.61	2.61	1.76	47.54	0.13	3.9	3.42	0.11	7.55	2.5	0.08	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	0.33	1.73	0.58	1.59	28.16	0.41	3.67	3.37	0.3	5.15	1.12	0.16	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	0.05	2.1	0.88	3.37	47.33	1.37	4.88	4.77	0.08	4.9	1.17	0.23	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	0.29	2.66	1.4	2.37	53.69	0.66	8.92	7.47	0.77	3.72	1.24	0.16	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	2.2	3.53	1.6	2.85	57.54	0.53	9.05	9.96	0.46	2.74	1.44	0.19	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	0.44	2.62	1.69	2.5	57.18	0.79	9.23	8.15	0.17	4.83	1.54	0.06	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	0	6.58	3.95	13.16	68.42	0	13.16	3.95	0	7.89	1.32	0	yes

Sheet1

Figure 11: NVMAX=9 output CSV.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	they	work	funct	i	leisure	ppron	past	relig	conj	quant	incl	Dic	we	verb	social	body	time	achieve	home	friend	class
2	0.36	1.07	25.62	2.85	1.07	6.76	1.42	0	2.85	0.36	1.78	46.62	0	6.05	6.05	0.71	4.27	0	0.36	0	yes
3	0	0.79	4.11	0.25	0.11	0.72	0	0.29	0.09	0.07	0	8.51	0.32	0.07	1.11	0.02	0.23	0.5	0	0	yes
4	0	1.86	41.26	5.58	1.49	10.04	2.23	0	3.35	2.23	3.35	70.26	0.74	17.47	7.43	2.6	4.09	1.49	1.12	0	yes
5	0.03	0.34	4.66	0.6	0.8	0.83	1.4	0	0.03	0.09	0.11	11.67	0.11	2.6	1.97	0.37	0.89	0.17	0.31	0	yes
6	0	0	5.26	5.26	0	5.26	0	0	0	0	0	10.53	0	5.26	0	0	5.26	0	0	0	yes
7	0.22	1.09	30.35	3.61	0.51	6.42	1.82	0.16	2.36	1.69	1.69	51.1	0.38	8.01	6.58	1.24	2.94	0.8	0.1	0.03	yes
8	0.05	0.55	31.49	3.02	2.77	5.57	0.82	0.55	2.74	1.48	2.83	51.91	0.33	7.69	5.65	0.03	5.16	1.04	0.03	0.33	yes
9	0.41	1.8	38.18	2.4	2.29	4.31	2.66	0.19	2.96	2.1	2.55	63.92	0.08	12.19	5.18	0.11	5.36	1.84	0.45	0.11	yes
10	0.63	1.46	35.67	5.93	2.77	8.41	2.24	0.39	2.38	1.9	2.04	62.39	0.29	10.4	5.34	0.49	6.37	1.41	0.24	0.05	yes
11	0	0.5	5.1	1.07	0.43	1.32	0	0.21	0.18	0.14	0	9.13	0	0.18	0.68	0.14	0.32	0.11	0.04	0.29	yes
12	0.43	1.5	33.74	1.76	1.07	4.91	2.31	0.4	3.64	1.5	3.21	53.47	0.46	7.31	6.07	0.46	3.38	0.92	0.17	0.14	yes
13	0.22	1.26	25.06	1.22	2.86	3.3	1.51	0.28	1.79	1.32	2.14	46.73	0.38	5.65	4.59	0.25	4.3	1.32	0.72	0.31	yes
14	0.19	0.66	29.11	1.32	1.48	3.93	1.32	0.19	2.77	1.95	2.33	47.86	0.63	7.74	5.95	0.19	4.59	1.26	0.09	0.09	yes
15	0.37	1.25	37.42	3.66	2.78	7.69	1.83	0.51	2.71	2.88	3.18	61.56	1.76	10.33	8.09	1.12	2.74	1.73	0.1	0.24	yes
16	0	0.09	0.26	0.01	0.08	0.01	0	0	0.01	0	0.01	1.13	0	0.03	0.12	0	0.16	0.01	0.02	0.02	yes
17	0.27	1.2	31.02	2.75	1.28	5.19	1.66	0.16	2.53	1.44	2.45	51.03	0.41	7.78	5.33	0.3	4.24	1.82	0.44	0.24	yes
18	0	2.44	18.29	3.66	3.66	4.88	1.22	0	0	0	1.22	43.9	0	4.88	2.44	1.22	3.66	1.22	0	0	yes
19	0.62	1.7	35.55	0.83	0.21	6.14	0.8	0.9	2.77	1.63	2.12	56.09	0.45	9.99	9.16	0.31	2.29	1.01	0.17	0	yes
20	0.2	0.99	29.78	4.93	1.08	7.4	1.82	0.74	1.78	1.28	1.92	50.44	0.35	9.47	5.57	1.08	3.21	1.28	0.05	0.1	yes
21	0.29	0.89	36.33	2.08	0.94	7.19	2.97	0.39	3.05	1.3	3.08	56.29	1.82	9.62	8.7	0.29	4.33	1.75	0.13	0	yes
22	0	0	9.62	0.64	0.64	3.85	0	0	0	2.56	0	18.59	0	6.41	3.85	0	0.64	0	0	0	yes
23	0	0	22.86	2.86	0	2.86	2.86	0	2.86	0	2.86	48.57	0	5.71	0	0	5.71	8.57	0	0	yes
24	0.21	1.84	27.06	0.87	2.13	1.9	1.47	0.05	1.61	2.61	1.76	47.54	0.13	3.9	3.42	0.11	7.55	2.5	0.08	0.05	yes
25	0.14	0.69	13.16	0.52	1.07	2.17	0.38	0.33	1.73	0.58	1.59	28.16	0.41	3.67	3.37	0.3	5.15	1.12	0.16	0.11	yes
26	0.1	1.43	24.71	0.47	3.35	3.11	0.31	0.05	2.1	0.88	3.37	47.33	1.37	4.88	4.77	0.08	4.9	1.17	0.23	0.13	yes
27	0.45	1.66	32.03	3.54	1.61	6.78	1.9	0.29	2.66	1.4	2.37	53.69	0.66	8.92	7.47	0.77	3.72	1.24	0.16	0.08	yes
28	0.46	0.87	31.39	2.05	0.8	5.78	0.72	2.2	3.53	1.6	2.85	57.54	0.53	9.05	9.96	0.46	2.74	1.44	0.19	0.19	yes
29	0.49	1.51	34.41	2.45	1.22	6.9	1.92	0.44	2.62	1.69	2.5	57.18	0.79	9.23	8.15	0.17	4.83	1.54	0.06	0	yes
30	0	3.95	43.42	7.89	0	7.89	1.32	0	6.58	3.95	13.16	68.42	0	13.16	3.95	0	7.89	1.32	0	1.32	yes

Sheet1

Figure 12: NVMAX=10 output CSV.

3.5 Classification Algorithms

We have used Weka Software tool. We have used some classification algorithms for check the accuracy of the prediction on the dataset which contains the most features (nvmax=10). The classification algorithms we use are Naïve Bayes, IBK(K-NN), Support Vector Machines (SVM), J48, and Random forest on the training dataset. The Outputs after compiling on the dataset have shown below,

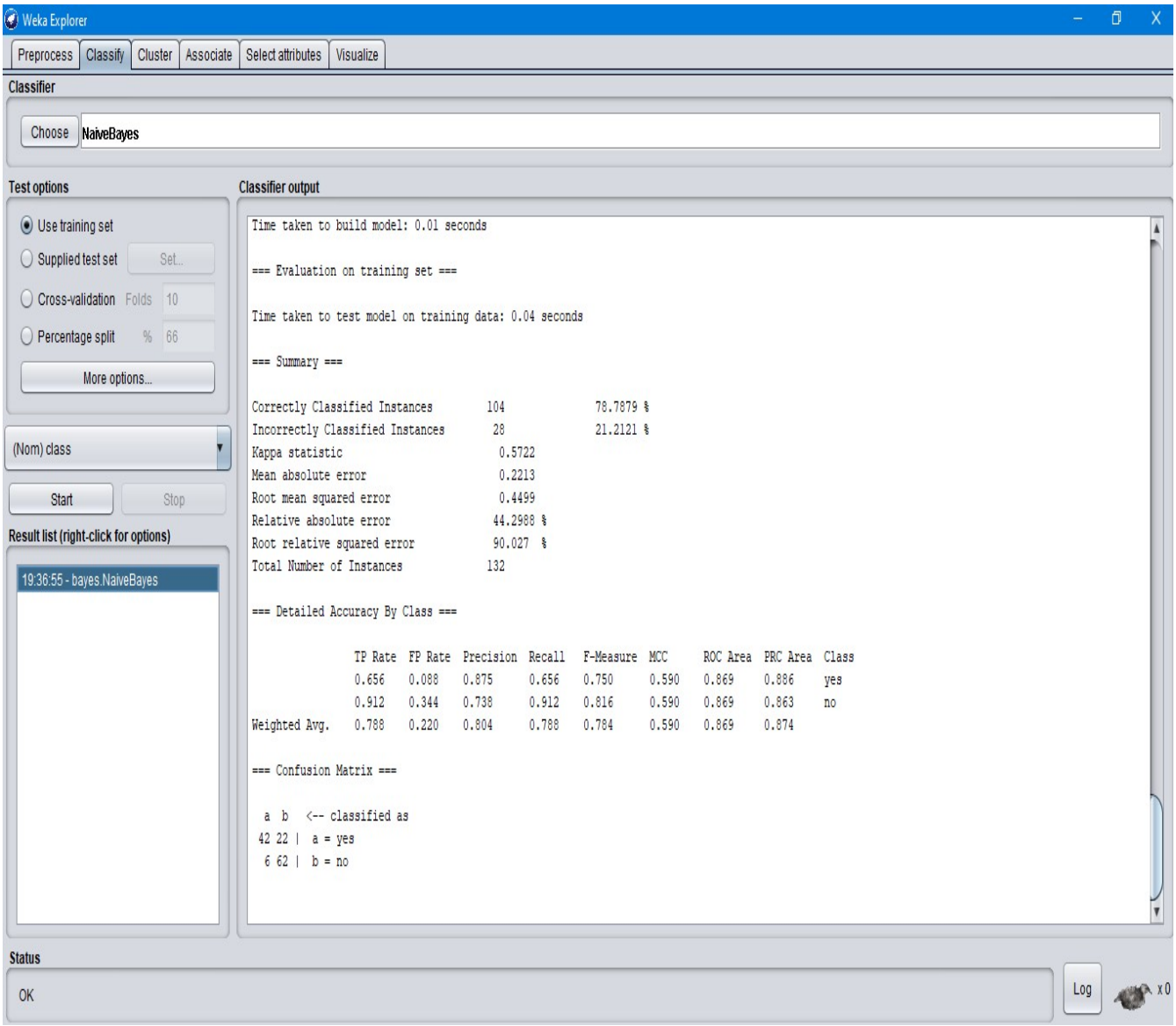


Figure 13: Naïve Bayes Classifier.

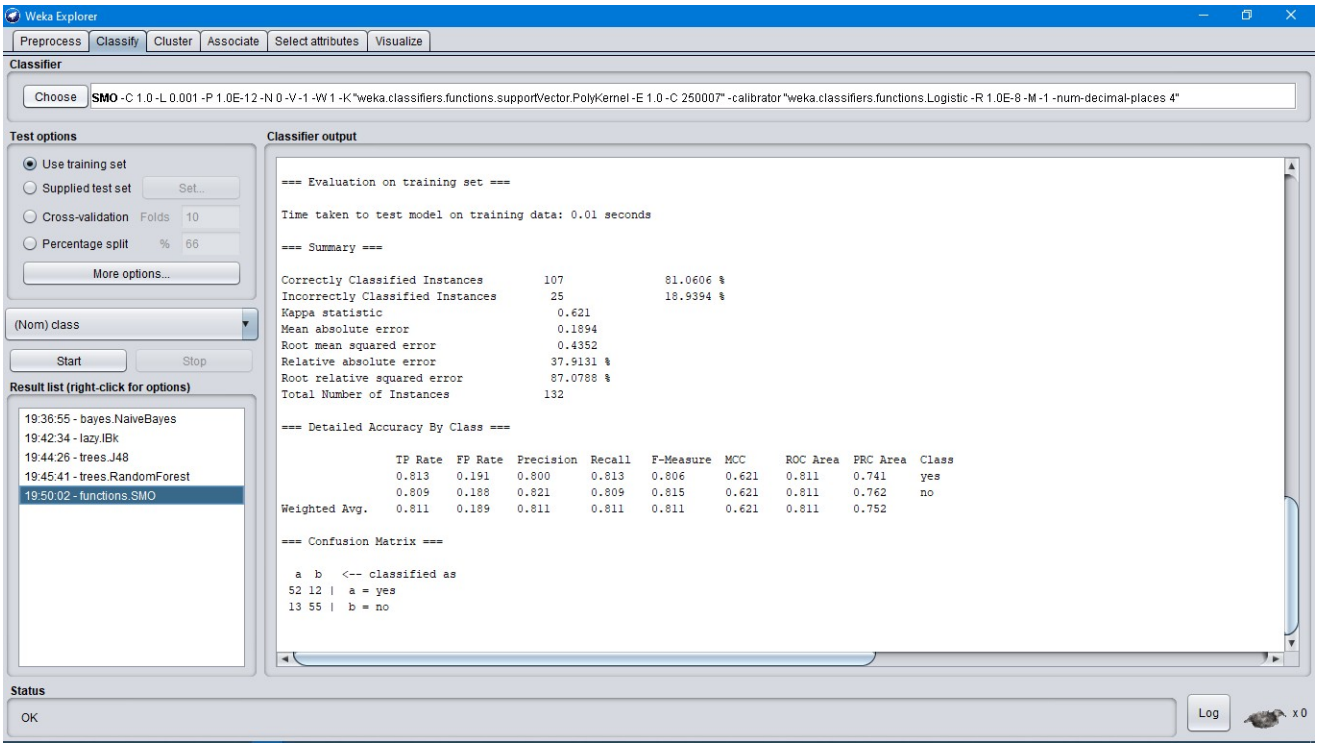


Figure 14: Support Vector Machines (SVM) Classifier.

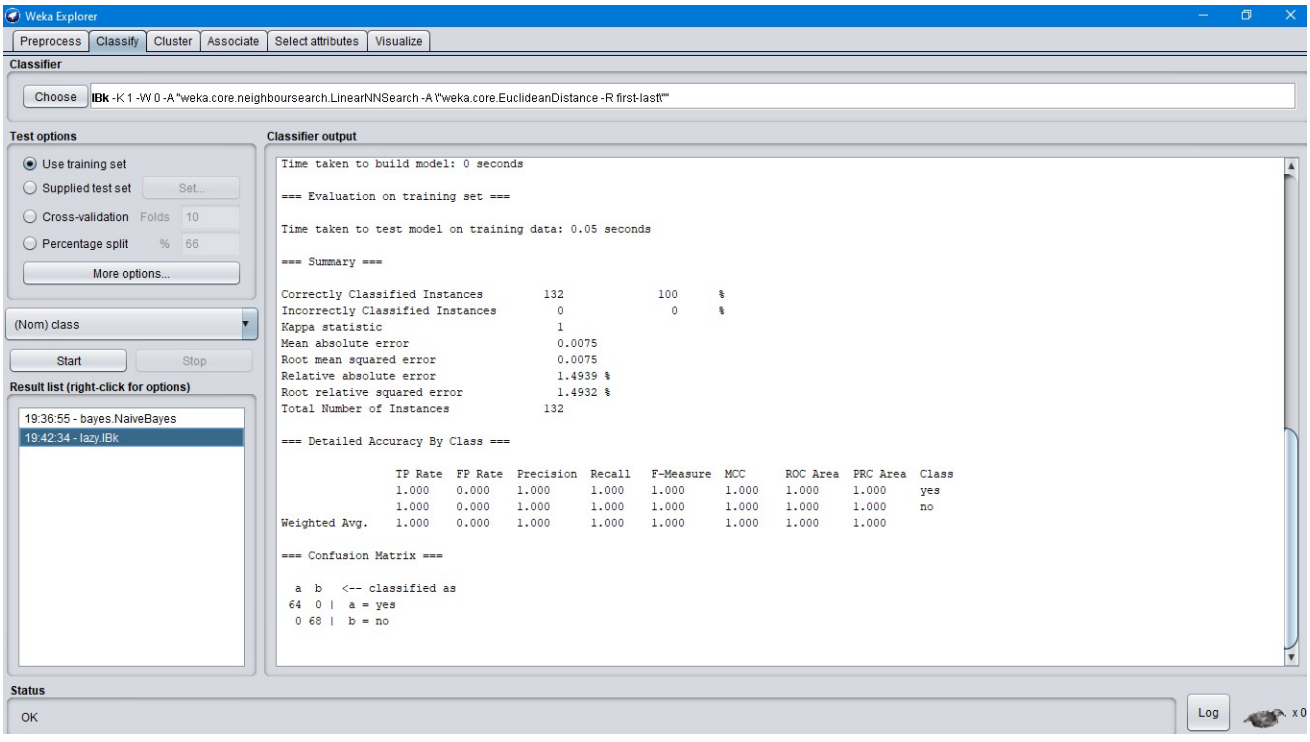


Figure 15: IBk (K-NN) Classifier.

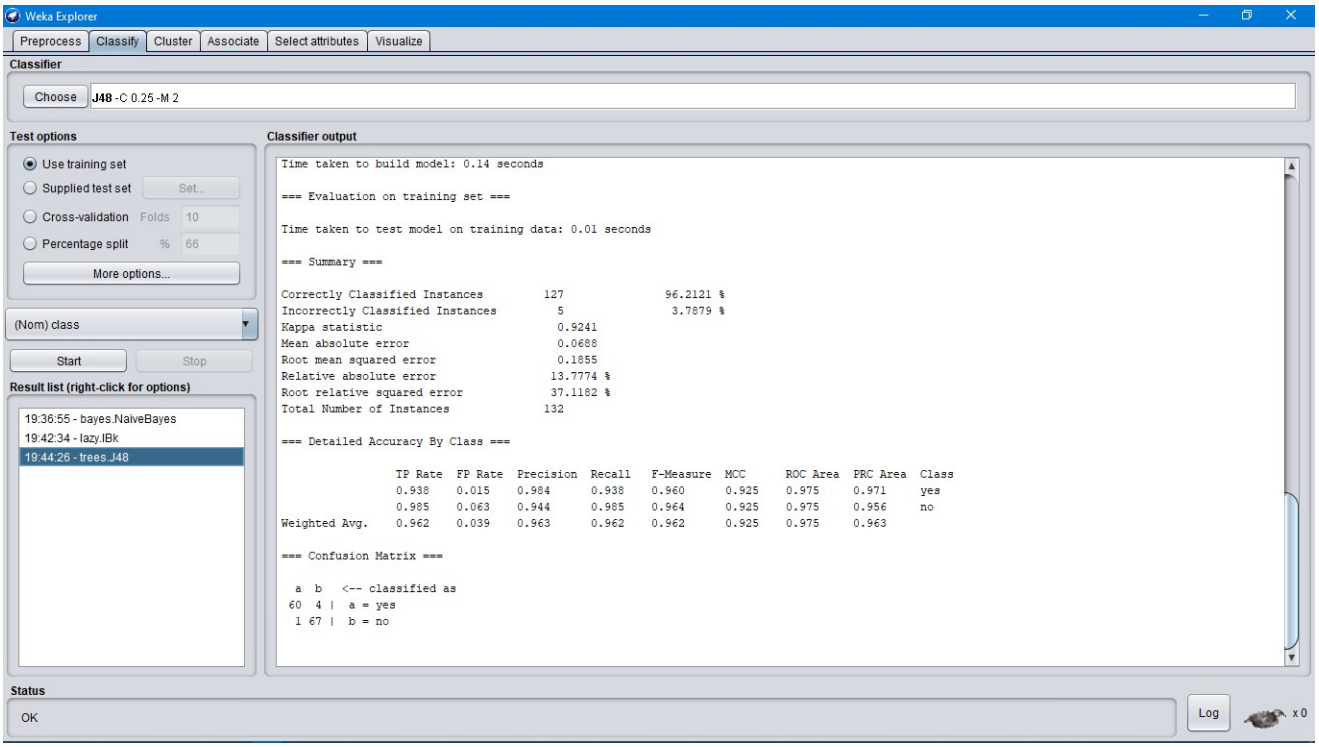


Figure 16: J48 Classifier.

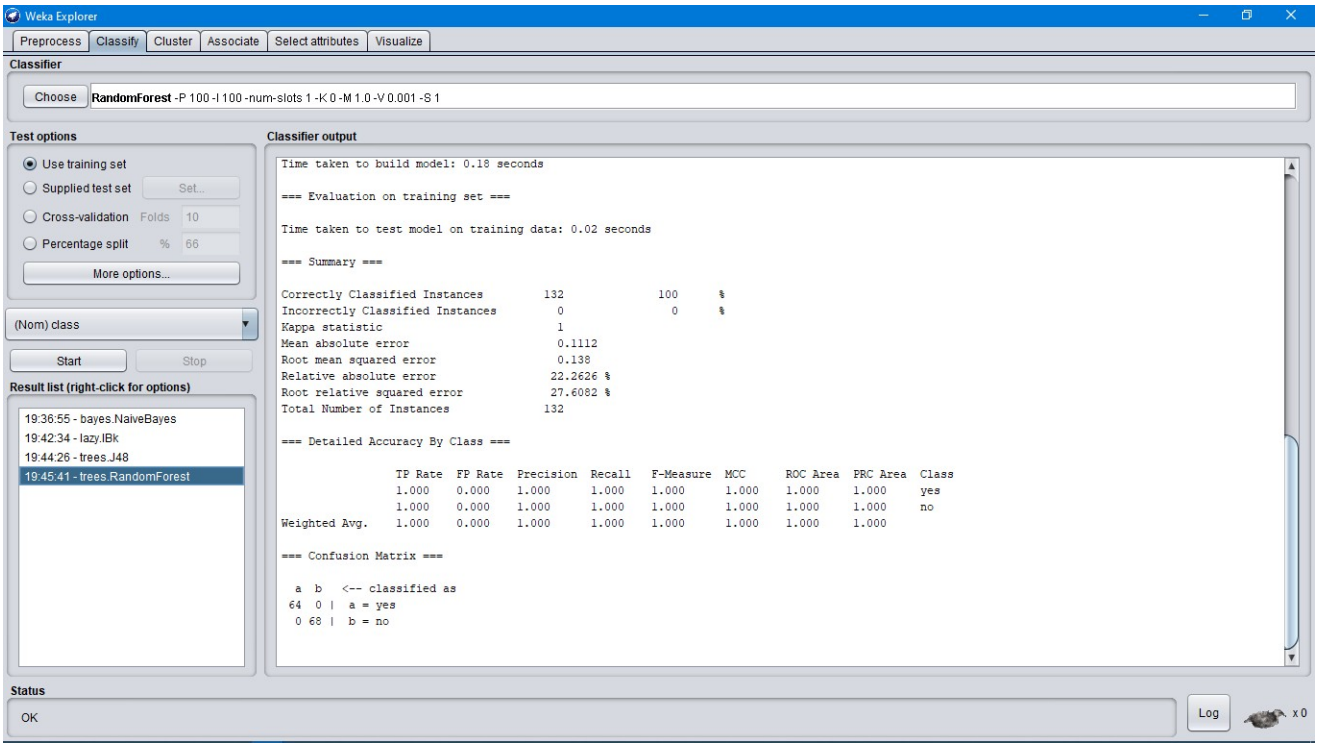


Figure 17: Random Forest Classifier.

Chapter 4

4.Result and Descriptive Analysis

4.1 Result Analysis

For nvmax=10, we get the highest 20 features which are **"**They, work, funct, I, leisure, ppron, past, relig, conj, quant, incl, Dic, we, verb, social, body, time, achieve, home, friend". These variables are mainly responsible psychologically for suicide. Some description of these features are given bellow -

they (They): It defines a third person in a grammatical point of view.

work (Work): It means doing something.

funct(Function Words): It indicates Informal, social ,Honest, depressed, low status, personal, etc.

i(I) : It defines a first person in a grammatical point of view.

Leisure(leisure): It defines spend lazy time.

ppron (Personal pronouns): This defines a grammatical person example: first , second and third person.

past (Past):It defines past events or incidences.

relig (Religion): It means church, mosque etc.

conj (Conjunctions'):It indicates conjunctions. example: And, but, for, nor, or, so, and yet.

quant (Quantifiers): Defines quantity such as some, many, few, a lot and no.

incl (Inclusive) : Example : And, with, include etc.

Dic: It includes Dictionary.
we: Defines grammatical third person WE.

verb (Verbs) : The presence of all activities, all events or all conditions is included in this variable.

social (Social): These are social terms used to refer to social activity and relationships. This is due to social care, social support.
body (Body): Different parts of a human body.

time (Space): End, until, season.

achieve (Achievement): Make money ,hero, or win

home (Home): It defines Apartment, kitchen, family etc.

friend(Friend) : These are social terms used to communicate social activities and relationships. Dude, this word is related to social problems.

There many classification algorithms in Weka. Above in chapter 3, we applied five classification algorithms on the dataset for testing and analyzing which are Naïve Bayes (NB), SVM, IBk (K-NN), J48, Random Forest. These classification models have ensured to test the correctness of the training dataset. These classification algorithms are applied on the dataset to predict the specific reason behind the suicidal case.

Naïve Bayes Classifier: Naïve Bayes Classifier is a machine learning classifier and very simple. It is also called as simple Bayes or independence Bayes. Bayes' theorem is used for classifying objects in naive Bayes classifier. Spam filters, text analysis etc are some famous use of this algorithm.

Support Vector Machines: SVM is an algorithm and its concepts are literally simple. A hyperplane is used in this classifier to separate data points with the margin of largest amount. For this, it is also called discriminative classifier. SVM works to find a line that can separate the particular data into two group and for this it uses an optimization process. It considers those variables or data instances in training dataset that are the most closest to the line that can separate the classes in best way. These instances are also known as support vectors.

IBk (K-NN) Classifier: In supervised learning, K-NN is one of the most well-known classifications. K-NN is used for problems of both classification and regression. But it is extensively used for classification. It stores all cases which are available in the training dataset is a non-parametric method. Majority of vote its k neighbors is used to classifies new cases. In testing, a new case is assigned which is most common in its K nearest neighbors to the class. In this case, it is measured by a distance function. These functions are Euclidea, Manhattan, Minkowski, and Hamming distance. Euclidea, Manhattan, and Minkowski are used for continuous variables. The last one Hamming distance is used for categorical variables.

Random Forest Classifier: It is the most popular method of supervised learning. It is well known for its classification and regression problems. Decision Trees are the ensemble of this algorithm. It is trained via the bagging method and this is the reason why it introduces randomness and diversity and lower variance as well as. It has several types of techniques like the Gini Index, Information Gain as the splitting criterion for splitting the feature nodes. In the testing, each of the trees individually can predict, and finally, it chooses the best classifier with the highest vote. This classifier is most useful for determining the feature importance.

J48 Classifier: The main use of this classifier is generating decision tree which is based on C4.5 algorithm. This algorithm is developed by Ross Quinlan. It is an extension of an algorithm which is Quinlan's earlier ID3. C4.5 can be used for classification which is used for generating decision trees. It is mentioned as a statistical classifier.

This research has records of 132 instances from twitter using API. For these 132 instances this research has found 71 variables. From this 71 variables it has selected only 20 variables which are mostly affected in this research and these instances make training dataset for this research. To check the accuracy of the prediction this research have used mentioned algorithms on the training dataset and it has got the percentage of accuracy which are shown in below.

Classifier	Accuracy	Number of Correctly Classified Instances	Number of Incorrectly Classified Instances
Naïve Bayes Classifier	78.7879 %	104	28
SVM	81.0606 %	107	25
IBk (K-NN)	100 %	132	0
J48	96.2121 %	127	5
Random Forest	100 %	132	0

Table 2: Based on Accuracy The Comparison of Different Classifiers.

From table 2 it is clarified that the five classification algorithms perform with their high accuracy. Among all of them, IBK (K-NN) and Random Forest have the highest accuracy with 100% and they have no incorrectly classified instances. Here J48 has the second-highest accuracy with 96.2121% and SVM have the third-highest accuracy. With an accuracy of 78.7879% the bottom of the list is Naïve Bayes. Because of the same accuracy error measurement is going to use for selecting the best classification model.

Here this research has two classifiers that have the same position with the highest accuracy of 100%. Values of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) are going to use to compare between IBK (K-NN) and Random Forest algorithm so that the best classification performance of the models can be found more correctly and one best classification algorithm can be selected. In this research, the purpose is to predict the reasons behind a suicide case.

Classifier	Accuracy	MAE	RMSE	RAE	RRSE
Naïve Bayes Classifier	78.7879 %	0.2213	0.4499	44.2988 %	90.027 %
SVM	81.0606 %	0.1894	0.4352	37.9131 %	87.0788 %
IBk (K-NN)	100 %	0.0075	0.0075	1.4939 %	1.4932 %
J48	96.2121 %	0.0688	0.1855	13.7774 %	37.1182 %
Random Forest	100 %	1112	0.138	22.2626 %	27.6082 %

Table 3: Comparison of five classifiers based on error measurement.

Both IBk (K-NN) and Random Forest have the same position in accuracy so these error values determine which classifier is best. From Figure 18, 19 and Table 3, it can be observed that the IBk (K-NN) classification model has the lowest errors measurement with the minimum errors in terms of Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and Root Mean Square Error (RMSE). Naïve Bayes Classifier has the highest errors measurement and others classifiers are at the middle based on error measurement.

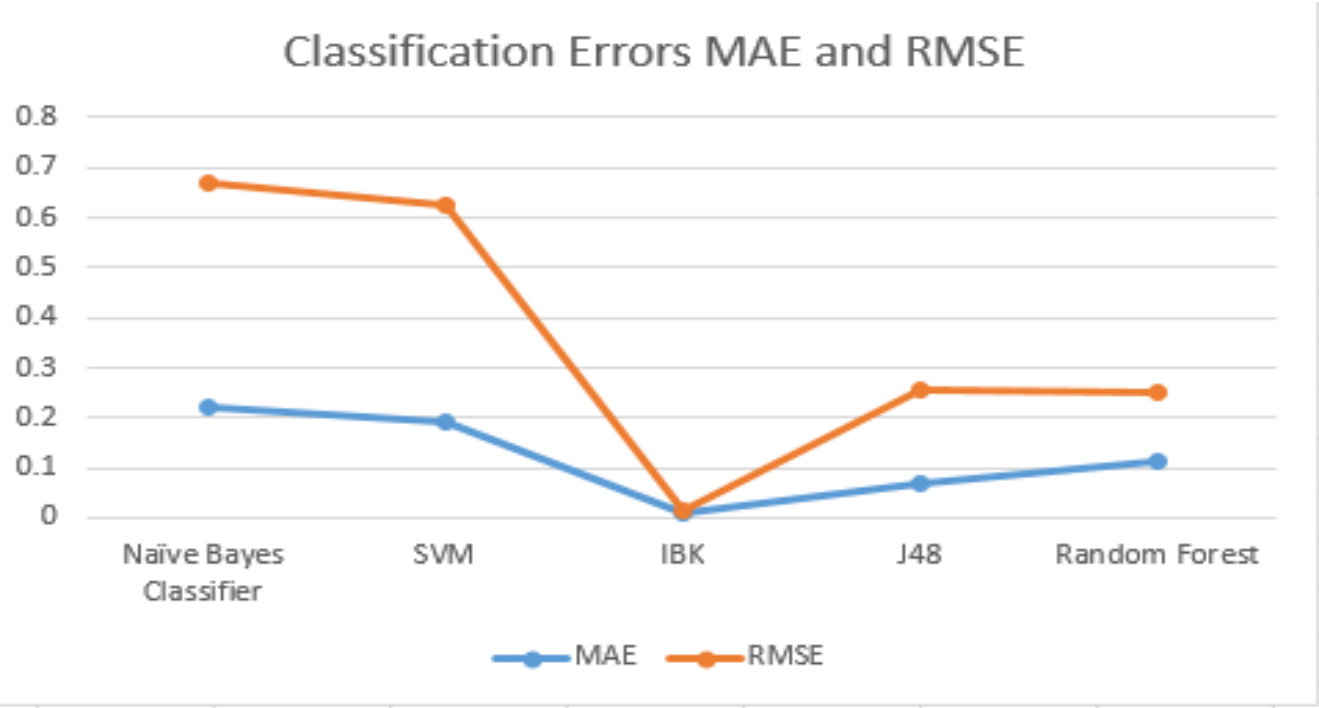


Figure 18: MAE and RMSE Metrics.

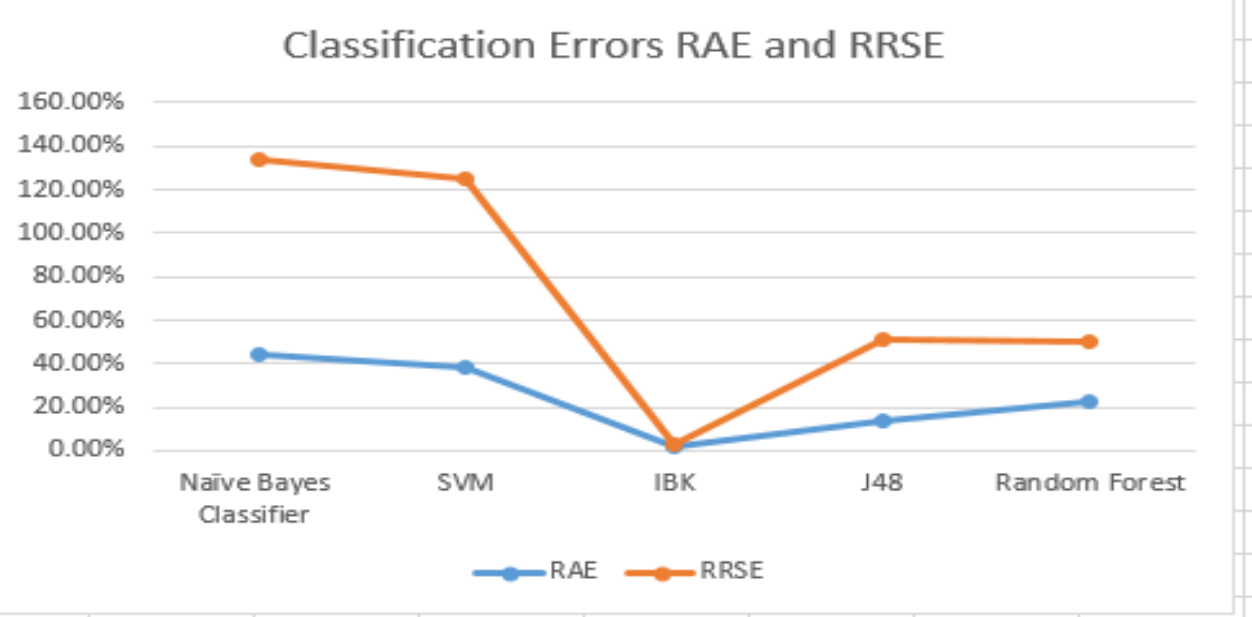


Figure 19: RAE and RRSE Metrics.

In IBk classifier and Random Forest classifier 1 is the value of The Kappa statistic. Support Vector Machine has the value of 0.621 in Kappa Statistic and 0.9241 is the value of Kappa statistic for J48. It means Most of the case models are statically significant.

Among the five classifiers based on the value of accuracy and error measurement model of IBk (K-NN) classification. Using IBk method for checking test data-set value whether it has predicted correctly or not in terms of the training dataset value. So, in WEKA 3.8 the Test dataset has uploaded and the option of test dataset has selected. Then in this case the IBk classifier has chosen and it showed the accuracy and the measurement of the error correctly (Figure 20).

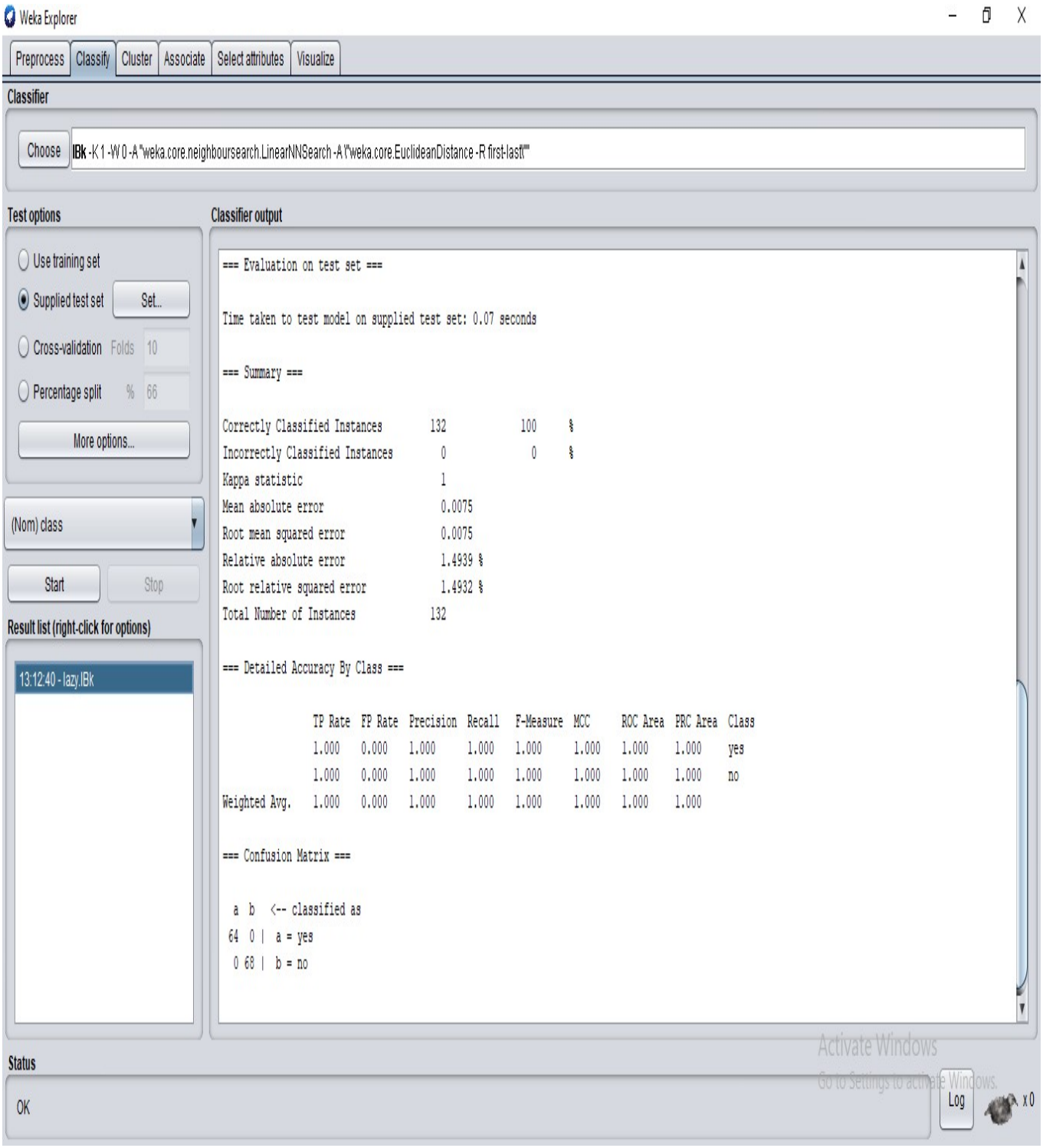


Figure 20: Value of the Test Set with Actual and Predicted Class.

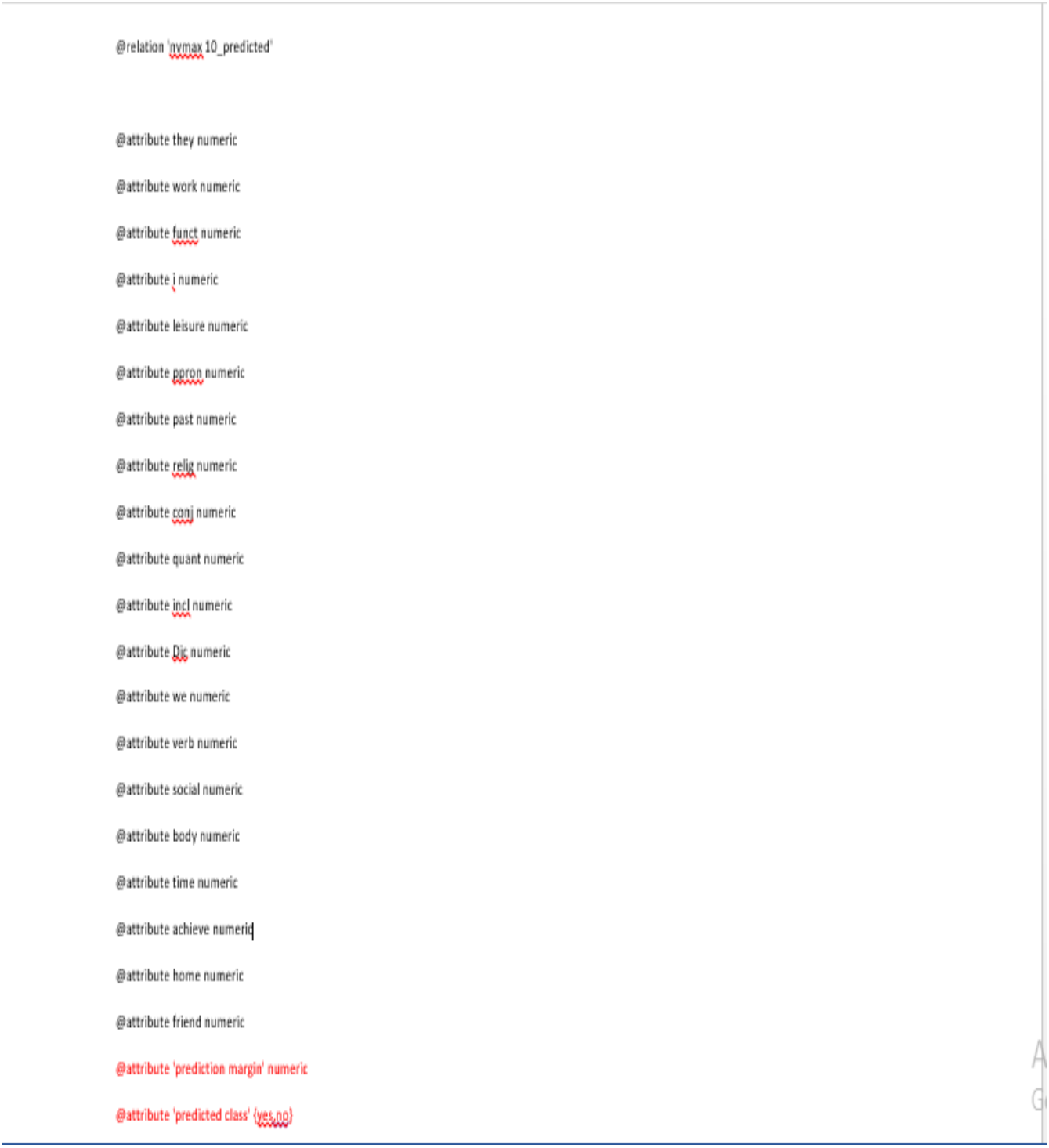


Figure: 21(a) - After Evaluating the Training Dataset the Result has shown

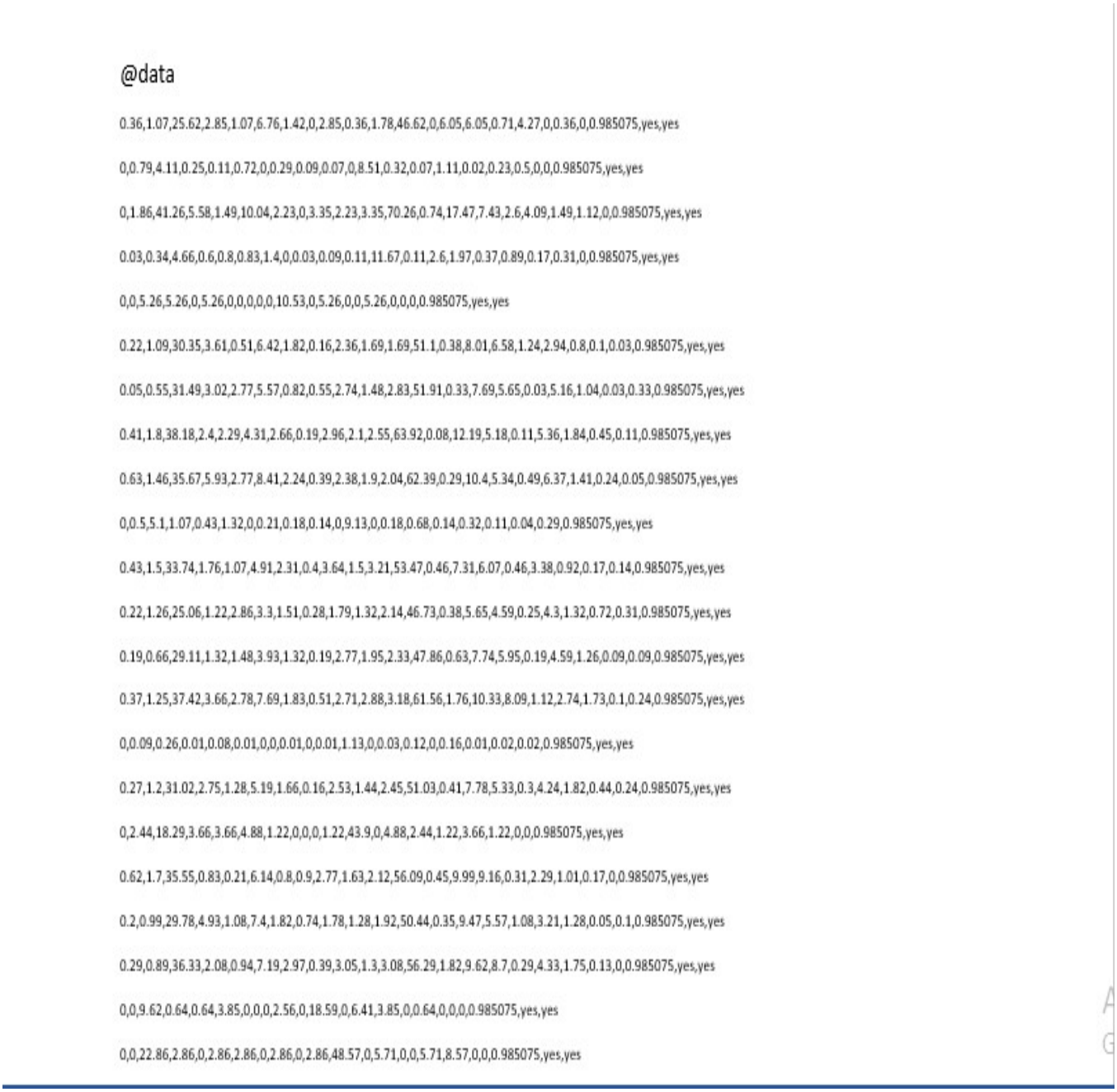


Figure: 21(b) - After Evaluating the Training Dataset the Result has shown

For observing and examining The ARFF file, it has checked its accuracy and error value (Figure 20). In this observation, most of the cases for predicting the suicide reasons are the same as variables of nvmax=10 except two cases (Figure: 21(a)). So, IBk has predicted the instances with the highest accuracy and lowest errors hence it can further be used to predict the reasons behind suicide.

4.2. Descriptive Analysis:

In this research paper, the variables have been categorized into two features these are yes and no that already mentioned before, to predict the reasons behind the suicidal cases.

After doing analysis and observations with the dataset it can be said that there are mostly affected 20 reasons behind the suicidal cases. In this research paper, there are 132 instances with 71 variables. Then after the analysis with nvmax=10 in R language finally this research has 20 variables those are they, work, funct, I, leisure, ppron, past, relig, conj, quant, incl, Dic, we, verb, social, body, time, achieve, home, friend. These are the major reasons behind most of the cases of suicide. Each of the words has some deep meaning and it mentioned earlier (Table: 1). This paper have used five classifiers and among those IBk (K-NN) is the best classifier model for this paper with its highest classifiers and lowest error management. So, for the suicidal cases, IBk (K-NN) is more preferable classifier it is proved.

In the end, it is clarified that if anyone wants to predict the suicidal reason, then these 20 reasons has the highest priority. There is a primary suggestion to everyone if someone has any symptoms among these 20 reasons then they should be given importance and care.

Chapter 5

5. Future Work and Conclusion

5.1. Future Work:

There is a considerable commitment to improving suicide prevention through the integration of AI into real life. Normal peoples and patient data can be collected from a multitude of sources including devices connected to the IoT, other mobile and smart technologies, social media, audio recordings, personal and written documents by clinicians, research databases, biological data, and EMRs. These data can be used to develop ML models in which suicidal patterns and suicidal behaviors, including risk and protective factors, can be identified and used to guide clinical management strategies and predictive analytics at the individual level.

For this research, the data was collected from twitter. We took about 132 peoples live data from their account using several tools and summarized the data to create a model. This model will give a solution and help to predict someone's suicidal activity more effectively and efficiently. The classification algorithm has used to predict a random people's suicidal activity based on some attributes value here. In future more work can be done to predict someone's suicidal activity using large number of data. On application can be developed with the ability to predict suicidal activity in terms of some academic and personal information. People can insert necessary values in that application and see the approximate outcome. Public will be able to know about their behavior and mental condition based on the prediction. Research can be done about the issues of the suicidal prediction system.

In addition, people's data can be used to support AI prediction efforts, which may be specifically targeted at individual users and entire populations. At the individual level, suicide prediction can help identify individuals in a crisis situation in order to offer emotional support, crisis and psychoeducational resources and rescue warning. At the population level, although real-time intervention is not feasible, algorithms can identify vulnerable groups or suicide hotspots that help inform about resource mobilization, policy reform, and advocacy efforts. Predicting AI-mediated risks is relatively early in its adoption and is likely to continue to expand its capacity.

5.2. Conclusion:

The chief aim of this study is to improve the accuracy of suicide rate predictions with advanced multiple variables prediction models. Several systematic reviews for suicide prevention strategies have focused principally on gatekeeper training, screening programs, public education, media education, and restricting access to lethal means [30]. If suicide high-risk days or periods could be predicted, various suicide prevention strategies could be optimized to reduce suicide number effectively after such vulnerable periods were declared for the population. Social media has revolutionized the way we communicate with the world and allowed us all to stay connected and self-expressed. Mixed anxiety depression and social media seem to exist in a vicious cycle, with one problem often stimulating another. Sometimes the periodicity of user publication was also considered to create a model for effective prediction of anxiety depression among users. The launch of social networks on the Internet has caused a sharp interaction between people. Although it positively affects society, it can also spread suicidal ideas, leading to infection effects. We used live Twitter data for our database as well as pre-tagged Twitter data and draw conclusions about the data to gain insight into the possible behavior of users experiencing suicidal thoughts. In this paper, a generalized model is designed that undoubtedly help the people to predict someone's suicidal activity in advance based on some attributes value and according to that prediction the people can take necessary steps. Subordinate classification of learning has limitations and cannot provide accuracy to the human level in predicting suicidal activity using textual data. Moreover, in Tweets collected prior to pre-processing, there is significant noise, which excludes approximately one-third of third-party data and news links. Later, a layer of expert suggestions can be added to the model to reduce the number of false positives. This will increase the accuracy of mood analysis to detect suicidal activity.

References:

[1] Cheng, Q., Li, T. M., Kwok, C. L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. *Journal of medical internet research*, *19*(7), e243.

[2] Kumar, A., Sharma, A., & Arora, A. (2019, March). Anxious Depression Prediction in Real-time Social Data. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*.

[3] Zhang, J., Jia, S., Wieczorek, W. F., & Jiang, C. (2002). An overview of suicide research in China. *Archives of Suicide Research*, *6*(2), 167-184.

[4] Lee, K. S., Lee, H., Myung, W., Song, G. Y., Lee, K., Kim, H., ... & Kim, D. K. (2018). Advanced daily prediction model for national suicide numbers with social media data. *Psychiatry investigation*, *15*(4), 344.

[5] Deshpande, M., & Rao, V. (2017, December). Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 858-862). IEEE.

[6] Chaudhary, L., Nair, V. V., & Prasad, I. (2019). Descriptive Analysis of Suicide Ideation on Twitter.

[7] Vioules, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, *62*(1), 7-1.

[8] de Andrade, N. N. G., Pawson, D., Muriello, D., Donahue, L., & Guadagno, J. (2018). Ethics and artificial intelligence: suicide prevention on Facebook. *Philosophy & Technology*, *31*(4), 669-684.

[9] Bae, S. M., Lee, S. A., & Lee, S. H. (2015). Prediction by data mining, of suicide attempts in Korean adolescents: a national study. *Neuropsychiatric disease and treatment*, *11*, 2367.

[10] Burnap, P., Colombo, W., & Scourfield, J. (2015, August). Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media* (pp. 75-84).

[11] Kumar, E. R., & Rao, A. K. R. (2019, February). Suicide Prediction in Twitter Data using Mining Techniques: A Survey. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 122-131). IEEE.

[12] Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, *113*, 65-72.

[13] Burnap, P., Colombo, G., Amery, R., Hodorog, A., & Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on Twitter. *Online social networks and media*, *2*, 32-44.

[14] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015, April). Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3187-3196).

[15] Mbarek, A., Jamoussi, S., Charfi, A., & Hamadou, A. B. (2019). Suicidal Profiles Detection in Twitter. In *WEBIST* (pp. 289-296).

[16] *Davood Astaraky,* "Linear Model Selection and Regularization - Subset Selection Methods."[online]
Available at:https://rpubs.com/davoodastaraky/subset#:~:text=The%20regsubsets()%20function%20(part,variables%20for%20each%20model%20size.

[Accessed july 8, 2020].

[17] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

[18] Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. *IJACSA) International Journal of Advanced Computer Science and Applications*, *8*(6), 19-25.

[19] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

[20] Fast, E., Chen, B., & Bernstein, M. S. (2016, May). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4647-4657).

[21] Sasso, M. P., Giovanetti, A. K., Schied, A. L., Burke, H. H., & Haeffel, G. J. (2019). # Sad: Twitter Content Predicts Changes in Cognitive Vulnerability and Depressive Symptoms. *Cognitive Therapy and Research*, *43*(4), 657-665.

[22] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

[23] These 5 social media habits are linked with depression | live science. [online]

Available at: https://www.livescience.com/62718social-media-habits-depression.html

[Accessed july 9,2020]

[24] Does social media cause depression? | child mind institute. [online]

Available at: https://childmind.org/article/issocial-media-use-causing-depression

[Accessed july 9,2020]

[25] Miller, D. D., & Brown, E. W. (2018). Artificial intelligence in medical practice: the question to the answer?. The American journal of medicine, 131(2), 129-133.

[26] Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence, 1, 316-334.

[27] Statista—The statistics portal for market data, market research and market studies. [online]

Available at: https://www.statista.com/

[Accessed july 10,2020]

[28] Internet Live Statistics (2018). [online]

Available at: https://www.internetlivestats.com/

[Accessed july 10,2020]

[29] Harvard, I. O. P. (2018). [online]

Available at: https://iop.harvard.edu/use-socia%20l-networking%20-technology

[Accessed july 10,2020]

[30] Clifford, A. C., Doran, C. M., & Tsey, K. (2013). A systematic review of suicide prevention interventions targeting indigenous peoples in Australia, United States, Canada and New Zealand. *BMC public health*, *13*(1), 463.