SYNOPSIS REPORT

On

SOCIAL MEDIA DATA CLASSIFICATION

Submitted by

PRATYUSH SHARMA (Enroll No. R103216072)

SHUBHAM KUMAR (Enroll. No. R103216094)

AYUSH CHATURVEDI (Enroll. No. R103216124)

Under the guidance of

Dr. T.P. Singh
Associate Professor

SCHOOL OF COMPUTER SCIENCE
UNIVERSITY OF PETROLEUM & ENERGY STUDIES
Bidholi Campus, Energy Acres, Dehradun – 248007.

May – 2019

CANDIDATES DECLARATION

We hereby certify that the project work entitled "SOCIAL MEDIA DATA CLASSIFICATION" in partial fulfilment of the requirements for the award of the Degree of "Bachelor of Technology in Computer Science and Engineering with Specialization in Business Analytics and Optimization" and submitted to the "Department of Informatics" at School of Computer Science, University of Petroleum and Energy Studies, Dehradun, is an authentic record of our work carried out during a period from January, 2019 to May, 2019 under the supervision of Dr. T.P Singh.

The matter presented in this project has not been submitted by us for the award of any other degree of this or any other University.

PRATYUSH SHARMA (Enroll No. R103216072)
SHUBHAM KUMAR (Enroll. No. R103216094)
AYUSH CHATURVEDI (Enroll. No. R103216124)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Date: May 20, 2019)

Dr. T.P Singh
Project Guide
Department of Informatics
School of Computer Science
University of Petroleum and Energy Studies
Dehradun - 248001

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide Dr. T.P Singh, for all advice, encouragement and constant support he has given us throughout this project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank our Head of the Department, Dr. T.P Singh, for his great support in doing this project Social media data classification.

We are also grateful to Dr. Manish Prateek, Director SCS, UPES for giving us the necessary facilities to carry out this project work successfully.

We would like to thank all our friends for their help and constructive criticism during this project work. Finally, we have no words to express our sincere gratitude to our parents who have shown us this world and for every support they have given us.

School of Computer Science

University of Petroleum & Energy Studies, Dehradun

Project Proposal Approval Form (2018-19)

Minor II

PROJECT TITLE: SOCIAL MEDIA DATA CLASSIFICATION

ABSTRACT

An uncommon growth in user-generated data, in particular social media and micro-blogging websites such as Twitter over the past decades. These resources offer a rich mine of marketing knowledge to organisations. The primary aim is to provide a method for analyzing opinion score in noisy twitter streams. Sentiment analysis of the tweets determine the different views of different people and their area of interest of vast population towards specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. Various machine learning methods for pre-processing Social media data and find those ones which impact on the building precise classifiers.

INTRODUCTION

In our daily life activities, we may require to make decision for which we seek out people advice or opinion. This decision can be anything from investing in a particular field to voting a country future. With the explosive growth of social media in the last few years, the web has drastically changed to the extent that nowadays billions of people all around the globe, which are freely allowed to conduct many

activities such as interacting, sharing, posting and manipulating contents, i.e. Use of the micro blogging platform like twitter. Sentiment analysis (aka opinion mining) is one class of computational techniques which automatically extracts and summarizes the opinions of such immense volume of data, which the average human reader is unable to process as shown in fig 1. Before the automation text process through NLP the analysis was done manually.

Fig1 Shown some Sentiment words

Sentiment analysis aims to identify and extract opinions and attitudes from a given piece of text towards a specific subject , Through the platforms like Face book, Yammer and Twitter, millions of status updates, posts and Tweet messages, which reflects people's present opinion and attitude towards particular agenda, are created and posted every day. However, Sentiment analysis of micro blogs like twitter is considered as an a lot more difficult issue because of the one of a kind qualities controlled by

micro blogs (for example short length of status updates and language varieties with provincial language impact).

PROBLEM STATEMENT

- The negative tweets which have the dark motives to other.
- A message where hate speech, trolling and social media bullying used, which become serious issues in these days.
- The argumentation which aims to identifying the reasons of such opinions and the overall reasoning path in general.

LITERATURE REVIEW

In this project we are trying to classify social media posts to make the experience of the users better by identifying toxic posts and comments. Here are the conclusions of some of the reference papers and publications that we review to make this project better and to know technologies we use in our system.
In the paper [citation2] by Yang and Pederson, they did comparative study of feature selection methods in statistical learning of text categorization. The focus was on aggressive reduction of dimensionality of data. Primarily five methods were evaluated. Document Frequency (DF), Information Gain (IG), Mutual Information (MI), CHI test, and Term Strength (TS). The conclusion of the paper found that IG and CHI were most effective, using IG with KNN algorithm on Reuters corpus yielded 98% accuracy, improvement over the basic algorithm implementation. Strong correlation among DF, IG and CHI were found. MI were least effective because towards bias for rare words. TS reduced 50% vocabulary but was not effective with large data. Lastly IG and CHI were most effective of the five but both had higher computational cost than DF, which was a little bit less effective but had huge advantage in computational performance.

In the paper [citation3] by Cohen and Singer, they did a context based text categorization study using two models. RIPPERS and sleeping-experts for phrases are evaluated on large text categorization problem. Both algorithms take into account context of a word ‘w’, how its presence or absence in a document affects its classification. However, both radically differ into forming the rules for context. Both algorithms are efficient on large, noisy corpora, running in linear or almost linear time. Both represent data directly, i.e. in the form of ordered list of tokens. Most common classifiers like Naïve Bayes only takes a word or word stem into account of context which is unrealistic. Combination of

words in actual world forms a context. Hence, we need to relax the rule of word contexts. Both algorithm work on that in a different way. Sleeping-experts take combination of words into account. Challenge here is to found useful complex features, given the enormous space of potential features. Second way to approach this is of RIPPER algorithm. RIPPER can be thought of as learning a disjunction of “contexts”, each context defined by a conjunction of simple terms. The principle technical challenge when this approach is followed is to learn these nonlinear classifiers efficiently. Final commonality between the two algorithms is that they form rules by themselves without any external input, learning from train data itself. By implementing the algorithms on Reuters data the performance of both algorithms was compared with Rocchio linear algorithm. It was found that Sleeping-experts was most accurate of the three with least error rates. RIPPER algorithm was also more effective than Rocchio but not by much margin. Precision of sleeping experts was 80, RIPPER-78 and Rocchio-72.

In paper [citation4] by El-Din, he did sentiment analysis on online research papers. The aim of the paper was to get sentiments of comments on those papers and hence find out papers that have positive feedbacks. ML system investigated the classification accuracy of Naïve Bayes algorithm. In addition, the research made a judgment of feature selection techniques like Bag of Words and TF-IDF for web scrapping. This research used three machine learning algorithms (Naïve Base Classifier, K-nearest neighbor, and random forest) to calculate the sentiments accuracy. In conclusion it was found that the random forest improves the performance of the classifier.

OBJECTIVES

In this project the following objectives has been tried to achieved: -

1. Preprocessing and exploration of data.
2. Extract features from the cleaned text using feature TF-IDF.
3. We use the predefine model using TF-IDF feature and test the model and sets to classify the unlabeled tweets.

METHODOLOGY

NLP is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language. Cleaning up the text data is necessary to highlight attributes that we're going to do in our machine learning system to pick up on[1] as shown in fig 2. The preprocessing of the text data is an essential step as it makes the raw text ready for mining Here in this project we are using the data from the twitter to analysis because now a day's twitter has over millions of user and the tweets from twitter really affect the users that are why we have the sentiments tweets from the twitter, and also show the visualization on the sentiments words.

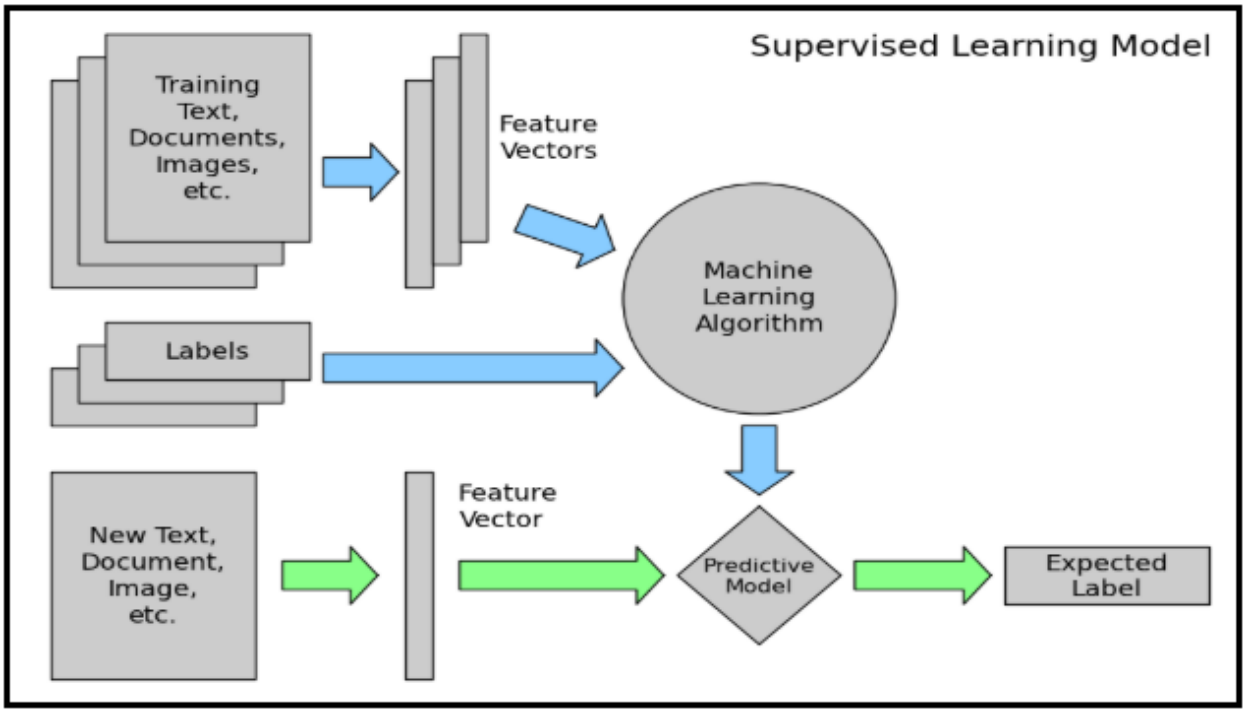


Fig 2 Getting Expected label with the help of ML Algorithms

To analyze a preprocessed data, it needs to be converted into features. Depending upon the usage, text features can be constructed using assorted techniques like in this we are using TF-IDF , we are using the predefine model to train, because the process itself is intuitive and its vectors are interpretable and the feature selection gives more accuracy and effectiveness.

The data cleaning include the 5 steps

1. Removing Twitter Handles (@user)
2. Removing Punctuations, Numbers, and Special Characters
3. Removing Short Words
4. Tokenization
5. Stemming

Fig 3. Data Cleaning

- **Data Preprocessing Algorithm :**
  1. Remove the @user pattern from the combined data set with the help of the vectorize function
  2. Getting rid of the punctuations, numbers and even special characters, ie a-z, A-Z and #
  3. Remove all the words having length 3 or less.
  4. Splitting a string of text into tokens i.e tokenization.
  5. Stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

- **Feature Extraction Algorithm**

  1. From the sklearn.feature_extraction.text we will import IT-IDF vectorizer library.
  2. After the vectorization of TF- IDF feature we will take clean preprocessed as an input.
  3. Then Analyze it.

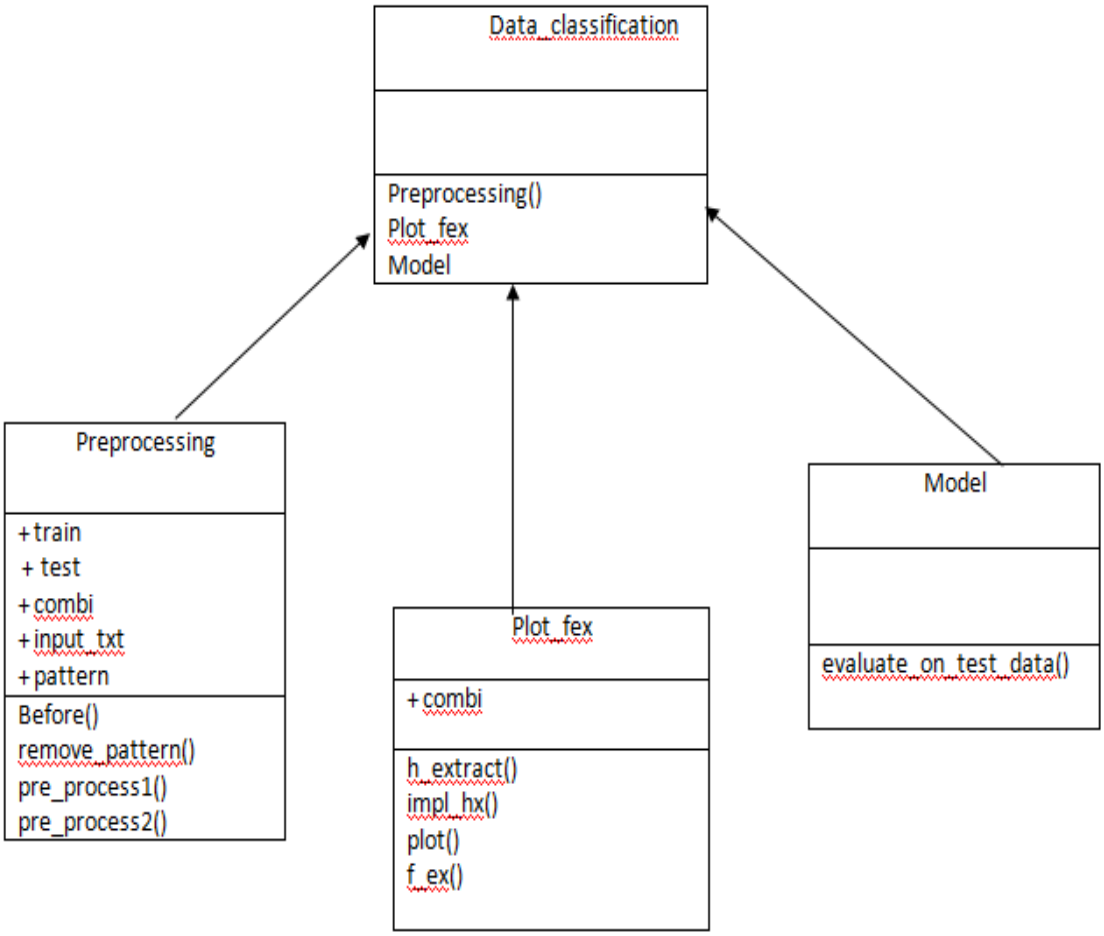
- **SVM Model Algorithm**

1. Read the clean csv file of tweets
2. Read the train csv file
3. Create object of plot fex.
4. From tfidf function inside plot_fex() call it by using object method and Save it in variable
5. Split the data as
    a. X_train = train_data_tfidf
    b. X_test = test_data_tfidf
    c. y_train = df_train["label"].values[:31962]
    d. y_test = df["label"].values[31962:]

For tfidf.

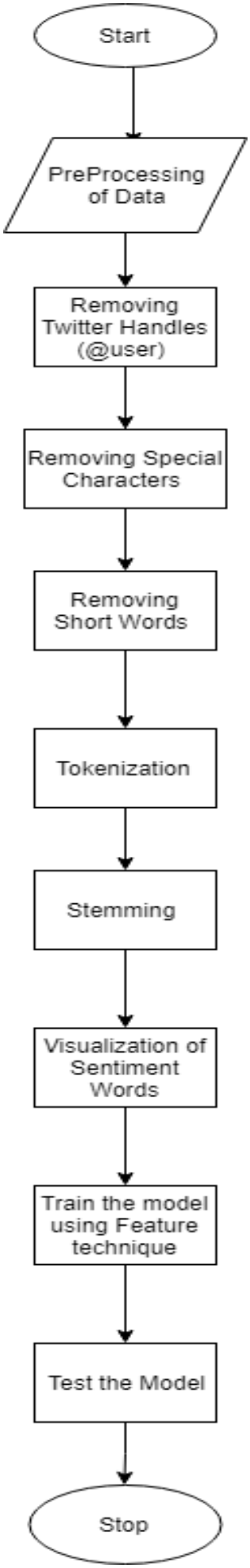
6. Import svm from sklearn
7 . Define a function to calculate accuracy
8. Tune the svm model
9 .Use the X_train,y_train to fit the model.
10. Evaluate accuracy

Class Diagram



I

Flow Chart


Start
PreProcessing of Data
Removing Twitter Handles (@user)
Removing Special Characters
Removing Short Words
Tokenization
Stemming
Visualization of Sentiment Words
Train the model using Feature technique
Test the Model
Stop

In this Project we will use Waterfall model, In this model, the software development activity is divided into different phases and each phase consists of series of tasks and has different objectives. This model widely used in the software industry. And easy to use In this development of one phase starts only when the previous phase is complete. Because of this nature, each phase of waterfall model is quite precise well defined. This model is simple and easy to understand and use and also easy to manage due to the rigidity of the model each phase has specific deliverables and a review process as shown in fig 6

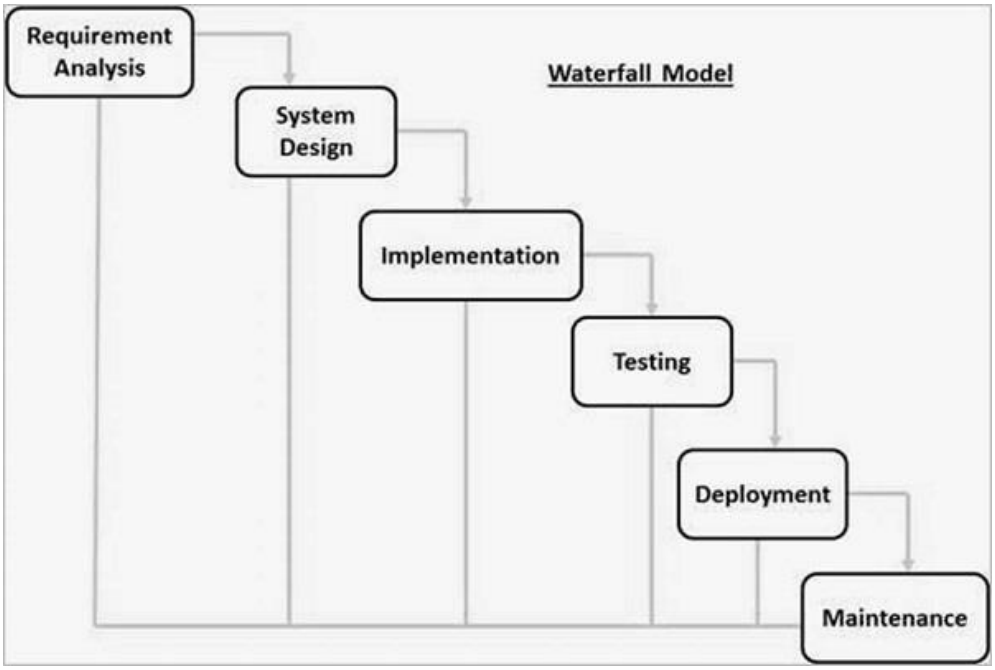


Fig 6. Process of Water fall model

Output

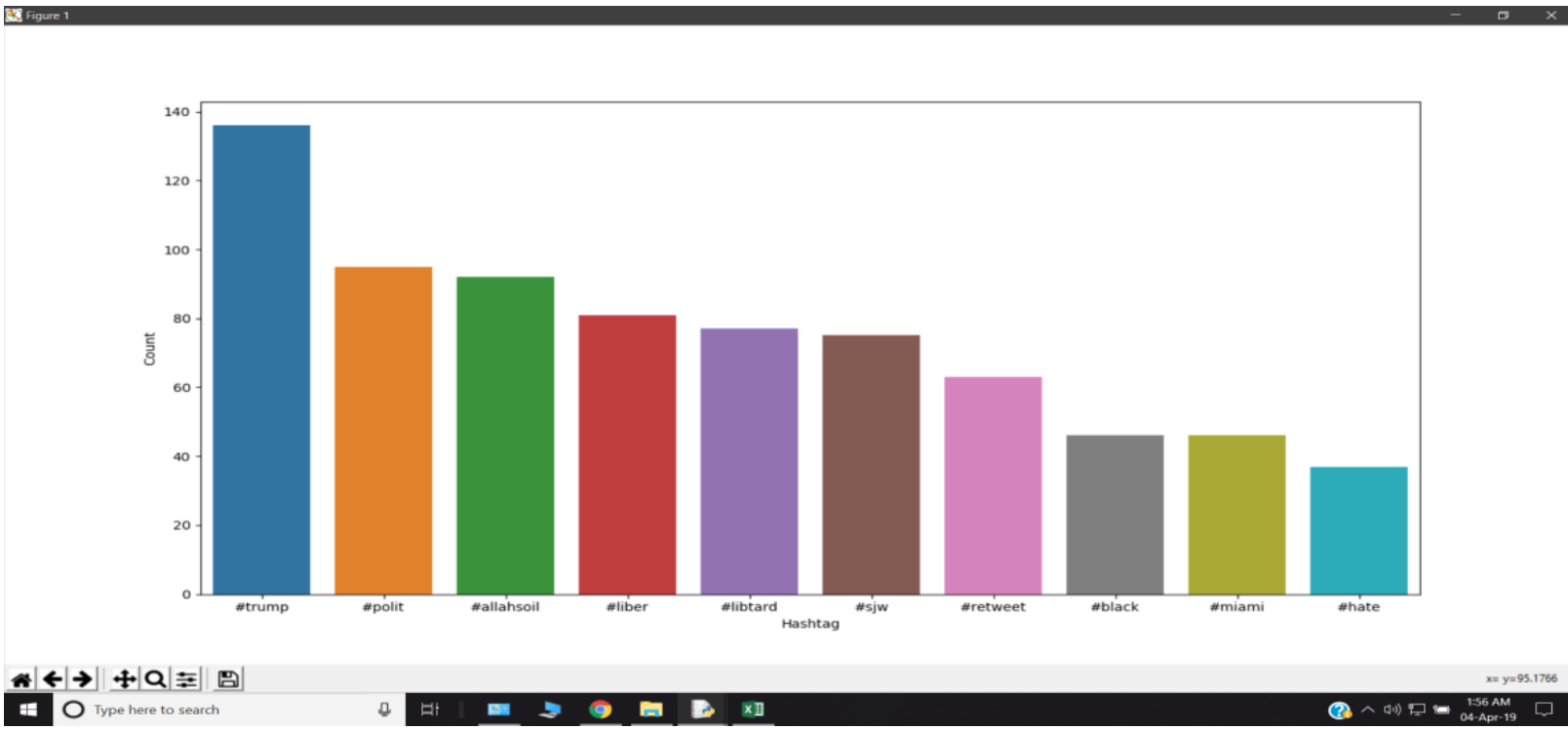


Figure 1
Count
Hashtag
#trump
#polit
#allahsoil
#liber
#libtard
#sjw
#retweet
#black
#miami
#hate
Type here to search
1:56 AM
04-Apr-19

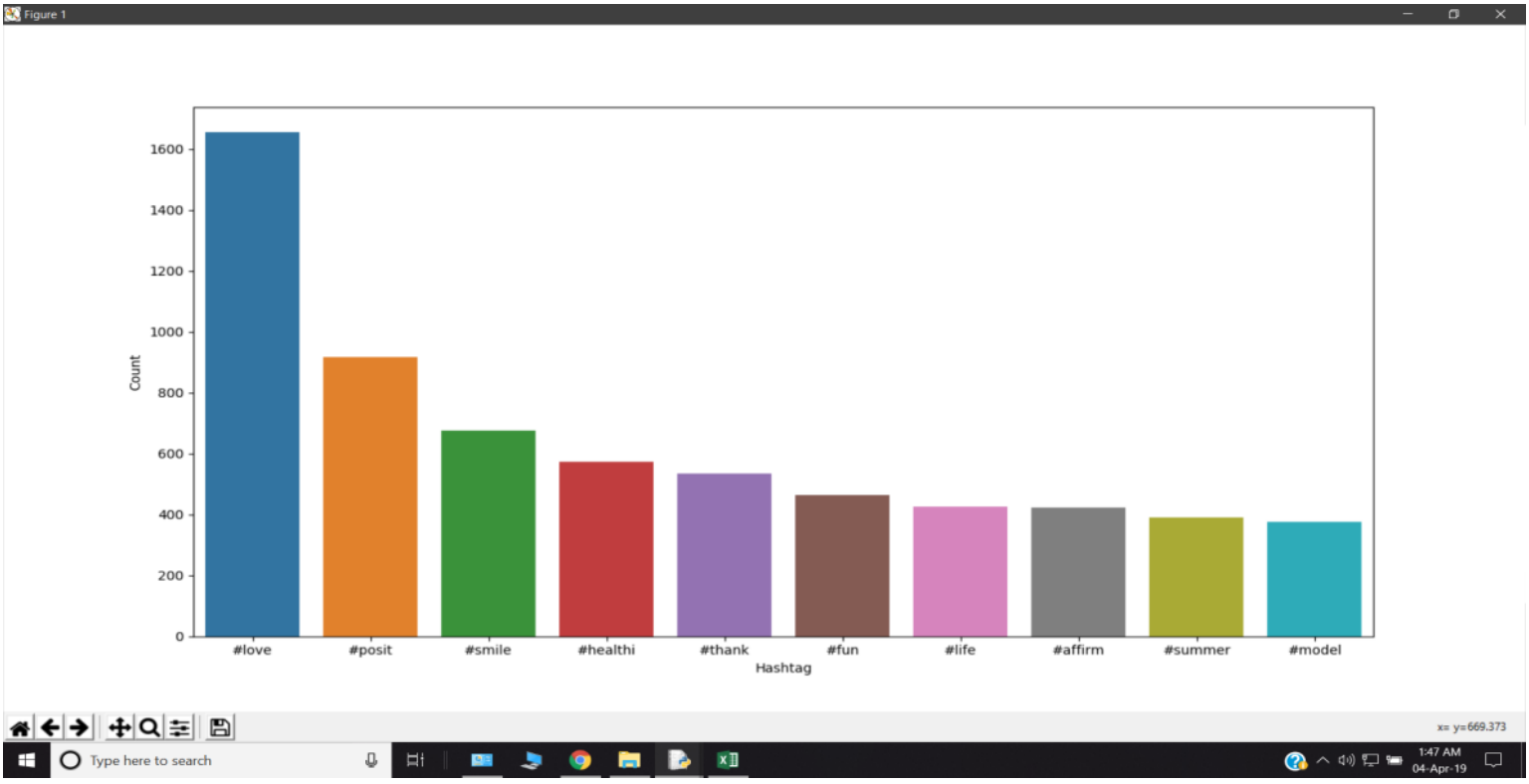
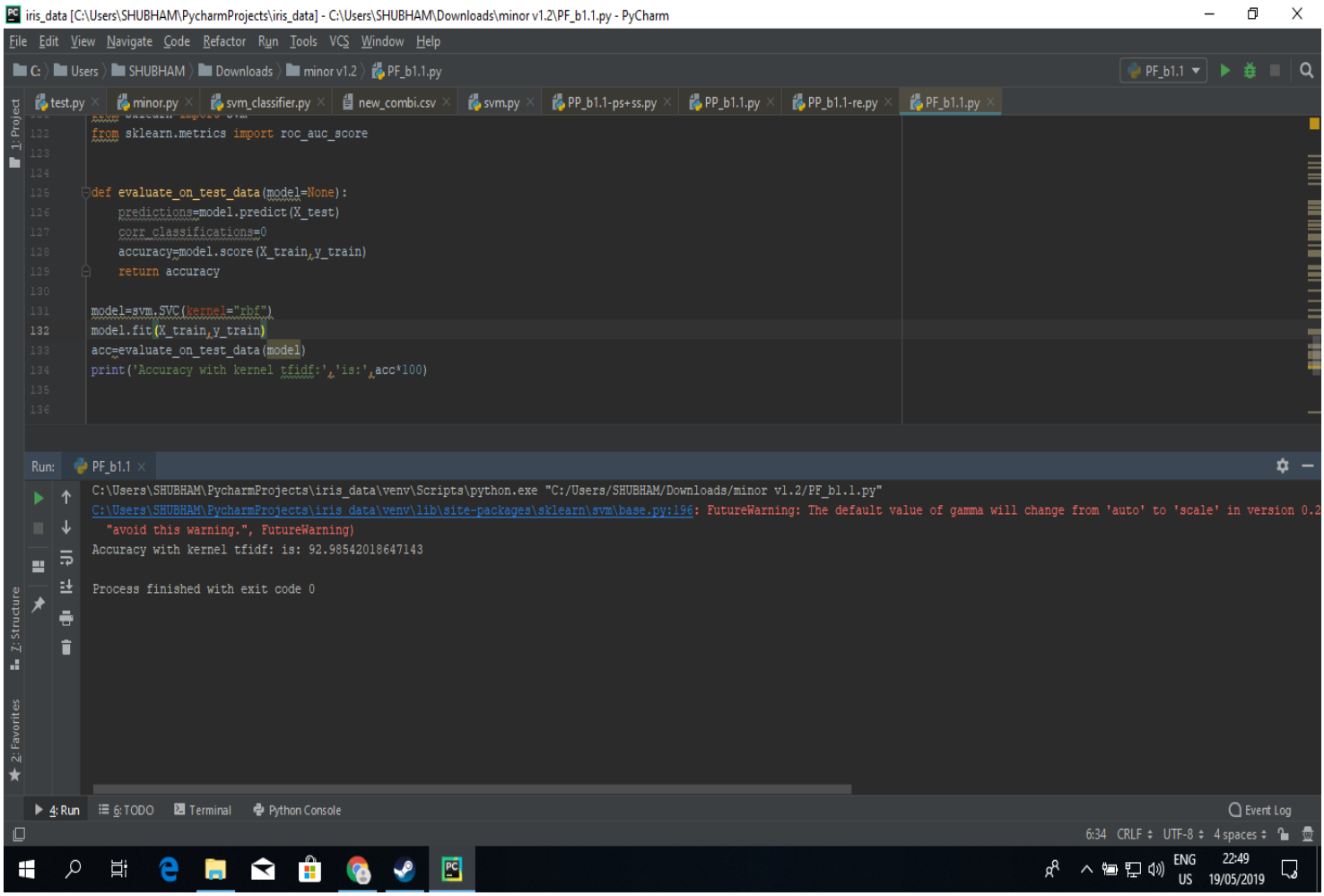


Figure 1
Count
Hashtag
#love
#posit
#smile
#healthi
#thank
#fun
#life
#affirm
#summer
#model
Type here to search
1:47 AM
04-Apr-19



Conclusion

1. Data was explored and pre-processed.
2. Features are extracted using TF-IDF method.
3. SVM model was successfully used for classification of data

SYSTEM REQUIREMENTS

Hardware Interface:

- 64-bit processor architecture supported by windows 10.
- Minimum RAM requirements for proper functioning of latest windows 10 is 2 GB.
- Required input as well as output devices.

Software Interface:

- The system is developed in ‘python 3.0+’ programming language.
- Windows 10

REFERENCES

[1] https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b

[2] Yiming Yang and Jan O. Pederson, 1997, ICML ’97 Proceedings of Fourteenth International Conference on Machine Learning Pages 412-420.

[3] William W. Cohen and Yoram Singer, April 1999, ACM Transactions on Information Systems, Vol. 17, No. 2.
[4] Doaa Mohey El-Din Mohamed Hussein, 2016, ‘Analyzing Scientific Papers Based on Sentiment Analysis’, Department of Computers and Information, Cario University.

Synopsis Draft verified by

Project Guide
(Name & Sign)

HOD
(Dept. of Systemics)