TEXT-TO-EMOTION PREDICTION USING SVM

ABSTRACT:

Text-to-emotion prediction is a necessary part of today's world as human are going online to a large extent and humans are prone to making mistakes also as with the large amount of data being generated chances of getting errors on the text leading for humans to interpret wrong emotions. This paper is based on Linear SVC() model to classify the text based on the 6 most prominent emotions i.e., anger, fear, joy, sadness, love, surprise . For this Paper Investigated many classification models to finally select this model .Used random forest , decision tree ,used MultinomialNB , GaussianNB, SVM with different kernels and along with hyperparameter tuning but LinearSVC yet giving the best results. Different data preprocessing techniques are applied for making the data noise free like stop word removal , digits removal ,punctuation marks , Vectorizations etc. finally the accuracy was taken into consideration for each model and F1 score and other parameters was considered to get to the final model to be deployed as a local host web server.

KEYWORD:

Emotions, Linear SVC, Text-to-emotion, MultinomialNB , GaussianNB, SVM.

INTRODUCTION:

We as a human can easily be able to interpret each other and convey our emotion to each other. There are two ways of doing so that is we can talk to each other and the other is we can write it down to do the same one more important thing is we can use facial expression also to get our emotions conveyed. When I talk about the emotions it opens a gateway to a large set of applications The first and foremost can be how we are interacting with the computers and how do a computer is able to understand our emotions. With the help of Text to emotion prediction we can make a computer understand us more accurately and precisely so as to interact to us in a betters way. We can use this technology to get the feedback on a large scale about certain set of people over a certain area just by monitoring their online comments and blogs and predict the happiness index and get to know more about the people in that area. We can also stop many unwanted things to happen like crimes just by monitoring the texts of some suspicious persons. We can make use of this technology to tackle the situations of suicidal thoughts in a person[1].

When we see today's world with this full of social media stuffs and there is a ton of data being generated on social media like (Facebook, Instagram, WhatsApp) when it comes to today's world Where tracking people is very difficult now a days as it is almost impossible manually. When we use this kind of technology we can make the work of company workers to get a smaller set of filtered data and help them get more chance to stop cyber-crimes as we can have a look at some of the malicious text and chats in which peoples are talking. We can also have one more Application as when a person is complaining about any sort of thing to the social media team the team members have to be very precise in understanding the context get back to the answer[1]. If any famous person gives some negative reviews about the company it just takes a day at max to get viral which can lead to a catastrophe for the company and its reputation.[1]
As of 2016, YouTube gives greater than 1 billion registered users. [4] we can see and use emotion analysis to get the reaction of different sets of people through this model.
The rest of the Paper is organized as follow :

1. Related Works
2. Proposed Model
3. Result Analysis
4. Conclusion
5. References and future scope

1. Related Works:

The works done in this field which I came across some of the papers which have done some textual analysis to predict one such paper worked on the (Ekman's model)[3].There are works on this field but model of them are discrete and they tend to have more specific goal for the paper but this paper mainly focuses on the best model suitable for this data set .Emotion detection of tweets[2], And from many of the previous research work fields I have seen that that they have used certain analysis on several models such as random forest , decision tree, SVM and they have attained the accuracy of less than 80%.Other than that there is none to my knowledge using machine learning models to gain more that this accuracy on this dataset.

Amira F. El Gohary et.a l [6] are detecting the emotions out of the Arabic children stories and predicting on these basic six emotions joy, fear, sadness, anger, disgust and surprise.

Their approach achieved 65 % accuracy for emotion detection.

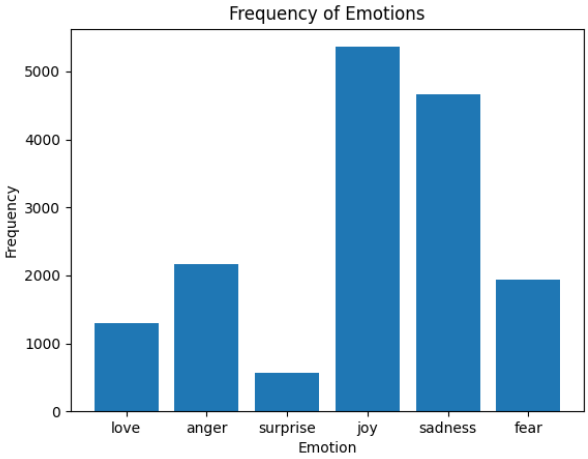


Fig .1

2. Classification algorithms:

(i) Multinomial Naïve Bayes :

This algorithm is used in many of the NLP applications naïve bayes uses the concept of bayes theorem and find the maximum likelihood of the features .It assumes that all the features are independent of each other which is very rare in todays world .

$$P(a|b) = \frac{p(c|a)p(a)}{P(b)}$$

Where :

a : class,

b : predictive document,

p(a) :probability for given class,

p(b) : probability for document ,

p(b|a): conditional probability when b is given,

p(a|b) : conditional probability with respect to document and class c.

(ii) LinearSVC :

Linear svc belongs to the algorithm set of support vector machines. It uses several mathematical formulas to make the best possible hyperplane to separate different classes mostly linearly separable.

Linear Equation :

w**T*x + b = u

w = Weight vector for the hyper plane

b = bias for the hyper plane

a) **Subject function** = y(i) * (w**T*x(i) + b) for all I

b) **The objective function :**

c) **Minimize function** = ½ * ||w||**2

Y(i) = labels for m^{th} data point

X(i) = feature vector

d) **Langragian Dual Form :**

L(w, b,@) = ½ * ||w||2 – sum(@(i) *(y(i)*(w**T*x(i) + b) – 1 __**

Here @ is a vector of Lagrange multipliers.

3. Proposed Model

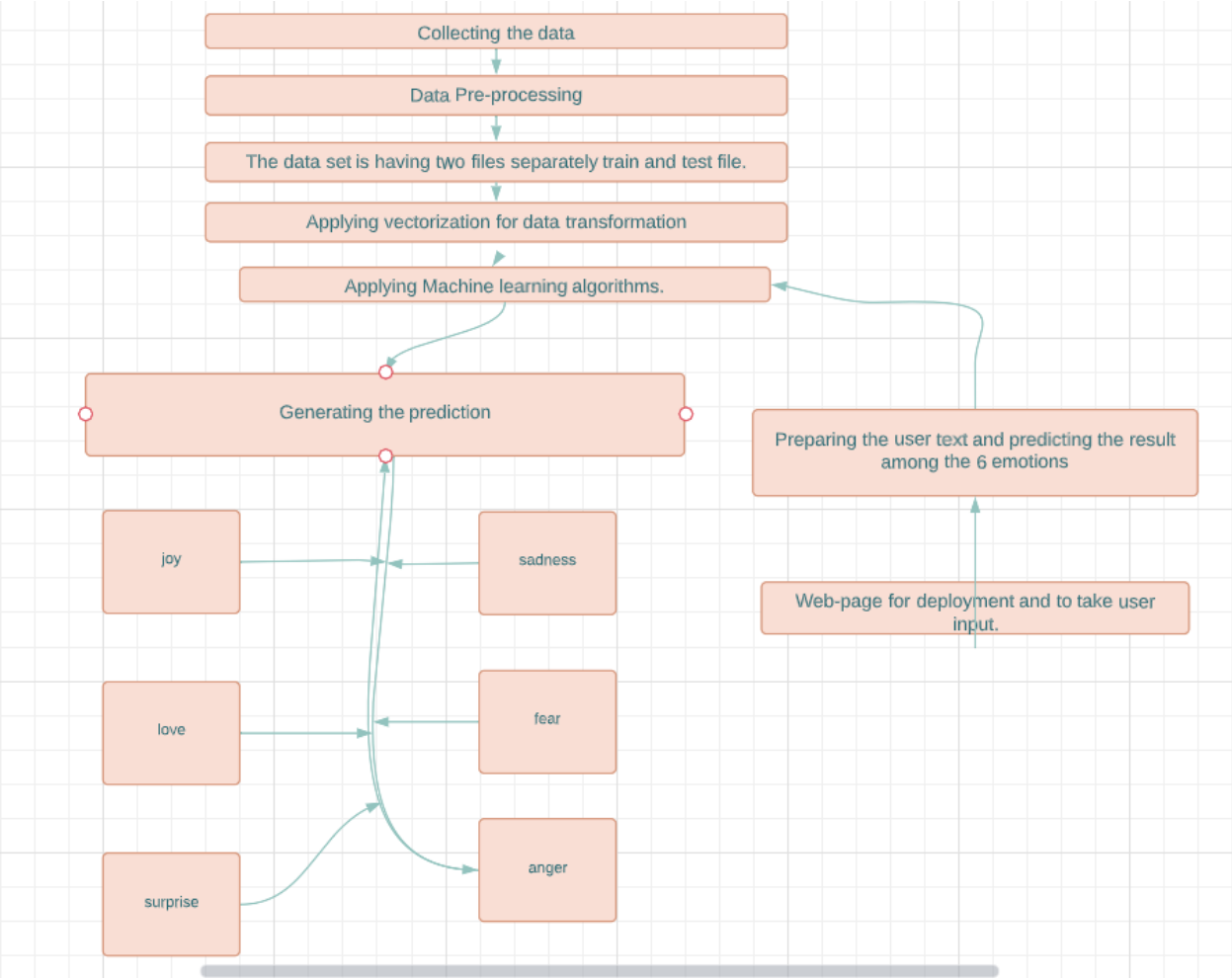


Fig : 2

3.1 Vectorizing :

th(d, t) = f(d, t) (1)

$$idf(t) = \{\log(N/df(t)\} \quad (2)$$

$$TFIDF = tf(d, t)idf(t) \quad (3)$$

(TFIDF) is like weights which can be used in text mining and taking information. [8]

F(d, t) It is the prevalence of the phase t in record d. [5]

This project is made using python language only and for the final representation and GUI is done using HTML and CSS to make a web page to deploy the project.

3.2 **Validating**

K means validation is a method to determine success of a system [9]

$$\text{Accuracy} = \frac{\text{sum of the right}}{\text{Sum of data}} \times 100\%$$

The model starts by importing the datasets from the csv file using pandas module and converting it into the Dataframe.

Then the data set goes through a series of cleaning and filtering to remove unwanted texts and characters such as special characters and numbers as they don't convey anything to the emotions . Removed and replaced all the short forms like(isn't) to "is not" . using different libraries such as (ntkl for Stop words).

Then transformed the texts to vectors for the further processing and to generate the features and classify the emotions as per the model which is used by me.

Then after the preprocessing loaded the **Linear SVC** from sklearn library and then

Started the training from the data set containing 16000 line of texts and its corresponding emotion.

After applying the training and model was getting the accuracy of : **89.14%.**

4. RESULT ANALYSIS

This paper discusses about several trained models and to choose the best model and performed several hyper parameters tuning to get the desired result

4.1 GaussianNB classifier:

After doing all the data cleaning and then applying the gaussian model and the accuracy, model is getting is : 34.8%

4.2 MultinomialNB classifier :

After doing all the data cleaning and then applying this model and the accuracy, model is getting is: 68.38%.

4.3 CategoricalNB classifier :

After doing all the data cleaning and then applying this model and the accuracy, model is getting is: 34.76%

4.4Random Forest classifier :

After doing all the data cleaning and then applying the gaussian model the accuracy, model is getting is: 89.14% But it was relatively very slow while training the model.

a. Using n_estimators = 400
   The accuracy decreases : 88.79%

4.5Using Ada boost boosting techniques:

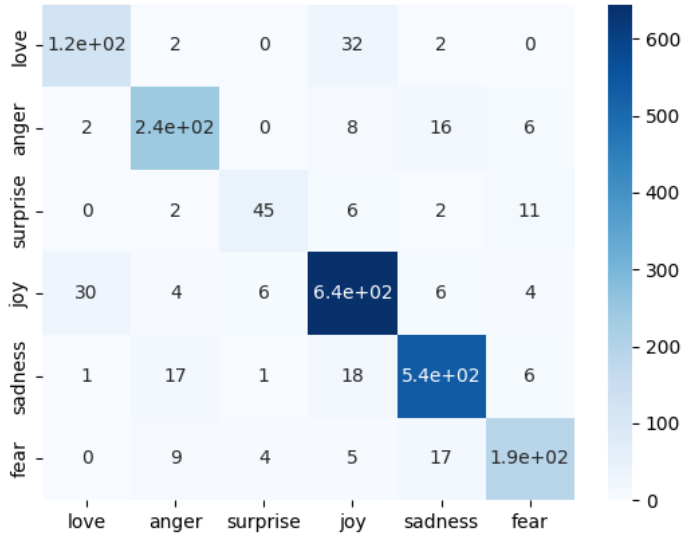
We have several boosting techniques like adaboost classifier ,attained the accuracy of : 77.79%

MODEL	Performance
GaussianNB classifier	34.8
MultinomialNB classifier	68.38%
CategoricalNB classifier	34.76%
Random Forest classifier	89.14%
Ada boosting	77.79%
Linear svc*	89.14%

Table :1

For Linear SVC()

F1_score =0 .8499



5. CONCLUSION AND FUTURE SCOPE

This research paper is dedicated to generate best possible accuracy for the given dataset and I have compared several classifications algorithms for the same to see which model is performing the best in its domain with all other conditions kept

So far by comparing all the algorithms Linear svc is the best for the same although random forest classifier gives the same accuracy but here also linear svc get on with the benefit as it takes relatively very less time to get trained So, final conclusion is that Linear svc performed best for this dataset. There is a very good and demanding future scope for this thing as day by day we are generating. It is very difficult for now to get the emotions from the several texts based data to make this work easier and to get the emotions of a large text that is being generated daily. We can use this technique to get the summarized analysis of several comments on the sites or channels such as YouTube channel or Instagram reels to predict the peoples response easily and effectively.

6. REFERENCES

[1] Text-based emotion prediction system using machine learning approach June 2020 IOP Conference Series Materials Science and Engineering 769(1):01202

[2]
https://www.researchgate.net/publication/327120380_Emotion_Detection_of_Tweets_using_Naive_Bayes_Classifier

[3] Agrawal, A., & An, A. (2012, 4-7 Dec. 2012). Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. Paper presented at the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

[4] Chen, Y.-L., Chang, C.-L., & Yeh, C.-S. (2017). Emotion classification of YouTube videos. Decision Support Systems, 101, 40-50. doi:10.1016/j.dss.2017.05.014

[5] Pal, K., & Patel, B. V. (2020, March). Data classi□cation with k-fold cross-validation and holdout accuracy estimation methods with 5 different machine learning techniques.

[6] Amira F. El Gohary, Torky I. Sultan, "A Computational Approach for Analyzing and Detecting Emotions in ArabicText", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, Issue 3,May-Jun 2013, pp.100-107

[7] Ma C J & Ding Z . S (2020,December).Improvement of k-nearest neighbor algorithm based on the double filtering .In 2020 %th International conference on Mechanical Contrlo and computer science Engineering (ICMCCE) .

[8] Yasmina, D., Hajar, M., & Hassan, A. M. (2016). Using YouTube Comments for Text-based Emotion Recognition. Procedia Computer Science, 83, 292-299. doi:10.1016/j.procs.2016.04.128

[9] Wahyono, I. D., Ashar, M., Fadlika, I., Asfani, K., & Saryono, D. (2019, October). A new computational intelligence for face emotional detection in ubiquitous. In 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)

TEAM MEMBERS:

1.DOKKU PAVAN-21BCE8451
2.GARIKIPATI VENKATA GUNA CHAITANYA-21BCE7607
3.GONDU SAI KIRAN-21BCE7067
4.TADALA ADITYA SRI VAMSI -21BCE7164