AIRLINE TWITTER SENTIMENT ANALYSIS

* US airline review on twitter in 2015

Akanksha Dewangan
Computer Science Department
Indraprastha Institute Of Information Technology, Delhi
Delhi, India
akanksha19049@iiitd.ac.in

***Abstract*—Aim of this project is to make models for 6 major U.S. airlines that have collectively customers review on which sentimental analysis has performed for concise feedback so that airlines can improve services of customer complains.There are 3 type of classifier performed in this project (1)K-Nearest Neighbors,(2)Multinomial Naive Bayes on Twitter data which is taken from Kaggle.Accuracy are good so that our model can use for prediction.The classifier trained using 80% training data and 20% test data.Accuracy of models are also compared in which Multinomial classifier is the best approach among all.**

***Index Terms*—1.CountVectorizer,2.TfidfVectorizer,3.K-Nearest Neighbors,4.Multinomial Naive Bayes.**

I. Introduction

Twitter is the best customer feedback and review platform hence its tweets on twitter is having a very high importance on the airlines feedback.It is assuming that travellers and straphangers are tweeted their experience on twitter websites.In this project there is use of experiences of flight share on twitter by the people's and give suggestions to airlines that how to improve the services on the basis of negative feedback. In the data set there are about 15000 tweets collected on various airlines reviews from 2015 and this tweets are available in the kaggle.Reviews are labeled as positive, negative or neutral according to polarity of the sentence.Percentage of negative reasons are very high in the dataset.By knowing the reviews we can suggest the airlines to improve the services on the basis of negative reasons.Main goal of this project is to build the model which predict whether a tweet is positive , negative or neutral and this is sentiment analysis on Airlines tweet data.

II. Methedology

Testing different models for natural language processing is an interesting thing to look for.By using the test data of preproccessed tweets we check and compare the result with already labeled sentiments and obtain model accuracy.Also we compare the accuracy of different models applied on the data.

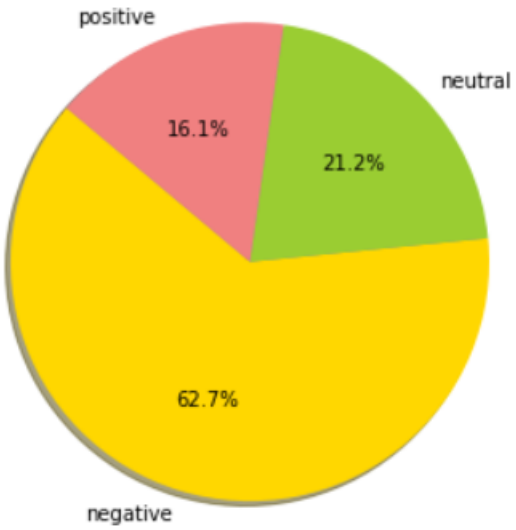
Identify applicable funding agency here. If none, delete this.

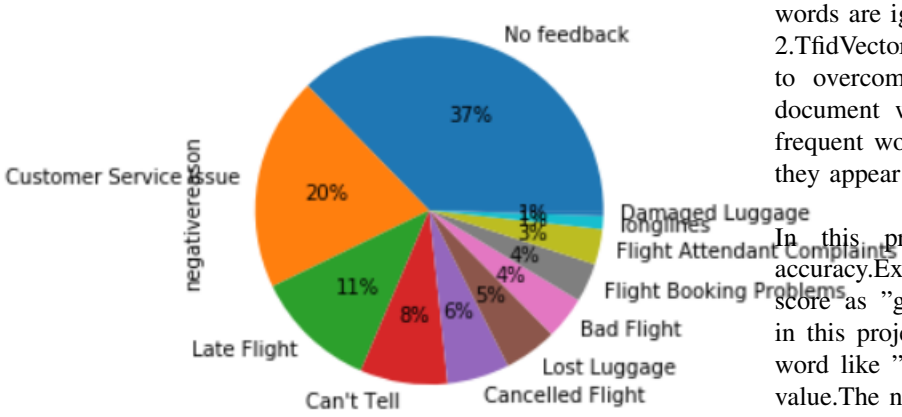
A. Data Collection

The data set used is taken from kaggle which is a standard dataset of "US airlines review from twitter of 2015" total size is about 15000 approx.

B. Data exploration

As we see in every airlines 16.1%,21.1%,62.7% are positive,negative and neutral tweets.Where negative tweets are highest that is in United.So we start exploring and see that in negative reasons which is given in data set among them Customer service issue is highest with 20% in visualization.Negative reasons are most in united and least negative reasons are in Virgin American.





C. Data Preproccessing

In this step i had remove the attribute who had more then 90% NULL values.After that in negative reasons 37.31% of data are remained NULL so this is replaced by the the relevant value "No Feedback" so that data should be cleaned without affecting the result.

As per observations there are several attributes are present in the data set which is irrelevant to me so i removed those ones on the domain knowlege and keep others as per my requirement. And the class labels which i observed are as per the polarity which i can find out using textblob so i remove all either polarity attribute and keep only 4 relevant attribute in which tweet is most important along with its class label.

1) Feature creation: In this step there is use of the concept of Natural language processing in which following steps are performed, It is done because we need cleaned tweet without any special character,meaningless words, and repeated words. a.Tokenization:In this step we breakdown the whole sentence into individual words. b.Removing Stop words:The words such as "the", "a", "an", "in" which will need in sentiment analysis are removed. c.Stemming:It is used to determine domain vocabularies in analysis. After that cleaning over tweet data new columns are formed "new_clean_tweet","new_clean_tweet" (length of total words in new_clean_tweet),"unigram tweet"(in this single words which or only relevant for sentiment analysis are present),"bigrams_tweet" (in this attribute bigrams(pair of words) of new_cleaned_attributes are present),"sentiment_in_binary" which is having numeric values for each labels positive,negative,neutral as 1,3,2 respectively so that it can pass in classifier for prediction.

2) Feature extraction: Bag of words:

as features in data set are all relevant words in which models are applied.Its a method of feature extraction, it describes occurence of words within the a sentence in a texted data.

In this project 2 methods are used to converting texted data into vectors or matrix because models can only works on numeric data.Following methods are:

1.CounterVectorization:

In this count of total number of items are present in the data which observed to be suppressed frequent words hence rare words are ignored result in less accuracy of model.

2.TfidVectorization:

to overcome the demerits of CounterVectorization,it take document weight of words as per definition.it deals with frequent words.it weight the word counts by how frequently they appear in document.

In this project there is use of 2 grams for good accuracy.Example: "She is not good", is having same score as "good".As each words in data having probability in this project while prediction we use two words as single word like "not good","is good" evaluates to the same score value.The number which assign to each word is called Score value.Stop words are also present which is having less values as compare to other.

III. Models

In this project following three classifications are used which will predict the labels of positive,negative and neutral classes of a tweets.In this project Scikit learn library are used and following models are used: 1.K-Nearest Neighbors(KNN) 2.Support Vector Machine 3.Multinomial Naive Bayes Model following section of models are elaborate about the above three classifier: why it is used in this data set?.

A. K-Nearest Neighbors

The K-nearest neighbor algorithm is a Supervised machine learning algorithm.It is very easy to implement and performs classification even though is complex.KNN does not assume anything about the data which in technical terms says non-parametric learning algorithm.It's called lazy learner.It did't have any specialised training phase.It uses the data point of whole data while training and and classifies the new data instances.It uses euclidean distance as a metric.It uses its nearest neighbor to predict the class label.

In text classification there is use of nltk libraries to generate similarities and similarity score which will taken out among all texts.In this model we find out the highest similarity score which among all the training data sets.

B. Multinomial Naive Bayes

It says to be baseline solution for sentimental anaylsis tasks.Naive bayes technique is find probabilities of classes which are assigned to document texts by using joint probabilities of all the labeled classes. Training multinomial classifier follow a sequence of technique as: vectorizer to transformer to Classifier is easy to work.It is using word frequency as a feature so on the basis of how many time that word will appear on the whole document it will classifies the tweet in

classes.The following is the basic probability formula for naive bayes classifier:

P(label|features)=P(label)*P(features|label)/P(features)

IV. Results

As a result of above sentimental analysis following accuracies are resulted

*TfidfVectorizer:
KNN classifier
Accuracy: 67.72%
precision: 59.19%
recall: 62.53%
f-score: 60.43%

multinomial classifier
Accuracy : 70.32%
precision: 79.83%
recall: 44.96%
f-score: 46.78%

*CountVectorizer:
KNN classifier
Accuracy: 52.04%
precision: 52.23%
recall: 55.68%
f-score: 49.69%

multinomial classifier
Accuracy : 78.65%
precision: 72.57%
recall: 70.42%
f-score: 71.34%

V. Conclusion

In this paper two different methods are applied but among them it is found that Multinomial classifier is accurate in both kind of vectorize.But Due to variation of the values of precision and recall in multinomial classifier in TfidfVectorizer we can say that this will return few results but the predicted results were always correct as compared to the tarining labels.

VI. Future Work

In this paper apart from comparing the results of two different way feature extraction, we can also analyse trigrams.We can also add more labels along with positive, negative or neutral sentiment.currently in this paper we analyse text but in future we can work with emoticons tweets also.We can also predict the reason of the sentiment by along with class labels positive,negative and neutral, like if we got negative sentiment for any tweet then we can also able to find its reasons from the ”negativereason” column in this data set.

References

[1] Kunal Lalwani, Kyle Lemaire, Darshan Mange, Donglin Lao, Mark Ledesma,https://github.com/kunal-lalwani/Twitter-US-Airlines-Sentiment-Analysis
[2] https://www.kaggle.com/carlolepelaars/predicting-sentiment-with-ml-80-accuracy
[3] Mandar Munagekar, Sai Harsha Nagalla Guide: Prof.Meiliu Lu
[4] https://github.com/kunal-lalwani/Twitter-US-Airlines-Sentiment-Analysis
[5] https://github.com/ruchitgandhi/Twitter-Airline-Sentiment-Analysis/blob/master/Twitter_Airline_Sentiment_Analysis.ipynb
[6] https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/
[7] https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/
[8] https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
[9] https://cse.iitk.ac.in/users/cs365/2015/_submissions/ajaysi/report.pdf
[10] Analytics vidya is used for many other informations also
[11] https://www.kaggle.com/parthsharma5795/comprehensive-twitter-airline-sentiment-analysis
[12] https://github.com/bsnpavan/NLP/blob/master/nlp_part-1_ml_way.py
[13] https://stackoverflow.com/questions/26266362/how-to-count-the-nan-values-in-a-column-in-pandas-dataframe
[14] https://www.geeksforgeeks.org/python-pandas-dataframe-fillna-to-replace-null-values-in-dataframe/
[15] https://www.kaggle.com/rashmitarouy01/us-airlines-twitter-sentiment-analysis
[16] https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/
[17] https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/
[18] https://stackoverflow.com/questions/43646877/python-extract-positive-words-from-a-string-using-sentiment-vader
[19] https://www.geeksforgeeks.org/text-preprocessing-in-python-set-1/
[20] https://stackoverflow.com/questions/18674064/how-do-i-insert-a-column-at-a-specific-column-index-in-pandas
[21] https://stackoverflow.com/questions/16327055/how-to-add-an-empty-column-to-a-dataframe
[22] https://stackoverflow.com/questions/53986877/pandas-iterate-over-a-row-and-adding-the-value-to-an-empty-column
[23] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[24] https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a
[25] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
[26] kaggle,stackoverflow,medium,towardscience were used alot while learning:https://towardsdatascience.com/text-classification-using-k-nearest-neighbors-46fa8a77acc5
https://www.kaggle.com/lbronchal/sentiment-analysis-with-svm
[27] https://www.google.com/search?q=is+it+possible+that+accuracy+is+high+and+precision+recall+is+lowoq=is+it+possible+that+accuracy+is+high+and+precision+recall+is+lowaqs=chrome..69i57.18517j0j7sourceid=chromeie=UTF-8
[28] https://www.programcreek.com/python/example/89260/sklearn.metrics.precision_recall_fscore_support