Sentiment Analysis Of Twitter Data (US Airline)

Bassem Ashraf Beshoy Amgad Hossam Mohammed

1 Introduction

Social Media Connections like Twitter ,Facebook and Instagram give Us Very Huge amount of Data from Users Write their Opinins about Products and services which give us alot of Ideas to do with it one of these ideas is Sentiment Analysis. Sentiments Analysis, means we can classify their opinins about a Particular service or Product in Nigative ,Positive and Natural.
in this paper we Use Sentiment Analysis yo Classify opinins about US airline it can help us to know the quality of this service and help managers to improve their company .

2 Realated Work

There is more compines Use Sentiment Analysis Like Intel, Twitter and IBM are now using sentiment-analysis software and similar technologies to determine employee concerns and, in some cases, develop programs to help improve the likelihood employees will stay on the job.

3 Methodology

In this paper we use tweets Written in english to see what people think about US airline in our project we get dataset from Kaggle link of dataset in Section **??** we get around **15.000** tweets talk anout US airline,after that we start clean Data and Build our models.

3.1 Preprocessing

after get our dataset we start clean data from not important words and unnecessary Symbols and delete usernames and Emojis after that we go to build model.

3.2 Model Building

After cleaning Dataset we want to calssify our Data in **3** Classes **(Positive, Nigative , Natural)** for 1000 tweets (already labeled) and train our models **Models :**
1- logistic regression
2-support victor machine(svm)
3-Decession tree
4-k neighbors (knn)
5-Naïve byes
To apply ensemple we use voting classifier between decession tree and logistic regression and in result section we will show a new accuracy.

4 Experments

we get data labeld as table below

Table 1: Labels of data set

tweets	Positive	Nigative	Natural
14640	16%	63%	21%

4.1 Preprocessing in Details

The whole idea of proprocessing is delete the Duplicated words which not effect on user felling.
in Table **1** we will see the Row tweet without any preprocessing which extracted from csv file now we Start first Step in preprocessing...

Table 2: Tweet Before preprocessing

Row tweet
@VirginAmerica it was amazing, and arrived an hour early. You're too good to me.

Step 1 :
in this setp we tokenize tweet which mean we will split it as words to remove unnecessary Symbols and redundunt words .

Table 3: split tweet

Tweet_tokenized
[, virginamerica, it, was, amazing, and, arrived, an, hour, early, you, re, too, good, to, me,]

Step 2 :
in this step we will delete Stopwords in english like 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself',....etc because it don't effect on felling of users when they write tweets. tweet became more clean in Table **4**

Table 4: Delete stopwords

tweet without stopwords
[, virginamerica, amazing,arrived, hour, early,good,]

Step 3 :
after Step 3 when we see data we found a word duplicated like "virginamerica" so we delete it tweet now more clean as we can see in Table **5**

Table 5: Delete Redundunt words

tweet without redundunt words
[, amazing, arrived, hour, early, good,]

Step 4 :
in stemming step we delete letters which not effect on tweet like 'ly , tion , ing' to make tweet mare clean Table **6**

Table 6: Stemming

Tweet_stemmed
[, amaz, arriv, hour, earli, good,]

Step 5 :
in last step we delete punctuations from tweet to get the last shape of tweet like in Table **7**

Table 7: Delete Panctuations

Tweet after preprocessing
amaz arriv hour earli good

5 Implementation

We used **Python** To implement this paper

5.1 Libraries an Dataset :

1-Sklearn:
to import CountVectorizer to vectrize our tweets , models such multinomial naive bayes,svm,DecissionTree,logistic

regression and linear regression,import metrices like classification_report metrices

2-Numpy and Pandas:
to make our processes on Data more easy

3-Dataset:
that was used -¿(twitts about airline in usa)data set from kaggle

5.2 the steps of the implemention :

1-import libraries
2-preprocessing: as we talked previously (delete duplicated words,imotions,stopwords and tokenize it)
3- preprocessing data : take the preprocessing twetts and vectorize it to matrix and labeled it
4-Split Data: from built in function in sklearn we split data into train and test for x and y and shuffle it.
5-Train models : After preprocessing data we vectorized tweets and pass it to models to train it
6-Enspmle learing: after fitting and print accuaracy will choose any 2 models and apply enspmle learing to improve accuracy
7-Evaluation Matrices : finally we use sklearn.metrices to print evaluation metices for each model to compare between them .

6 Results

After train our models we apply evaluation matrecis on it to test and get these results for every model.

Table 8: Knn Results

Knn	precision	recall	f1-score
0	0.54	0.65	0.59
1	0.43	0.52	0.47
2	0.84	0.74	0.79

Table 9: Logistic regression Results

LR	precision	recall	f1-score
0	0.83	0.60	0.69
1	0.66	0.47	0.55
2	0.81	0.94	0.87

Table 10: Decession Tree Results

DT	precision	recall	f1-score
0	0.61	0.53	0.57
1	0.43	0.39	0.41
2	0.78	0.83	0.80

Table 11: SVM Results

SVM	precision	recall	f1-score
0	0.83	0.60	0.70
1	0.72	0.41	0.52
2	0.79	0.96	0.87

Table 12: NaiveBayes Results

NaiveBayes	precision	recall	f1-score
0	0.92	0.17	0.29
1	0.75	0.17	0.27
2	0.69	0.99	0.81 height

we get acuuracy for models with test tweets as in next table

Table 13: Accuracy

Model	Accuracy
KNN	68.35%
LR	78.94%
DT	69.08%
SVM	78.92%
NavieBayes	69.72%

we use **ensemple learning** between Logistic regression with accuracy (78.94%) and Decession Tree withh accuracy (69.08%) we use voting classifier model and accuracy became (72.50%)

7 Conclusion and FutuerWork

Sentiment analysis is a field of study for analyzing opinions expressed in text in several social media sites. Our proposed model used several algorithms to enhance the accuracy of classifying tweets as positive, negative and neutral.and it can help managers and oweners in make decision and know opinins of customers to develop their services to get more positive reactions from customers .

8 Refrences

Dataset link :https://www.kaggle.com/crowdflower/twitter-airline-sentiment
Book:[Ian H. Witten,Eibe Frank]Data Mining Practical Machine Learning Tools and Techniques
Book: [Chris Albon] Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning
Book: Python Data Science Handbook Essential Tools for Working with Data