ÇANKAYA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

CENG 408
Sentiment Analysis of the Feedback from Airplane Passengers

FINAL REPORT

Merve Şirin – 201626404
İrem Şahingöz - 201611051

2020-2021

Contents

Abstract

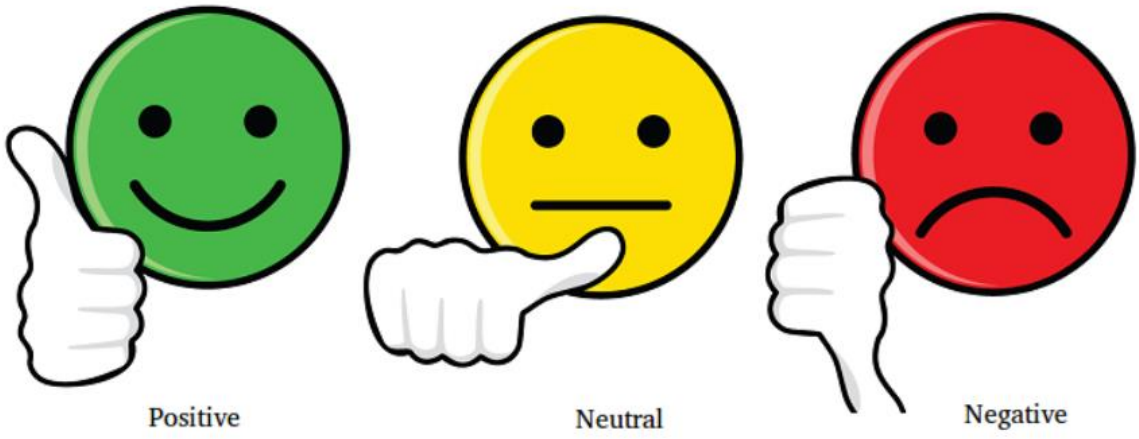
Airlines provide a wide range of services to their passengers such as meal service on board, luggage service, online check-in, etc. These services are aimed to attract more passengers and put the airline in a better position to compete with other companies.

From passengers perspective however, not all services are at an acceptable level. They complain about the delays, quality of service, etc. For an airline understanding the nature of these complains or satisfactions is of primary importance.

In this project, the tweets written by passengers of American airlines will be used to classify the services as good, bad, neutral. In addition, the co-occurrence of the reactions to these services will be investigates.

1. Introduction

By analyzing emotionally the comments made by American airport passengers on Twitter, we classified these comments as positive, negative, and neutral. Our aim to carry out this project is to determine the satisfaction level of airport passengers and to enable airport companies to provide better service. As a result of the evaluations, we made this project with Machine Learning algorithms, to help airport companies to provide better service to passengers by improving their services in a more accurate direction. We needed a lot of data to get a better accuracy rate on our project. We met our data needs thanks to the data set provided by our advisor. For our project, we used the SVM classifier, which we observed to give better accuracy than other methods. We performed Data Collection, Data Preparation, Text Preprocessing, Sentiment Annotation operations respectively. While doing Data Preparation, we performed the filtering process. While doing Text Preprocessing, we did the cleaning, normalizing, tokenizing, stopwords removal, and stemming operations. While doing Sentiment Annotations, we divided the emotions into positive, negative, and neutral. In order to achieve a high rate of accuracy, we paid attention to the use of effective methods and to the correct order of the operations we do. You can examine the methods and libraries we use in more detail in the following sections.



2. Sentiment Analysis

Sentiment analysis is a field of study that analyzes the ideas, feelings, evaluations, attitudes and feelings of entities such as products, services, organizations, individuals, problems, events, issues and their characteristics. [1] Sentiment analysis mainly focuses on ideas that express or imply positive or negative emotions. Research in sentiment analysis has a significant impact only on NLP (Natural Language Processing). It can also have a profound influence on management sciences, political science, economics and social sciences because they are all influenced by people's opinions. With the rapid growth of social media today, individuals and organizations are increasingly using the content on this media to make decisions. Automatic sentiment analysis systems are needed because long blogs and forum posts always contain a large volume of opinion text that cannot be easily deciphered.

3. Problem Definition of Sentiment Analysis

The biggest problem in sentiment analysis is analyzing implied or sarcastic sentences, because even if these sentences are negative in words, they may actually mean something positive in general. This great dilemma is the biggest problem facing synthesis analysis. Because here we have to go for a more rational analysis by using smarter artificial intelligence, not a purely word-oriented analysis. Only in this way, we can distinguish these sarcastic or suggestive sentences and perform a correct analysis.

Our main goal in this project is to make a very accurate sentiment analysis and to be distinguished by using artificial intelligence applications.

3.1. Opinion Definition

The most important feature of our opinions and thoughts is that they are subjective. The reason is that they contain only the commenter's thoughts. Therefore, researchers or product sellers collect much more person's opinions and try to understand what society generally thinks.

It is also reliable in terms of assessment and evaluation, and since many people from many platforms with twitter, facebook and forum applications express their opinions and thoughts that they share every day, we have a lot of opinion datas on that topic.

Twitter posts have short sentences compared to other forum comments and include some internet slangs in these sentences. But It is also easy to review this data because it is short. Because these comments are short, it means that they do not contain too much unnecessary information rather than forum applications.

Tweets have been useful for users with these features. We will also use twitter comments for sentiment analysis in our project.

4. Sentiment Analysis Process

4.1. Data Collection

Datas were collected from twitters. In this project 14640 tweets are used.

4.2. DataPreperation

This process includes filtering English comments on twitter. Because we want to focus on only English comments for our sentiment analysis, we filtered all tweets by using only English comments.

This process includes comment query. We use data which domains '#airline, #American Airline, #airline service'. These are our hastags. We used them to filter datas.
Also, we filter duplicated and irrelevant comments.
After that, we have datas for sentiment analysis. In figure 1 , First 29 datas are seen in cvs form.

4.3. Text Preprocessing:

This process have 4 basic steps:

I. Cleaning

II. Tokenizing

III. Stopwords removal

IV. Stemming

	A
1	tweet_id,airline_sentiment,airline_sentiment_confidence,negativereason,negativereason_confidence,airline,airline_sentiment_gold,name,negativereason_gold,retweet_count,text,tweet_coord,tweet_created,tweet_location,user_timezone
2	570306133677760513,neutral,1.0,,Virgin America,,cairdin,,0,@VirginAmerica What @dhepburn said.,,2015-02-24 11:35:52 -0800,,Eastern Time (US & Canada)
3	570301130888122368,positive,0.3486,,0.0,Virgin America,,jnardino,,0,@VirginAmerica plus you've added commercials to the experience... tacky.,,2015-02-24 11:15:59 -0800,,Pacific Time (US & Canada)
4	570301083672813571,neutral,0.6837,,,Virgin America,,yvonnalynn,,0,@VirginAmerica I didn't today... Must mean I need to take another trip!,,2015-02-24 11:15:48 -0800,Lets Play,Central Time (US & Canada)
5	57030103: they have little recourse",,2015-02-24 11:15:36 -0800,,Pacific Time (US & Canada)
6	570300817074462722,negative,1.0,Can't Tell,1.0,Virgin America,,jnardino,,0,@VirginAmerica and it's a really big bad thing about it,,2015-02-24 11:14:45 -0800,,Pacific Time (US & Canada)
7	570300767074181121,negative,1.0,Can't Tell,0.6842,Virgin America,,jnardino,,0,"@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.
8	it's really the only bad thing about flying VA",,2015-02-24 11:14:33 -0800,,Pacific Time (US & Canada)
9	570300616901320704,positive,0.6745,,0.0,Virgin America,,cjmcginnis,,0,"@VirginAmerica yes, nearly every time I fly VX this â€œear wormâ€™t go away :)",,2015-02-24 11:13:57 -0800,San Francisco CA,Pacific Time (US & Canada)
10	570300248553349120,neutral,0.634,,,Virgin America,,pilot,,0,"@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP",,2015-02-24 11:12:29 -0800,Los Angeles,Pacific Time (US & Canada)
11	570299953286942721,positive,0.6559,,,Virgin America,,dhepburn,,0,"@virginamerica Well, I didn't…but NOW I DO! :-D",,2015-02-24 11:11:19 -0800,San Diego,Pacific Time (US & Canada)
12	570295459631263746,positive,1.0,,,Virgin America,,YupitsTate,,0,"@VirginAmerica it was amazing, and arrived an hour early. You're too good to me.",,2015-02-24 10:53:27 -0800,Los Angeles,Eastern Time (US & Canada)
13	570294189143031808,neutral,0.6769,,0.0,Virgin America,,idk_but_youtube,,0,@VirginAmerica did you know that suicide is the second leading cause of death among teens 10-24,,2015-02-24 10:48:24 -0800,1/1 loner squad,Eastern Time (US & Canada)
14	57028972: 3 pretty graphics. so much better than minimal iconography. :D,,2015-02-24 10:30:40 -0800,NYC,America/New_York
15	57028958: I haven't €,,2015-02-24 10:30:06 -0800,NYC,America/New_York
16	570287408438120448,positive,0.6451,,,Virgin America,,mollanderson,,0,@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies again! U take all the #stress away from travel http://t.co/ahlXHhKiyn,,2015-02-24 10:21:28 -0800,,Eastern Time
17	570285904809598977,positive,1.0,,,Virgin America,,sjespers,,0,@VirginAmerica Thanks!,,2015-02-24 10:15:29 -0800,"San Francisco, CA",Pacific Time (US & Canada)
18	570282469121007616,negative,0.6842,Late Flight,0.3684,Virgin America,,smartwatermelon,,0,@VirginAmerica SFO-PDX schedule is still MIA.,,2015-02-24 10:01:50 -0800,"palo alto, ca",Pacific Time (US & Canada)
19	570277724385734656,positive,1.0,,,Virgin America,,ItzBrianHunty,,0,@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but great things about Virgin America. #29DaysToGo,,2015-02-24 09:42:59 -0800,west covina,Pac
20	570276917301137409,negative,1.0,Bad Flight,1.0,Virgin America,,heatherovieda,,0,@VirginAmerica I flew from NYC to SFO last week and couldn't fully sit in my seat due to two large gentleman on either side of me. HELP!,,2015-02-24 09:39:46 -0800,this pl
21	570270684619923457,positive,1.0,,,Virgin America,,thebrandiray,,0,I â¤ flying @VirginAmerica. â˜ºğŸ‘,,2015-02-24 09:15:00 -0800,Somewhere celebrating life. ,Atlantic Time (Canada)
22	570267956648792064,positive,1.0,,,Virgin America,,JNLpierce,,0,@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly with only you.,,2015-02-24 09:04:10 -0800,Boston	Waltham,Quito
23	570265883513384960,negative,0.6705,Can't Tell,0.3614,Virgin America,,MISSGJ,,0,@VirginAmerica why are your first fares in May over three times more than other carriers when all seats are available to select???,,2015-02-24 08:55:56 -0800,,
24	570264145116819457,positive,1.0,,,Virgin America,,DT_Les,,0,@VirginAmerica I love this graphic. http://t.co/UT5GrRwAaA,"[40.74804263, -73.99295302]",2015-02-24 08:49:01 -0800,,
25	570259420287868928,positive,1.0,,,Virgin America,,ElvinaBeck,,0,@VirginAmerica I love the hipster innovation. You are a feel good brand.,,2015-02-24 08:30:15 -0800,Los Angeles,Pacific Time (US & Canada)
26	57025882: LAS non stop permanently anytime soon?,,2015-02-24 08:27:52 -0800,"Boston, MA ",Eastern Time (US & Canada)
27	570256553502068736,negative,1.0,Customer Service Issue,0.3557,Virgin America,,ayeevickiee,,0,@VirginAmerica you guys messed up my seating.. I reserved seating with my friends and you guys gave my seat away ... ğŸ˜¡ I want free internet,,2015-02-24 0
28	570249102404923392,negative,1.0,Customer Service Issue,1.0,Virgin America,,Leora13,,0,@VirginAmerica status match program. I applied and it's been three weeks. Called and emailed with no response.,,2015-02-24 07:49:15 -0800,,
29	570239632807370753 negative 1.0 Can't Tell 0.6614 Virgin America meredithilynn 0 @VirginAmerica What happened ? ur vegan food options?! At least say on ur site so i know I won't be able 2 eat anything for next 6 hrs #fail 2015-02-24 07:11:37 -0800

In Figure 1: Collected First 29 Datas

We collected 14640 tweets and we used them for our Project. All tweets collected in a Excel file.

4.3.1. Cleaning

```
In [1]: #importing libraries
        import re
        import warnings
        import nltk
        from nltk.tokenize import TweetTokenizer #for tokenize text
        from nltk.stem.snowball import SnowballStemmer # for Stemming word
        #from nltk.stem.lancaster import LancasterStemmer
        import pandas as pd
        import nltk
        from nltk.tokenize import TweetTokenizer #for tokenize text
        from nltk.stem.snowball import SnowballStemmer # for Stemming word
        #from nltk.stem.lancaster import LancasterStemmer
```

```
In [2]: %matplotlib inline
        warnings.filterwarnings("ignore")
```

```
In [3]: tweet = pd.read_csv("/Users/mervesirin/Desktop/Tweets.csv") #reading tweet.csv file using pandas

        tweet
```

First of all, we read our Excel file which contains tweets . To read it, we need to import pandas.

Out[3]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN
...	...	...	...	...	...	...	...

Our process is text preperating so we want to focus on texts . We rewrite our tweets within 3 parts. They are text , sentiment and datatime. It is easy to analysize them.

```
In [4]: df=tweet.iloc[:,[10,1,12]]

        df.columns = ['text', 'sentiment','datetime'] #only two column from data text and sentiment
        data= df
        data
```

Out[4]:

	text	sentiment	datetime
0	@VirginAmerica What @dhepburn said.	neutral	2015-02-24 11:35:52 -0800
1	@VirginAmerica plus you've added commercials t...	positive	2015-02-24 11:15:59 -0800
2	@VirginAmerica I didn't today... Must mean I n...	neutral	2015-02-24 11:15:48 -0800
3	@VirginAmerica it's really aggressive to blast...	negative	2015-02-24 11:15:36 -0800
4	@VirginAmerica and it's a really big bad thing...	negative	2015-02-24 11:14:45 -0800
...	...	...	...
14635	@AmericanAir thank you we got on a different f...	positive	2015-02-22 12:01:01 -0800
14636	@AmericanAir leaving over 20 minutes Late Flig...	negative	2015-02-22 11:59:46 -0800
14637	@AmericanAir Please bring American Airlines to...	neutral	2015-02-22 11:59:15 -0800
14638	@AmericanAir you have my money, you change my ...	negative	2015-02-22 11:59:02 -0800
14639	@AmericanAir we have 8 ppl so we need 2 know h...	neutral	2015-02-22 11:58:51 -0800

14640 rows × 3 columns

pandas.DataFrame.iloc:

Purely integer-location based indexing for selection by position. .iloc[] is primarily integer position based (from 0 to length-1 of the axis), but may also be used with a boolean array.

4.3.2. Tokenizing

A token is, very simply, a piece of data that stands in for another, more valuable piece of information. Tokens have virtually no value on their own - they are only useful because they represent something bigger. Tokenization is the process of removing sensitive data from your business systems by replacing it with an undecipherable token and storing the original data in a secure cloud data vault. We use tokenizing to remove every thing exept texts.

In [5]:
```
# removing every thing except text
data['text']=data['text'].str.replace("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|([0-9])","")
# now tokenize text
data['text']=data['text'].apply(nltk.word_tokenize)
data
```

Out[5]:

	text	sentiment	datetime
0	[What, said]	neutral	2015-02-24 11:35:52 -0800
1	[plus, youve, added, commercials, to, the, exp...	positive	2015-02-24 11:15:59 -0800
2	[I, didnt, today, Must, mean, I, need, to, tak...	neutral	2015-02-24 11:15:48 -0800
3	[its, really, aggressive, to, blast, obnoxious...	negative	2015-02-24 11:15:36 -0800
4	[and, its, a, really, big, bad, thing, about, it]	negative	2015-02-24 11:14:45 -0800
...	...	...	...
14635	[thank, you, we, got, on, a, different, flight...	positive	2015-02-22 12:01:01 -0800
14636	[leaving, over, minutes, Late, Flight, No, war...	negative	2015-02-22 11:59:46 -0800
14637	[Please, bring, American, Airlines, to, BlackB...	neutral	2015-02-22 11:59:15 -0800
14638	[you, have, my, money, you, change, my, flight...	negative	2015-02-22 11:59:02 -0800
14639	[we, have, ppl, so, we, need, know, how, many,...	neutral	2015-02-22 11:58:51 -0800

14640 rows × 3 columns

4.3.3. Stemming

In linguistic morphology and information retrieval , stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root

A stemmer for English operating on the stem *cat* should identify such strings as *cats*, *catlike*, and *catty*.

A stemming algorithm might also reduce the words *fishing*, *fished*, and *fisher* to the stem *fish*. The stem need not be a word, for example the Porter algorithm reduces, *argue*, *argued*, *argues*, *arguing*, and *argus* to the stem *argu*.

Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

```
In [6]: #Stemming each word
stemmer = SnowballStemmer('english')
data['text']=data['text'].apply(lambda x: [stemmer.stem(y) for y in x])
data
```

Out[6]:

	text	sentiment	datetime
0	[what, said]	neutral	2015-02-24 11:35:52 -0800
1	[plus, youv, ad, commerci, to, the, experi, ta...	positive	2015-02-24 11:15:59 -0800
2	[i, didnt, today, must, mean, i, need, to, tak...	neutral	2015-02-24 11:15:48 -0800
3	[it, realli, aggress, to, blast, obnoxi, enter...	negative	2015-02-24 11:15:36 -0800
4	[and, it, a, realli, big, bad, thing, about, it]	negative	2015-02-24 11:14:45 -0800
...	...	...	...
14635	[thank, you, we, got, on, a, differ, flight, t...	positive	2015-02-22 12:01:01 -0800
14636	[leav, over, minut, late, flight, no, warn, or...	negative	2015-02-22 11:59:46 -0800
14637	[pleas, bring, american, airlin, to, blackberri]	neutral	2015-02-22 11:59:15 -0800
14638	[you, have, my, money, you, chang, my, flight,...	negative	2015-02-22 11:59:02 -0800
14639	[we, have, ppl, so, we, need, know, how, mani,...	neutral	2015-02-22 11:58:51 -0800

14640 rows × 3 columns

4.3.4. Stopwords Removal

To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK. In the script above, we first import the stopwords collection from the nltk. corpus module. Next, we import the word_tokenize() method from the nltk.

```
In [7]: # removing stopword
stopwords = nltk.corpus.stopwords.words('english')
data['text']=data['text'].apply(lambda x: [y for y in x if y not in stopwords])
data
```

Out[7]:

	text	sentiment	datetime
0	[said]	neutral	2015-02-24 11:35:52 -0800
1	[plus, youv, ad, commerci, experi, tacki]	positive	2015-02-24 11:15:59 -0800
2	[didnt, today, must, mean, need, take, anoth, ...	neutral	2015-02-24 11:15:48 -0800
3	[realli, aggress, blast, obnoxi, entertain, gu...	negative	2015-02-24 11:15:36 -0800
4	[realli, big, bad, thing]	negative	2015-02-24 11:14:45 -0800
...	...	...	...
14635	[thank, got, differ, flight, chicago]	positive	2015-02-22 12:01:01 -0800
14636	[leav, minut, late, flight, warn, communic, mi...	negative	2015-02-22 11:59:46 -0800
14637	[pleas, bring, american, airlin, blackberri]	neutral	2015-02-22 11:59:15 -0800
14638	[money, chang, flight, dont, answer, phone, an...	negative	2015-02-22 11:59:02 -0800
14639	[ppl, need, know, mani, seat, next, flight, pl...	neutral	2015-02-22 11:58:51 -0800

14640 rows × 3 columns

5. Feature Construction

5.1. Word Cloud Representation of Most Frequency Words

```
In [9]: neg_tweets = data[data['sentiment'] == 0]
neg_string = []
for t in neg_tweets['text']:
    neg_string.append(t)
neg_string = pd.Series(neg_string).str.cat(sep=' ')
from wordcloud import WordCloud

wordcloud = WordCloud(width=1600, height=800,max_font_size=200).generate(neg_string)
import matplotlib.pyplot as plt
plt.figure(figsize=(12,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

In [14]:
```
pos_tweets = data[data['sentiment'] == 2]
pos_string = []
for t in pos_tweets['text']:
    pos_string.append(t)
pos_string = pd.Series(pos_string).str.cat(sep=' ')
from wordcloud import WordCloud

wordcloud = WordCloud(width=1600, height=800,max_font_size=200).generate(pos_string)
import matplotlib.pyplot as plt
plt.figure(figsize=(12,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

In [19]:
```
neutral_tweets = data[data['sentiment'] == 1]
neutral_string = []
for t in neutral_tweets['text']:
    neutral_string.append(t)
neutral_string = pd.Series(neutral_string).str.cat(sep=' ')
from wordcloud import WordCloud

wordcloud = WordCloud(width=1600, height=800,max_font_size=200).generate(neutral_string)
import matplotlib.pyplot as plt
plt.figure(figsize=(12,10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

In [15]:
```
pos_tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2363 entries, 1 to 14635
Data columns (total 3 columns):
text          2363 non-null object
sentiment     2363 non-null int64
datetime      2363 non-null object
dtypes: int64(1), object(2)
memory usage: 73.8+ KB
```

In [16]:
```
neg_tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9178 entries, 3 to 14638
Data columns (total 3 columns):
text          9178 non-null object
sentiment     9178 non-null int64
datetime      9178 non-null object
dtypes: int64(1), object(2)
memory usage: 286.8+ KB
```

In [22]:
```
neutral_tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3099 entries, 0 to 14639
Data columns (total 3 columns):
text          3099 non-null object
sentiment     3099 non-null int64
datetime      3099 non-null object
dtypes: int64(1), object(2)
memory usage: 96.8+ KB
```

In [11]:
```
for t in neg_tweets.text[:200]:
    if 'bag' in t:
        print (t)
```

```
everything fine lost bag
landed LAX hour I Your Late Flight bag check business travel friendly nomorevirgin
I cant check add bag Your website isnt working Ive tried desktop mobile
Let scanned passengers leave plane told someone remove bag st class bin uncomfortable
must traveler miss flight Late Flight check bag I missed morning appointments lost business sfolax
Late Flight bag check lost business I missed flight AM apt Three people flight exp
Is normal receive reply Central Baggage baggageissues smh
Were flight Vegas Boston today checked online bag count didnt register Can I fix somehow
luggage severely dentedmissing wheel coming baggage claim SAN Luggage agent Miranda I think wasnt help
Had spend hours worrying items carryon would brokenstolen since I couldnt carry plane lock bag
All group E told room bins I got plane room least bags row
Thanks making flight LAX JFK nightmare forcing check carry bag gate
I sure I drive total hours get bag Id like explain debacle one wants talk
```

In [13]:
```
for m in neg_tweets.text[:200]:
    if 'time' in m:
        print (m)
```

```
first fares May three times carriers seats available select
Hey first time flyer next week excited But Im hard time getting flights added Elevate account Help
need change flight thats scheduled hours min wait time phone Im calling intern Help
please provide status flight I cant imagine time Web indicates dude weather andor Dallas
Flight BOS gt LAS tomorrow Cancelled Flightled No notification wait times hour Will rebook another airline
hold times call center bit much
cant check site looks like every time loads
Avis rental continue button doesnt work website book car Tried times phone This sucks
I tried You offered charge additional k new ticket stranded Thurs st time last time
flight scheduled pm departure still says time plane gate Any update long delay
internet great thing I emailing executives company maybe respond timely manner
```

In [20]:
```
for m in pos_tweets.text[:200]:
    if 'awesome' in m:
        print (m)
```

```
know would amazingly awesome BOSFLL PLEASE I want fly
awesome I flew yall Sat morning Any way correct bill
completely awesome experience last month BOSLAS nonstop Thanks awesome flight depart time VAbeatsJblue
Thanks much awesome support guys rock
awesome deals DALAUS way
Flying LAX SFO looking awesome movie lineup I actually wish I long haul
another awesome new plane flight extremely nice Captain Steve Connolly
```

In [21]:
```
for m in pos_tweets.text[:200]:
    if 'customer' in m:
        print (m)
```

```
amazing customer service RaeAnn SF shes best customerservice virginamerica flying
absolute best team customer service ever Every time I fly Im delighted Thank
best customer service rep world irmafromDallas takes cake
thank checking tickets purchased customer happy
Thanks great customer service today amp helping get travel sorted
INCREDIBLE customer service Ive ever experienced So refreshing
thanks Yall best customer service left industry
Thanks taking care MR Happy customer
Wow What deal Again plus seats available Easy change make customer happy
Definitely compliment I really thought bag lost sent another airport In end I happy customer
Thanks Karen Salisbury IAH amazing customer service Found daughters bag lost UA Made day
I left comment customer care Thanks contacting
```

6. Sentiment Annotation

```
sentiment_objects = [TextBlob(tweets_data) for tweets_data in tweets_data['text']]

sentiment_objects[0].polarity, sentiment_objects[0]
```

In this code statement, We created textblob objects of the tweets. Using the TextBlob library, our aim is to understand whether the text contains positive or negative content.

```
sentiment_values = [[tweets_data.sentiment.polarity, str(tweets_data)] for tweets_data in sentiment_objects]

sentiment_values[0]
```

In this code statement, We created list of polarity values and tweet tex.

```
sentiment_df = pd.DataFrame(sentiment_values, columns=["polarity", "tweets"])

sentiment_df.head()
```

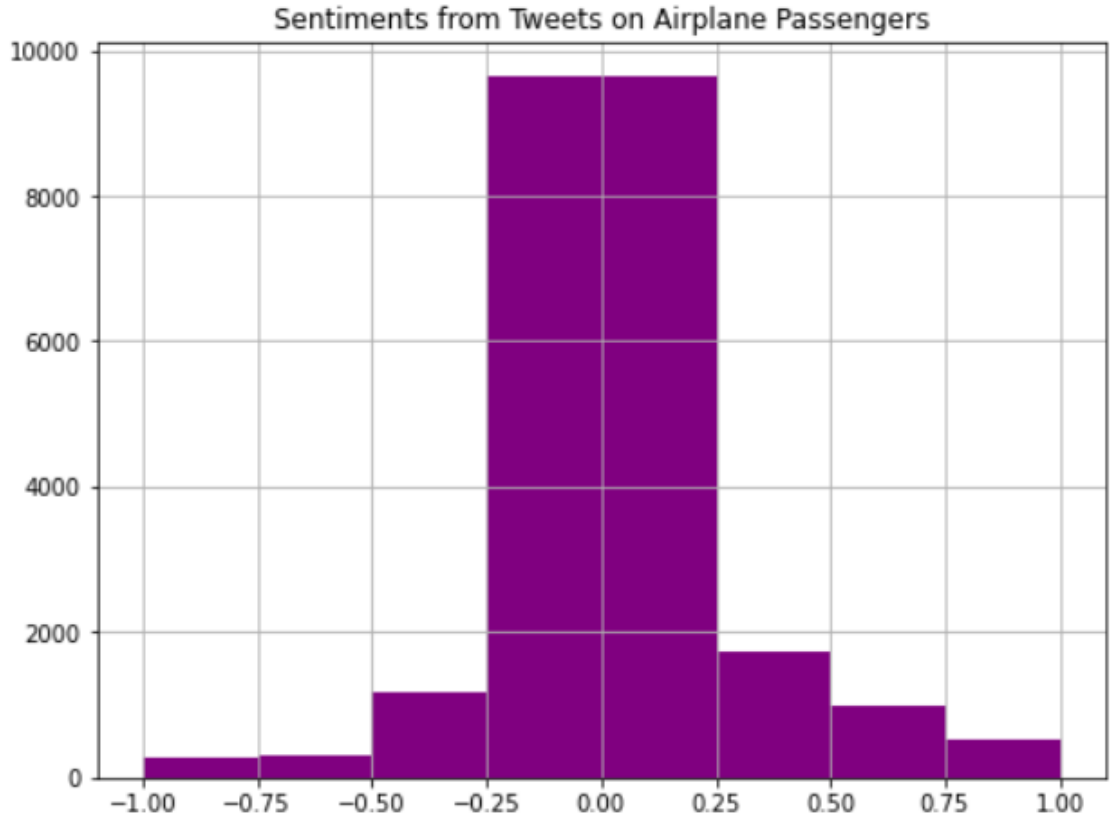
	polarity	tweets
0	0.000000	@VirginAmerica What @dhepburn said.
1	0.000000	@VirginAmerica plus you've added commercials t...
2	-0.390625	@VirginAmerica I didn't today... Must mean I n...
3	0.006250	@VirginAmerica it's really aggressive to blast...
4	-0.350000	@VirginAmerica and it's a really big bad thing...

In this code statement, We created data frame containing the polarity value and tweet text.

```
fig, ax = plt.subplots(figsize=(8, 6))

sentiment_df.hist(bins=[-1, -0.75, -0.5, -0.25, 0.25, 0.5, 0.75, 1],
             ax=ax,
             color="purple")

plt.title("Sentiments from Tweets on Airplane Passengers")
plt.show()
```



In this code statement, We plotted histogram of the polarity values.

```
sentiment_df = sentiment_df[sentiment_df.polarity != 0]

sentiment_df.head()
```

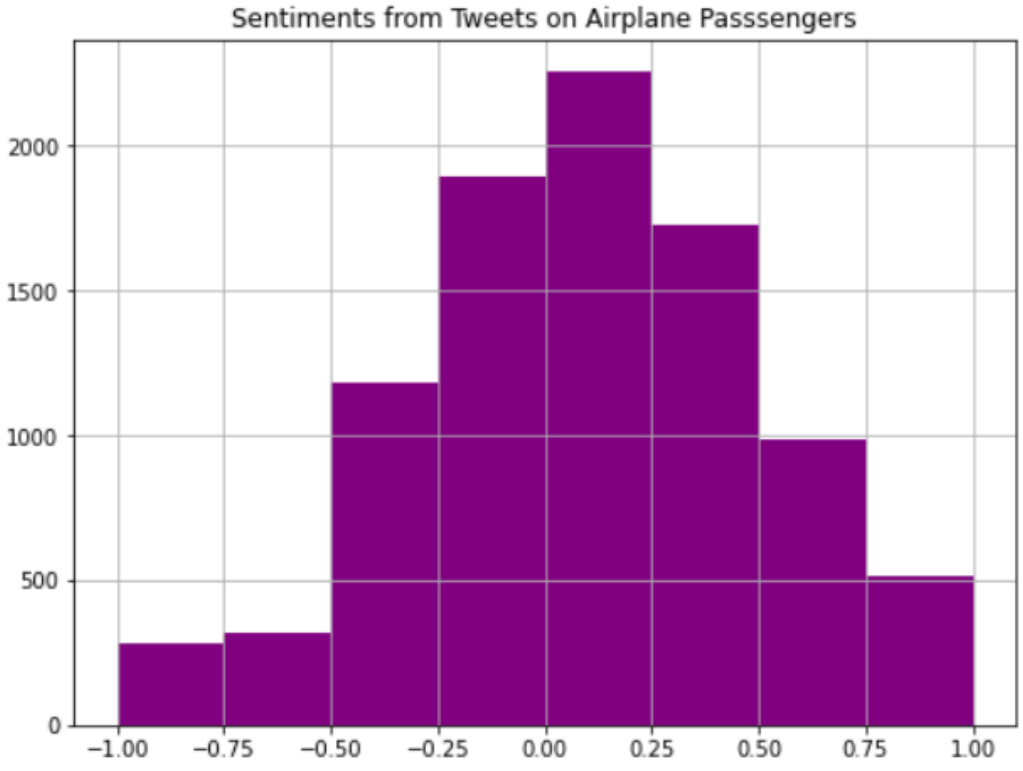
	polarity	tweets
2	-0.390625	@VirginAmerica I didn't today... Must mean I n...
3	0.006250	@VirginAmerica it's really aggressive to blast...
4	-0.350000	@VirginAmerica and it's a really big bad thing...
5	-0.208333	@VirginAmerica seriously would pay $30 a fligh...
6	0.466667	@VirginAmerica yes, nearly every time I fly VX...

In this code statement, We removed polarity values equal to zero.

```
fig, ax = plt.subplots(figsize=(8, 6))

sentiment_df.hist(bins=[-1, -0.75, -0.5, -0.25, 0.0, 0.25, 0.5, 0.75, 1],
             ax=ax,
             color="purple")

plt.title("Sentiments from Tweets on Airplane Passsengers")
plt.show()
```



In this code statement, We plotted histogram with break at zero.

7. Supervised Classification

```
VectorizedData = cntVectorizer.fit_transform(tweets_data.text)

IndexedData = hstack((np.array(range(0,VectorizedData.shape[0]))[:,None], VectorizedData))
```

Transform is used to adapt the operation 0 and 1 (representing the documents in the index) we have done above to our data. cntVectorizer converts matrices into 0 and 1. For example, counts words that are used, such as 2 times good, 0 sad, 1 happy, etc.

```
def sentiment(emotion):
    return {
        'negative': 0,
        'neutral': 1,
        'positive' : 2
    }[emotion]

targets = tweets_data.airline_sentiment.apply(sentiment)
```

Positive, negative, and neutral emotion words are numbered 0, 1, and 2, respectively. Then we adapted this to our data.

```
from sklearn.model_selection import train_test_split
data_train, data_test, targets_train, targets_test = train_test_split(IndexedData, targets, test_size=0.2, random_state=0)

data_train_index = data_train[:,0]
data_train = data_train[:,1:]
data_test_index = data_test[:,0]
data_test = data_test[:,1:]
```

We divided the train and test parts into 80% train and 20% test data. data_train_index is all the rows and only the 0. column. data_train is all the rows and all the columns except the 1. column. data_test_index is all the rows and only the 0. column. data_test is all the rows and all the columns except the 1. column.

```
from sklearn.svm import SVC

model = SVC()
model.fit(data_train, targets_train)
predictions = model.predict(data_test)
```

We trained our SVM model with our train data. In other words, our model is an SVC (Support Vector Classifier) classifier produced with a linear kernel.

```
from sklearn.metrics import accuracy_score

acc_score = accuracy_score(targets_test, predictions)
acc_score
```

```
0.7810792349726776
```

We adapted our model, which we created using the sklearn library using train data, to our model under the name of predictions by fitting it with test data. We adapted this estimate with our targets test data and made an accuracy score account using the sklearn library.

References

[1] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[2] Zhai, Zhongwu, Bing Liu, Hua Xu, and Peifa Jia. Clustering Product Features for Opinion Mining. in Proceedings of ACM International Conference on Web Search and Data Mining (WSDM-2011). 2011. 385. Zhai, Zhongwu, Bing Liu, Hua Xu, and Peifa Jia. Constrained LDA for Grouping Product Features in Opinion Mining. in Proceedings of PAKDD-2011. 2011.

[3] Zhang, Lei and Bing Liu. Extracting Resource Terms for Sentiment Analysis. in Proceedings of IJCNLP-2011. 2011a.

[4] Liu, Bing. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2006 and 2011: Springer.

[5] 296. Sarawagi, Sunita. Information extraction. Foundations and Trends in Databases, 2008. 1(3): p. 261-377.

[6] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.