An approach: Airlines service feedback using Sentiment Analysis

Anuj Khasgiwala
Utah State University
4205 Old Main Hill
Logan, Utah 84321
anujkhasgiwala@aggiemail.usu.edu

Shobhit Mundra
Utah State University
4205 Old Main Hill
Logan, Utah 84321
smundra42@gmail.com

ABSTRACT

Whenever we travel using airlines, we are sometimes happy and sometimes have issues with the services provided by airlines industry. In airline service industry, it is very hard to collect feedback given by customers using questionnaires to airlines, but social media like Twitter provides service to customers to share their voice. We can apply Sentiment Analysis to do customer tweets sentiment analysis. However, little research has been done in the domain of Twitter sentiment classification about airline services. In this paper, we are using Non-negative matrix factorization (NMF or NNMF), and collected a dataset of more than 10,000 tweets which are passengers reviews in the form of tweets and compared the reviews of year 2015 and 2017. The results showed that the airline companies are paying close attention to the reviews submitted by the passengers.

KEYWORDS

Twitter data mining, Sentiment Analysis, Airline service comparison, Tweets scrapping, Topic Modeling, NNMF, NMF, Non-negative Matrix Factorization, Vectors.

1 INTRODUCTION

Public transportation plays very important role in traveling from one place to another. With the increase of population and expenses of everything, people are using public transportation and local communities. Public transportation consists of variety of modes, some of them are Buses, Cars, Airlines. Customers are the most important part of this industry, because every mode of transportation depends on their customers —without customers, these modes of transportation would not exist.

Providing the best services should be the biggest motive of the public transportation. This can be done by making the customer feel more special with a personalized experience or sending a follow up e-mail, for example. Emirates, for instance, provides the so called fiKnowledge —driven Inflight Servicefi, which makes it possible for the airline crew to review previous trips customers have taken with the carrier before. This is how airlines will know about customerfis preferences and issues that may have occurred during their previous travels. Based on these, improvements can be made and personalized service can be provided to all the customers.

Currently most of the airlines providers has feedback portal on their websites. Passenger can post their thoughts, ideas about their journey. This thoughts & ideas are regarding the services like costs, food, delay time in time of arrival & departure etc. These reviews page can be used by other passengers to decide to travel this airlines or not. But the airlines do not release these customers reviews and feedback to the customers.

In such scenarios, social media and the third-party airlines provider proves to be very effective. Social media provides the facility to all the users to share their views or voices open for all the customers. These can be used to create a new rating services. Sentiment analysis is one of the most trusted and helpful technique. Out of so many social media services available, we have selected Twitter.

Sentiment[1] classification techniques can help researchers and decision makers in airline companies to better understand customers feeling, opinions and satisfaction. Researchers and decision makers can utilize these techniques to automatically collect customers' opinions about airline services from various micro-blogging platforms like Twitter. Business analysis applications can be developed from these techniques as well. Sentiment Analysis is done using five classifiers including Nave Bayesian classifier, Support Vector Machine (SVM) classifier, Bayesian Network classifier, C4.5 Decision Tree classifier and Random Forest classifier. In this paper, we are using Non-negative matrix factorization (NMF or NNMF), to perform topic modeling. Topic modeling works on TF-IDF technique to find the top topics of the documents.

To obtain or compare with the ground truth, we have collected the airline reviews from kaggle & twitter compare. Kaggle is used for 2015 data, while the twitter is used for the most recent tweets. In the reviews collected from the twitter, we have categorized the words into different reasons like customer services, food, time delay etc.

The paper is organized as follows. In section 1, the motivation for this research are explained and the objective is introduced. In section 2, the relevant previous work are summarized. Section 3 presents the data preprocessing,including the data collection, and data pre-processing and cleaning with the sentiment analysis values. In section 4, visual & interesting analysis is done. In section 5, methods that we have used to achieve our goal, Section 6 provides the conclusion, which summarizes our findings from this research.

2 RELATED WORKS

Sentiment classification is a division of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine[2]. But the special characters of sentiment expression in language make it very different from standard factual-based textual analysis [3].

The simplest way to do sentiment classification is using the Lexicon-based approach [1], which calculates the sum of the number of the positive sentiment words and the negative sentiment words appearing in the text file to determine the sentiment of the text file. The weakness of this approach is poor recognition of affect when negation is involved [3].

Big Data social data analysis has been very popular [9]. Because Twitter provide public access to its streaming and historical data, it has become a very popular data source for sentiment analysis and many work has been done in this area. J.Read used emoticons, such as ':—)' and ':—(', to collect tweets with sentiments and to categorize them into positive tweets and negative tweet. They adopted Nave Bayesian approach and the Support Vector Machine approach, both of which reached accuracy up to 70% [4]

Little work has been done on twitter sentiment classifications about airline services. Conventional sentiment classification approaches, such as Nave Bayesian approach, have been applied to some tweet data and the performance was not bad [5]. Twitter as the data source to analyze consumersﬁ communications about airline services. They studied tweets from three airline brands: Malaysia Airlines, JetBlue Airlines and SouthWest Airlines. They adopted conventional text analysis methods in studying Twitter usersﬁ interactions and provided advices to airline companies for micro-blogging campaign. In their research, they didnﬁt adopt sentiment classification on tweets, which will be more salient for airline services companies to understand what customers are thinking. This handbook suggests retrieving real time tweets from Twitter API with queries containing airline companiesﬁ names. The sentiment lexicons in this method are not domain specific and there is no data training process or testing process. By matching each tweet with the positive word list and the negative word list, and assigning scores based on matching result to each tweet, they can be classified as positive or negative according to the summed scores. The accuracy is unknown since it is not considered in this book. In our work, this method was applied and tested with labeled data. It can yield inaccurate testing results because sentiment classifications are highly domain specific. Adeborna et al adopted Correlated Topics Models (CTM) with Variational Expectation-Maximization (VEM) algorithm [6]. Their lexicons for classification were developed with Airline Quality Rating (AQR) criteria. In Sentiment detection process, the performances of the SVM classifier, the Maximum Entropy classifier and Naive Bayesian classifier were compared and Naive Bayesian classifier was adopted. Besides that, tweets are categorized by topics using the CTM with the VEM algorithm. In our work, more than 100,000 tweets are collected, and NNMF, which is much less biased. Besides that, their work did not present details about the classification approaches and comprehensive evaluations. However, our work not only contains the analysis of tweets with different sentiments but also includes the comparison of the performance of different airlines.

3 DATA PREPROCESSING

3.1 Data Collection

We required 2 different kinds of datasets, in which one is considered as the ground truth to verify the validity and authenticity of our solution. The first dataset is the collection of the tweets for the year 2015 for all the airlines provided by Kaggle as shown in the **Figure 1**. This dataset consisted of all the tweet details along with the sentiment analysis of each tweet and the major reason for the tweet.

tweet_id	airline_se	airline_se	negativer	negativer	airline	airline_se	name	negativer	retweet_c	text	tweet_co	tweet_cre	tweet_loc	user_timezone
5.7E+17	neutral	1			Virgin America		cairdin		0	@VirginAmerica Wh		########		Eastern Time (US & Canada)
5.7E+17	positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plu		########		Pacific Time (US & Canada)
5.7E+17	neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I di		########	Lets Play	Central Time (US & Canada)
5.7E+17	negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's		########		Pacific Time (US & Canada)
5.7E+17	negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and		########		Pacific Time (US & Canada)
5.7E+17	negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginA		########		Pacific Time (US & Canada)
5.7E+17	positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes		########	San Franci	Pacific Time (US & Canada)
5.7E+17	neutral	0.634			Virgin America		pilot		0	@VirginAmerica Rea		########	Los Angel	Pacific Time (US & Canada)
5.7E+17	positive	0.6559			Virgin America		dhepburn		0	@virginamerica Wel		########	San Diego	Pacific Time (US & Canada)
5.7E+17	positive	1			Virgin America		YupitsTate		0	@VirginAmerica it w		########	Los Angel	Eastern Time (US & Canada)
5.7E+17	neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did		########	1/1 loner	Eastern Time (US & Canada)
5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I &I		########	NYC	America/New_York
5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica Thi		########	NYC	America/New_York
5.7E+17	positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @v		########		Eastern Time (US & Canada)
5.7E+17	positive	1			Virgin America		sjespers		0	@VirginAmerica Tha		########	San Franci	Pacific Time (US & Canada)
5.7E+17	negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SFC		########	palo alto,	Pacific Time (US & Canada)
5.7E+17	positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So		########	west covi	Pacific Time (US & Canada)
5.7E+17	negative	1	Bad Flight	1	Virgin America		heatherovieda		0	@VirginAmerica I fl		########	this place	Eastern Time (US & Canada)
5.7E+17	positive	1			Virgin America		thebrandiray		0	I âœ¸ flying @VirginA		########	Somewhe	Atlantic Time (Canada)
5.7E+17	positive	1			Virgin America		JNLpierce		0	@VirginAmerica you		########	Boston | \	Quito
5.7E+17	negative	0.6705	Can't Tell	0.3614	Virgin America		MISSGJ		0	@VirginAmerica why		########		
5.7E+17	positive	1			Virgin America		DT_Les		0	@VirginA	[40.74804	########		
5.7E+17	positive	1			Virgin America		ElvinaBeck		0	@VirginAmerica I lo		########	Los Angel	Pacific Time (US & Canada)
5.7E+17	neutral	1			Virgin America		rjlynch21086		0	@VirginAmerica will		########	Boston, M	Eastern Time (US & Canada)
5.7E+17	negative	1	Customer	0.3557	Virgin America		ayeevickiee		0	@VirginAmerica you		########	714	Mountain Time (US & Canada)
5.7E+17	negative	1	Customer	1	Virgin America		Leora13		0	@VirginAmerica stat		########		
5.7E+17	negative	1	Can't Tell	0.6614	Virgin America		meredithjlynn		0	@VirginAmerica Wh		########		
5.7E+17	neutral	0.6854			Virgin America		AdamSinger		0	@VirginAmerica do		########	San Franci	Central Time (US & Canada)
5.7E+17	negative	1	Bad Flight	1	Virgin America		blackjackpro911		0	@VirginA	[42.36101	########	San Mateo, CA & Las Vegas, NV	
5.7E+17	neutral	0.615		0	Virgin America		TenantsUpstairs		0	@VirginA	[33.94540	########	Brooklyn	Atlantic Time (Canada)
5.7E+17	negative	1	Flight Boo	1	Virgin America		jordanpichler		0	@VirginAmerica hi!		########		Vienna
5.7E+17	neutral	1			Virgin America		JCervantezzz		0	@VirginAmerica Are		########	California	Pacific Time (US & Canada)
5.7E+17	negative	1	Customer	1	Virgin America		Cuschoolie1		0	@VirginA	[33.94209	########	Washingt	Quito
5.7E+17	negative	1	Customer	1	Virgin America		amanduhmccarty		0	@VirginAmerica awa		########		Pacific Time (US & Canada)
5.7E+17	positive	1			Virgin America		NorthTxHomeTeam		0	@VirginA	[33.21450	########	Texas	Central Time (US & Canada)

Figure 1: 2015 Tweets above

The second dataset we scraped, consisted of the most recent tweets of users for all airlines as shown in the **Figure 2**. We used 'Tweepy' library of python for extracting the dataset from Twitter. The tweepy library requires 'Oauth' which requires 'Consumer Key' and 'Consumer Secret' with 'Access Token'. The tweepy api takes 'query parameters', 'since' and 'items' which is number of tweets required. We set the item value as 10,000 tweets due to the request restrictions by 'Twitter'.

3.2 Data cleansing and preprocessing

The datasets obtained using both the process were almost cleaned and preprocessed. The only problem was in 2017 dataset we had to remove "#", "@" and "hyperlinks" for calculating the polarity or sentiments of the twitter tweets text.

For calculating the sentiment of the twitter tweets text we used 'Textblob' and 'nltk', python libraries. The csv file is read using csv library and the text column is then processed to remove "#", "@" and "hyperlinks". The sentiment analysis tokenizes a sentence and remove stop words from the text and returns the polarity value of each sentence. We attached the polarity value for each tweet as shown in the **Figure 3**.

These values are then categorized into:

(1) Positive - values greater than 0
(2) Negative - values less than 0

tweetID	tweetText	tweetRet	tweetFav	tweetSou	tweetCre	userID	userScree	userName	userCreat	userDesc	userFollo	userFrien	userLocati	userTimezone
8.52E+17	RT @Mike	1	0	Twitter fo	########	4.11E+08	guille_vel	Guille!	########	Las mejor	157	152	Guatemal	Central Time (US & Canada)
8.52E+17	RT @xado	54671	0	Twitter fo	########	1.15E+09	jackiee_m	less-ooh-	########	hey I foun	265	123	somewhe	Pacific Time (US & Canada)
8.52E+17	RT @xado	54671	0	Twitter fo	########	1.64E+08	hawkguys	jo the ho	########	this was s	3549	104		Central Time (US & Canada)
8.52E+17	How the S	0	0	IFTTT	########	7.20E+17	DRL_USAN	Daily New	########		1653	1216		
8.52E+17	Hey @Del	0	0	Twitter W	########	16227429	RipeInc	Len Roma	########	Owner of	8576	6955	Albuquer	Mountain Time (US & Canada)
8.52E+17	Which UN	0	0	Twitter fo	########	3.99E+09	Peacehav	Peacehav	########	the 'erber	162	499	City of Brighton & Hove	
8.52E+17	https://t.	0	0	IFTTT	########	33799339	Reeeemix	Reeeemix	########	Reeeemix	1762	4	USA	Central Time (US & Canada)
8.52E+17	RT @OhSc	3	0	Twitter fo	########	4.15E+08	erianna16	Erianna	########		970	423		Pacific Time (US & Canada)
8.52E+17	RT @xado	54671	0	Twitter fo	########	2.47E+09	iminesn	inÃªs	########		599	377		Lisbon
8.52E+17	United Air	0	0	WordPres	########	2.88E+09	theamedp	The Amec	########	Breaking	1189	0	DiyarbakÄ	Pacific Time (US & Canada)
8.52E+17	RT @fruit	1	0	Twitter fo	########	1.13E+09	uhlektruh	emo emu	########	pnw gay.	263	299		Pacific Time (US & Canada)
8.52E+17	It's a prob	0	0	Twitter fo	########	2.27E+09	Laurab4re	L. M. Blair	########	I punch N	3685	3426		
8.52E+17	RT @Steve	684	0	Twitter fo	########	2.52E+09	charliehin	Timothyð	########	Jeremiah	359	249		
8.52E+17	Why must	0	0	Twitter fo	########	2.86E+09	rezigler	Richard Zi	########		24	83		
8.52E+17	RT @matt	27	0	Twitter fo	########	15753253	jeffsleasm	Jeff Sleas	########	Skinny kic	1159	2121	216	America/Chicago
8.52E+17	RT	16	0	Twitter fo	########	3.71E+08	alltheway	MayItwee	########	Shakamak	342	249	Jasonville	
8.52E+17	RT @xado	54671	0	Twitter fo	########	2.84E+09	RICECAKE	jimin the	########	your local	1251	74	srâ,,¢ mal	Pacific Time (US & Canada)
8.52E+17	Why do ai	0	0	Twitter fo	########	1.74E+08	kwansfull	derek kwa	########	building p	415	175	los angele	Pacific Time (US & Canada)
8.52E+17	RT @xado	54671	0	Twitter fo	########	8.01E+08	wvngg	wvngg	########	living for l	596	573	Isla Vista,	Arizona
8.52E+17	RT @rollca	2	0	Twitter fo	########	19262912	careerfed	CareerFec	########	Gun Owne	1618	2234	DC-SFO-S	Eastern Time (US & Canada)
8.52E+17	Everyone	0	0	Twitter fo	########	2.43E+09	hawtcoco	habibi	########	it's a habi	145	167	Portland,	Arizona
8.52E+17	Me critica	0	0	Facebook	########	7.7E+08	nrcoleptc	A Narcole	########		17	0		
8.52E+17	#United A	0	0	SocialCha	########	2.27E+09	Videos_Fc	Videos Fo	########	Celebrity	176	414	United Sta	Arizona
8.52E+17	RT @Abuk	12	0	Twitter fo	########	3.03E+09	moirbad	Mohamed	########	Entrepren	306	83	Somalia	
8.52E+17	@GibiAsm	0	0	Twitter fo	########	2.31E+09	Blank_Le_	Blank	########	Like most	39	84		
8.52E+17	RT @fill3u	1612	0	Twitter fo	########	2.33E+08	DarkSamu	Patrick H.	########	Gamer/A	579	227	SR388	Atlantic Time (Canada)
8.52E+17	This is the	0	0	Hootsuite	########	15576134	teresajenl	Teresa Jer	########	"The very	2539	2313	Salt Lake (	Mountain Time (US & Canada)
8.52E+17	This is the	0	0	Hootsuite	########	1.28E+08	WriteOnP	Write On!	########	Write On!	1127	1496	Salt Lake (	Mountain Time (US & Canada)
8.52E+17	This	0	0	TweetDec	########	2.92E+09	bayside_j	BAYSIDE J	########	With a ho	592	290	Mumbai	Pacific Time (US & Canada)
8.52E+17	RT	172	0	Twitter fo	########	4.59E+08	MattEyre1	Matt Eyre	########	Handsom	242	380	Chaddert	Pacific Time (US & Canada)
8.52E+17	@wow_ai	0	0	Twitter W	########	31046184	toddbatt	Todd Batt	########		56	22		Eastern Time (US & Canada)
8.52E+17	People ar	0	0	Facebook	########	17755770	LibertysFi	Jerri Lynn	########		125	117		Central Time (US & Canada)
8.52E+17	RT @NotF	33	0	Twitter fo	########	6.05E+08	Spookkyle	Killer King	########	âšï¸âœï¸â	176	232	Morioh	
8.52E+17	These are	0	0	dlvr.it	########	7.88E+08	BerkleyBe	Berkley B	########	Bringing t	1622	182	Doghouse	Eastern Time (US & Canada)
8.52E+17	So what h	0	0	Twitter fo	########	1.42E+09	FR3DWOR	FR3D WOI	########	mood pro	748	291	San Antonio, TX	

Figure 2: 2017 Tweets above

tweetID	tweetText	tweetRet	tweetFav	tweetSou	tweetCre	userID	userScree	userName	userCreat	userDesc	userFollo	userFrien	userLocati	polarity
8.52E+17	RT : Empir	1	0	Twitter fo	########	4.11E+08	guille_vel	Guille!	########	Las mejor	157	152	Guatemal	Neutral
8.52E+17	RT : I troll	54671	0	Twitter fo	########	1.15E+09	jackiee_m	less-ooh-	########	hey I foun	265	123	somewhe	Positive
8.52E+17	RT : I troll	54671	0	Twitter fo	########	1.64E+08	hawkguys	jo the ho	########	this was s	3549	104		Positive
8.52E+17	How the S	0	0	IFTTT	########	7.20E+17	DRL_USAN	Daily New	########		1653	1216		Neutral
8.52E+17	Hey . Bran	0	0	Twitter W	########	16227429	RipeInc	Len Roma	########	Owner of	8576	6955	Albuquer	Positive
8.52E+17	Which UN	0	0	Twitter fo	########	3.99E+09	Peacehav	Peacehav	########	the 'erber	162	499	City of Bri	Positive
8.52E+17	2 Floor Ti	0	0	IFTTT	########	33799339	Reeeemix	Reeeemix	########	Reeeemix	1762	4	USA	Negative
8.52E+17	RT : They	3	0	Twitter fo	########	4.15E+08	erianna16	Erianna	########		970	423		Negative
8.52E+17	RT : I troll	54671	0	Twitter fo	########	2.47E+09	iminesn	inÃªs	########		599	377		Positive
8.52E+17	United Air	0	0	WordPres	########	2.88E+09	theamedp	The Amec	########	Breaking	1189	0	DiyarbakÄ	Neutral
8.52E+17	RT : Can y	1	0	Twitter fo	########	1.13E+09	uhlektruh	emo emu	########	pnw gay.	263	299		Negative
8.52E+17	It's a prob	0	0	Twitter fo	########	2.27E+09	Laurab4re	L. M. Blair	########	I punch N	3685	3426		Negative
8.52E+17	RT : Unite	684	0	Twitter fo	########	2.52E+09	charliehin	Timothyð	########	Jeremiah	359	249		Positive
8.52E+17	Why must	0	0	Twitter fo	########	2.86E+09	rezigler	Richard Zi	########		24	83		Negative
8.52E+17	RT : Mono	27	0	Twitter fo	########	15753253	jeffsleasm	Jeff Sleas	########	Skinny kic	1159	2121	216	Neutral
8.52E+17	RT :	16	0	Twitter fo	########	3.71E+08	alltheway	MayItwee	########	Shakamak	342	249	Jasonville	Negative
8.52E+17	RT : I troll	54671	0	Twitter fo	########	2.84E+09	RICECAKE	jimin the	########	your local	1251	74	srâ,,¢ mal	Positive
8.52E+17	Why do ai	0	0	Twitter fo	########	1.74E+08	kwansfull	derek kwa	########	building p	415	175	los angele	Neutral
8.52E+17	RT : I troll	54671	0	Twitter fo	########	8.01E+08	wvngg	wvngg	########	living for l	596	573	Isla Vista,	Positive
8.52E+17	RT : Van H	2	0	Twitter fo	########	19262912	careerfed	CareerFec	########	Gun Owne	1618	2234	DC-SFO-S	Neutral
8.52E+17	Everyone	0	0	Twitter fo	########	2.43E+09	hawtcoco	habibi	########	it's a habi	145	167	Portland,	Positive
8.52E+17	Me critica	0	0	Facebook	########	7.7E+08	nrcoleptc	A Narcole	########		17	0		Negative
8.52E+17	United Air	0	0	SocialCha	########	2.27E+09	Videos_Fc	Videos Fo	########	Celebrity	176	414	United Sta	Positive
8.52E+17	RT : Unite	12	0	Twitter fo	########	3.03E+09	moirbad	Mohamed	########	Entrepren	306	83	Somalia	Positive
8.52E+17	@GibiAsm	0	0	Twitter fo	########	2.31E+09	Blank_Le_	Blank	########	Like most	39	84		Positive
8.52E+17	RT : So Pe	1612	0	Twitter fo	########	2.33E+08	DarkSamu	Patrick H.	########	Gamer/A	579	227	SR388	Neutral
8.52E+17	This is the	0	0	Hootsuite	########	15576134	teresajenl	Teresa Jer	########	"The very	2539	2313	Salt Lake (	Neutral
8.52E+17	This is the	0	0	Hootsuite	########	1.28E+08	WriteOnP	Write On!	########	Write On!	1127	1496	Salt Lake (	Neutral
8.52E+17	This	0	0	TweetDec	########	2.92E+09	bayside_j	BAYSIDE J	########	With a ho	592	290	Mumbai	Neutral
8.52E+17	RT : Good	172	0	Twitter fo	########	4.59E+08	MattEyre1	Matt Eyre	########	Handsom	242	380	Chaddert	Positive
8.52E+17	@wow_ai	0	0	Twitter W	########	31046184	toddbatt	Todd Batt	########		56	22		Negative
8.52E+17	People ar	0	0	Facebook	########	17755770	LibertysFi	Jerri Lynn	########		125	117		Negative
8.52E+17	RT : Unite	33	0	Twitter fo	########	6.05E+08	Spookkyle	Killer King	########	âšï¸âœï¸â	176	232	Morioh	Neutral
8.52E+17	These are	0	0	dlvr.it	########	7.88E+08	BerkleyBe	Berkley B	########	Bringing t	1622	182	Doghouse	Positive
8.52E+17	So what h	0	0	Twitter fo	########	1.42E+09	FR3DWOR	FR3D WOI	########	mood pro	748	291	San Anton	Neutral

Figure 3: Sentiment Analysis on 2017 Tweets above

(3) Neutral - values equal to 0

4 ANALYSIS

After performing the analysis on the data set we created a donut chart as shown in the **Figure 4** to find the major reason for negative reviews of airlines. We found that 31.71% of people are dissatisfied by the customer services offered by the airlines,18.14% of people are dissatisfied by the arrival & the departure delay in flight timings, 12.97% people have no reason but they are not satisfied. Other reasons are canceled flights, lost luggage, bad flight, flight booking problems, flight attendants problems.

From the bar graph as shown in the **Figure 5** we can analyze the number of positive, negative and neutral tweets in both the datasets. But we can't predict that which year airlines performed worst or which year performed the best because the number of tweets are different in both the year. Also, its very difficult to find whether the services of airlines were improved or did airlines put any efforts to improve their services. Therefore we need to work on some other analytics to analyze and find if users are more satisfied with the airlines services.

But with the pie chart as shown in the **Figure 6**, we can see that number of positive reviews outperforms the number of negative reviews from year 2015 to 2017. With this, we can conclude that airline services are paying attention to their passenger reviews and improving their services.

One interesting graph which shows the number of tweeters and the number of tweets tweeted by them can be seen in **Figure 7**.

5 METHOD

5.1 Topic Modeling/NNMF

NNMF stands for Non-negative matrix factorization. Non-negative Matrix Factorization (NMF) is a feature extraction algorithm. NMF is useful in scenarios where there are many attributes and the attributes are ambiguous or have weak predictability. On combining attributes, NMF can create meaningful patterns, topics, or themes.. NMF is based on linear algebra.

NMF is often useful in text mining. In a text document, same word can be used and occur at different places and in different context.

NMF decomposes multivariate data by creating a user-defined number of features. Each feature is a linear combination of the original attribute set; the coefficients of these linear combinations are non-negative.

NMF decomposes data matrix R in the dot product of two low rank matrices P and Q so that R is approximately equal to P times Q. NMF uses an iterative procedure of updating initial values of P and Q. The procedure terminates when the error converges or the specified number of iterations is reached.

Factored Matrices:

(1) P - it is called featured matrix in which rows represent feature and column for column in R
(2) Q - it is called weight matrix in which columns represent weights and row for row in R

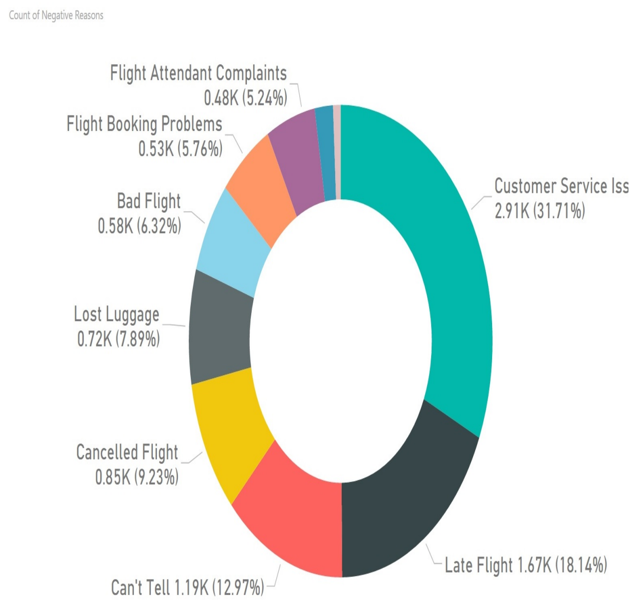

Figure 4: Top reasons for negative tweets

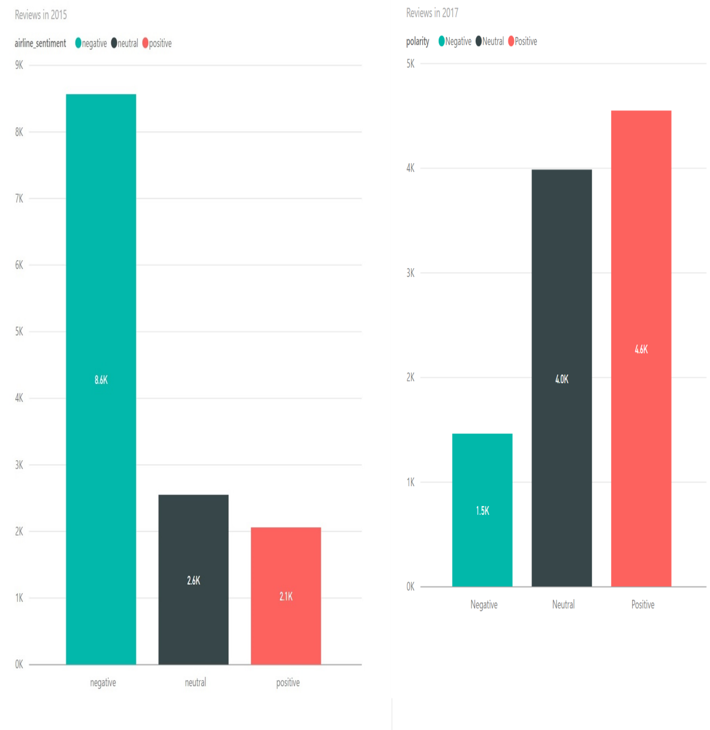

Figure 5: Comparison of number of positive, negative and neutral tweets in 2015 and 2017

The NMF is called non-negative matrix factorization because it always returns features and weights with no negative values. Henceforth, all features must be either positive or zero values to make it positive.

The NMF is highly related to both supervised and unsupervised clustering methodologies, but actually it is closely related to other clustering algorithms.

The 'Gradient Descent' is a technique used to factor which tries to minimize the errors in the modeling. Firstly, we calculate the squared error of out product. Then we calculate the gradient descent to figure out the direction to converge the error. We find the gradient by taking the differential of error for each element of the matrix. After calculating the gradient descent we try to update each element in P and Q matrix using learning rate α and then try to converge the error.

Algorithm 1 *Calculate*P&Q^T

```
Initialize:
P & Q with random small numbers
while max_steps do
    while row, col in R do
        if R[row][col] > 0 then
            compute error of element
            compute gradient from error
            update P & Q with new entry
        end if
        compute total error
        if error ¡ threshold then
            break
        end if
    end while
end while
return P, Q^T
```

On applying topic modeling on positive and negative tweet text, NNMF provides us with 2 different files. These files consists of top 10 positive and negative topics of the document. We can customize the number of top topics.

The NNMF is implemented in python language with the use of 'sklearn', 'nltk', and 're' libraries. The classes 'TF-IDF', 'NMF' are used from 'sklearn', while 'word_tokenize' and 'stopwords' are used from 'nltk' library.

Figure 6: Comparison of polarity in 2015 and 2017

To implement this we have first read each csv document and converted them to a vector or matrix after tokenizing each tweet and removing unwanted symbols and stop words from the tweets. These vectors are then transformed and converted to features and then passed to NMF to extract top topics of the document. The output of the NMF for top topics is: top 10 negative topics as shown in the **Figure 8** and top 10 positive topics as shown in the **Figure 9**.

6 REFERENCES

[1] Yun Wan, Dr. Qigang Gao. "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis".
[2]Cambria, Erik, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Statistical approaches to concept-level sentiment analysis." IEEE Intelligent Systems 3 (2013): 6-9.
[3] Adeborna, Esi, and Keng Siau. "An approach to sentiment analysisfi?!the case of airline quality rating." (2014).
. [4] Benjamin Bengfort. Beginners Guide to Non-Negative Matrix Factorization
[5] Albert Au Yeung. Matrix Factorization: A Simple Tutorial and Implementation in Python
[6]Read, Jonathon. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification." Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005
. [7] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus

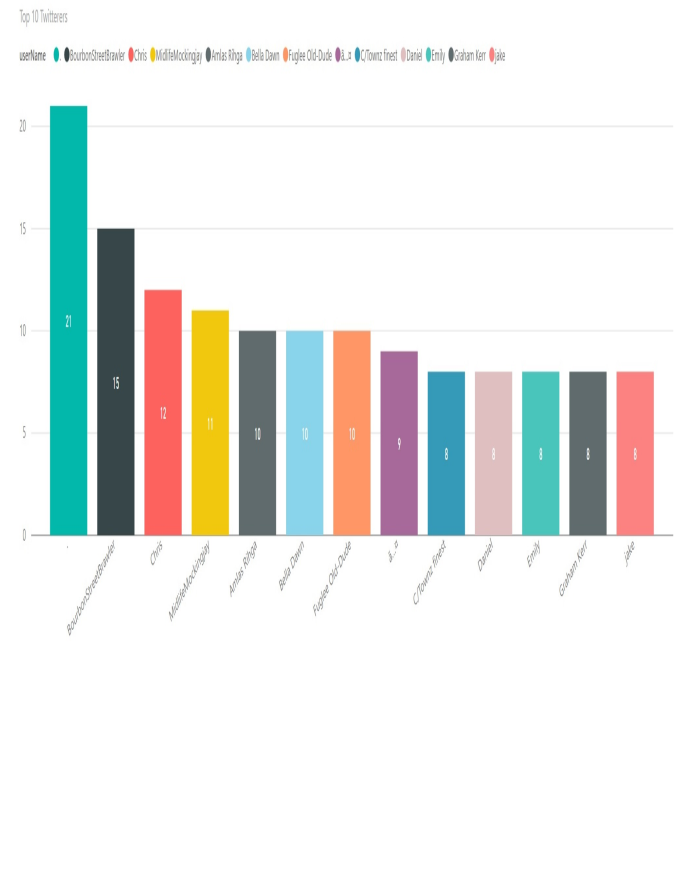

Figure 7: Top tweeters

forSentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.
. [8] Aneesha Bakharia "Topic Modeling with ScikitLearn"

Topic 0:
handcuffs threat first class forced https overbooked passenger plane united
Topic 1:
hate airlines find united pissed prove opportunity outraged right reasons
Topic 2:
kick likely least overbooked plane airlines time random passenger picked
Topic 3:
angry world whole easynews feedly united airlines buzz continues internet
Topic 4:
simple trick million turned doctor find turn found loses unitedairlinesmottos
Topic 5:
airlines united asian drag dragged flight passenger mocked plane video
Topic 6:
airline service customer tackle reputation promised poor united trend outrageous
Topic 7:
troubled ties attacks breaking past united airlines fine names hard
Topic 8:
ever newunitedairlinesmottos unite united never trend outrageous scandal http check
Topic 9:
pepsi spicer sean hold beer fuck marry kill take long

Figure 8: Top 10 negative topics

Topic 0:
trolled dead fucking response funny southwest airlines youtube first fine
Topic 1:
united airlines feature cool video ever seating best flights pleased
Topic 2:
wants plane leader city cops following video removing passengers theblaze
Topic 3:
launch campaign good time airlines useful protect next ways policy
Topic 4:
virgin take never airlines ever innocent visit trauma assaulted customer
Topic 5:
package deal know united airlines need better good david call
Topic 6:
kill happy youtube finding footage following followed flythefriendlyskies flying flyers
Topic 7:
spicer sean pepsi meme internet wrapped amazing united drinking picture
Topic 8:
dragged footage emerged forcefully flight united airlines overbooked says passenger
Topic 9:
forcibly passengers removing bill would prohibit remove pentagon awards assad

Figure 9: Top 10 positive topics