

# Artificial Intelligence Internship

## Project Report

Project ID: SPS\_PRO\_147



**PROJECT TITLE:**

**Sentiment Classification and Opinion Mining  
on Airline Reviews**

**Team Members:**

Nikhilesh Maremanda

Anvita Chebrolu

Adarsh Chandra Sekhar

G Rutwiz Gangadhar

# INDEX

1. Introduction
  - a. Overview
  - b. Purpose
2. Literature Survey
  - a. Existing Problem
  - b. Proposed Solution
3. Theoretical Analysis
  - a. Block Diagram
  - b. Hardware/Software designing
4. Experimental Investigations
5. Flowchart
6. Result
7. Advantages & Disadvantages
8. Applications
9. Conclusion
10. Future Scope
11. Bibliography
12. Appendix

# INTRODUCTION

## A. OVERVIEW

Every piece of information shared on Social Media carries an emotion, sentiment or feeling. These emotions can be positive, negative and neutral. All these emotions may come from a travel trip, restaurant trip, exhibitions, movies, elections, hospital visits etc. These emotions carry some hidden information related to comfort/discomfort in related areas. Hence, there is a good scope of analyzing this information to detect the patterns of the emotions. This analysis can help us to understand the emotions of the people in respective domain and the reasons behind it. Air travel is one of the most convenient modes for long distance travel at both national and international level. There are many airline service providers (ASPs) around the world. The competitive world motivates the airlines company to attract the customers. However, a traveller considers several points before selecting any airline. These points can be airfare, travel time, number of stoppages, number of baggage allowed, and existing customer feedback etc. Therefore, all ASPs are working in all these customer service areas to improve their facility and in-flight comfort in order to attract the customers. It is very important to understand the needs and comfort level of customers i.e. customer satisfaction during the flight. Therefore, customer feedback is very important for any airline industry. There could be several possible ways to collect the customer feedback. The most easiest and traditional way is the customer feedback form available during the journey. However, most of the passengers do not show any interest in filling feedback forms. The most convenient way for the passengers to share their opinions is the social media instead of feedback form. Social media provides a platform where a user can freely express his feedbacks on any issues they observed during flight. Twitter is one of the popular platforms worldwide. The information from Twitter can be utilized to develop a recommender system. In addition, travellers are more comfortable in sharing their views about travel experiences on Twitter. A variety of major issues affects the emotions of a passenger in air travel. These issues can be cabin crew behaviour, food quality, loss of baggage, seat comfort, flight delay, airfare etc. All these issues may give rise to both positive and negative emotions. Also, if there is a continuous trend of negative tweets for an airline, then it may put a negative impact to the economic growth of the airline company. Therefore, it is important to understand the issues that give rise to negative tweets so that the respective airline company can take appropriate action on time. Therefore, we required some tools and techniques that are able to handle such a large number of tweet database and can provide insights to help airline industry. Machine learning technologies made it possible to analyze huge database and to develop highly accurate prediction or classification models. The study opted to develop a classification model for three categories of sentiments i.e. positive, neutral and negative. Artificial Neural Networks were trained on the pre-processed tweets. Further, convolutional neural network (CNN) is trained on the data and its performance were compared with the best model among NLP and ANN models. A result shows that CNN outperformed all other models in terms

of accuracy and performance. Further, association rule mining is used to map the relationship between several issues related to passenger's comfort during flight with the nature of emotions (positive, neutral or negative).

---

## B. PURPOSE

Sentiment analysis is an important approach to extract emotions from any textual information i.e. online articles, product review, movie reviews, Twitter data etc. Twitter data is usually contains information about a person's opinion on any miscellaneous topic. Air travel is also one of these hot topics that are widely spread on Twitter. Air passengers usually share their travel experience on Twitter. This information can be useful if analyzed using machine learning techniques and can provide insights that helps to understand the comfort level of the passenger in the flight.

# LITERATURE SURVEY

---

## A. EXISTING PROBLEM:

Our day-to-day life has always been influenced by what people think. Ideas and opinions of others have always affected our own opinions. The explosion of Web 2.0 has led to increased activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. As a result there has been an eruption of interest in people to mine these vast resources of data for opinions. Sentiment Analysis or Opinion Mining is the computational treatment of opinions, sentiments and subjectivity of text. In this report, we take a look at the various challenges and applications of Sentiment Analysis.

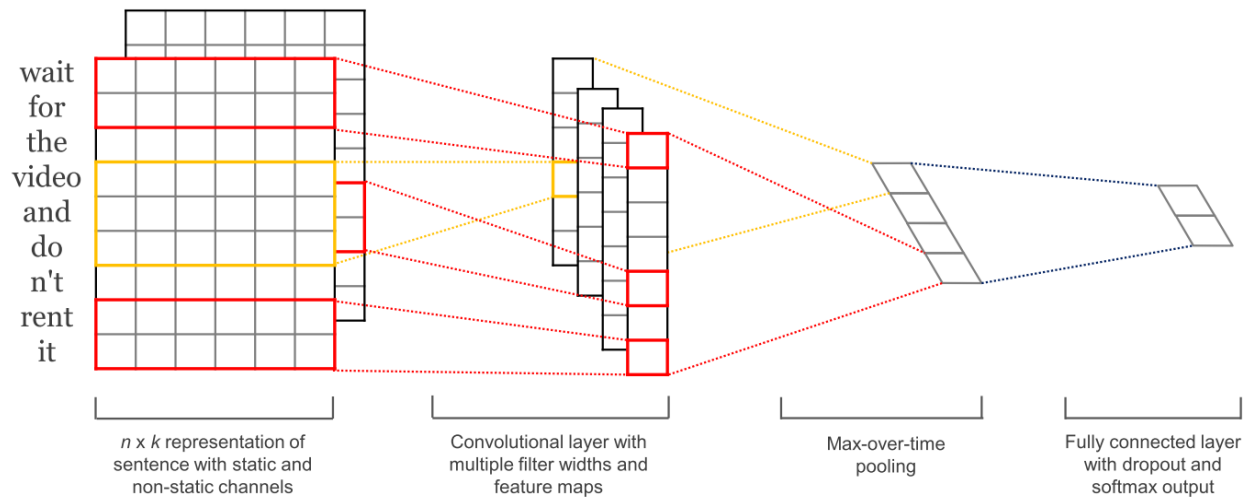
---

## B. PROPOSED SOLUTION:

Convolutional neural network (CNN) is trained on the data and its performance were compared with the best model among SVM (support vector machine) and ANN models. Results shows that CNN outperformed all other models in terms of accuracy and performance. Further, association rule mining is used to map the relationship between several issues related to passenger's comfort during flight with the nature of emotions (positive, neutral or negative).

# THEORETICAL ANALYSIS

## A. BLOCK DIAGRAM



The first layer is the input feature layer. For each sentence or paragraph, the rows in its feature matrix correspond to each of its word in the same order. More concretely, each row is a feature vector of the corresponding word. The second layer is convolution layer. We applied convolution on the input layer using multiple filters with different sizes. The idea behind this network is to use convolution layer to capture the local feature among words (similar to N-gram features), use two fully connected layers to learn high-level features appropriate for specific classification tasks, and use pooling layer to prevent over fitting and deal with different lengths of text data

## B. HARDWARE/SOFTWARE DESIGNING

✚ **Hardware requirements:** Laptop

✚ **Software requirements:**

- Twitter (Any Social Media platform)
- Python – 3.6
- Keras – 2.2.4
- Tensorflow – 1.14.0
- Spyder

# EXPERIMENTAL INVESTIGATIONS

We have used Twitter dataset from kaggle.

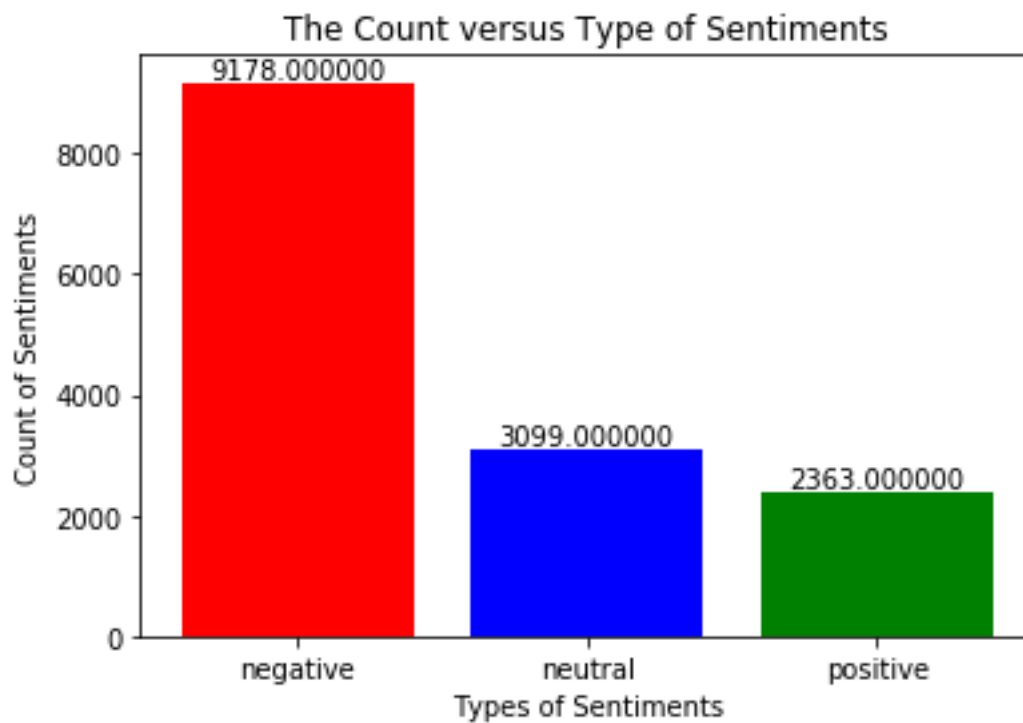
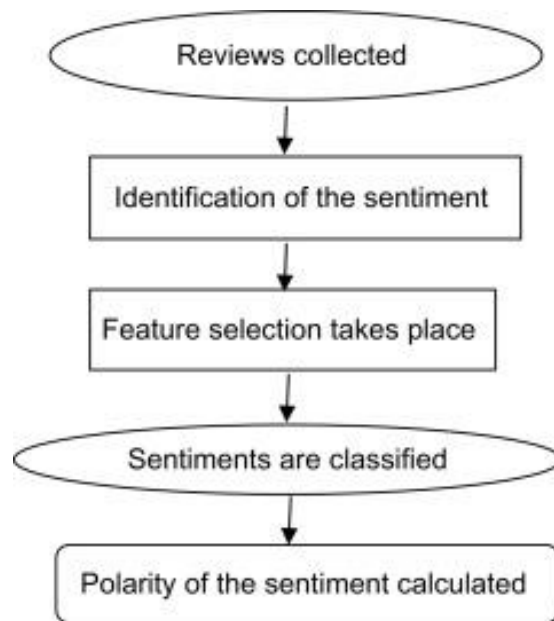
The dataset which we have worked on, has only two main columns which important for analyzing the data. One is airline-sentiment and other one is text, which is basically the tweets tweeted by several users. We have a dataset of 14,000 rows all together.

We have imported all the libraries and packages for model training. We started to train our model but firstly removing all the unnecessary attributes from the dataset. And also removal of unnecessary symbols from the twitter text from the text attribute.

tweet_id	airline_sentiment	negative	airline	name	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
5.7E+17	neutral	1	Virgin America	cairdin	0	@VirginAmerica What @dh	#####			Eastern Time (US & Canada)
5.7E+17	positive	0.3486	Virgin America	jnardino	0	@VirginAmerica plus you've	#####			Pacific Time (US & Canada)
5.7E+17	neutral	0.6837	Virgin America	yvonnalynn	0	@VirginAmerica I didn't tod	#####		Lets Play	Central Time (US & Canada)
5.7E+17	negative	1	Bad Flight	0.7033 Virgin America	jnardino	0	@VirginAmerica it's really a	#####		Pacific Time (US & Canada)
5.7E+17	negative	1	Can't Tell	1 Virgin America	jnardino	0	@VirginAmerica and it's a r	#####		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	Virgin America	cjmcginnis	0	@VirginAmerica yes, nearly	#####		San Francisco CA	Pacific Time (US & Canada)
5.7E+17	neutral	0.634	Virgin America	pilot	0	@VirginAmerica Really miss	#####		Los Angeles	Pacific Time (US & Canada)
5.7E+17	positive	0.6559	Virgin America	dhepburn	0	@virginamerica Well, I didn	#####		San Diego	Pacific Time (US & Canada)
5.7E+17	positive	1	Virgin America	YupitsTate	0	@VirginAmerica it was ama	#####		Los Angeles	Eastern Time (US & Canada)
5.7E+17	neutral	0.6769	Virgin America	idk_but_youtube	0	@VirginAmerica did you kn	#####		1/1 loner squad	Eastern Time (US & Canada)
5.7E+17	positive	1	Virgin America	HyperCamilax	0	@VirginAmerica I &3 pret	#####		NYC	America/New_York
5.7E+17	positive	1	Virgin America	HyperCamilax	0	@VirginAmerica This is such	#####		NYC	America/New_York
5.7E+17	positive	0.6451	Virgin America	mollanderson	0	@VirginAmerica @virginme	#####			Eastern Time (US & Canada)
5.7E+17	positive	1	Virgin America	sjespers	0	@VirginAmerica Thanks!	#####		San Francisco, CA	Pacific Time (US & Canada)
5.7E+17	negative	0.6842	Late Flight	0.3684 Virgin America	smartwatermelon	0	@VirginAmerica SFO-PDX sc	#####	palo alto, ca	Pacific Time (US & Canada)
5.7E+17	positive	1	Virgin America	ltzBrianHunty	0	@VirginAmerica So excited	#####		west covina	Pacific Time (US & Canada)
5.7E+17	negative	1	Bad Flight	1 Virgin America	heatherovieida	0	@VirginAmerica I flew from	#####	this place called NYC	Eastern Time (US & Canada)
5.7E+17	positive	1	Virgin America	thebrandiray	0	1 to6 flying @VirginAmeri	#####		Somewhere celebrat	Atlantic Time (Canada)
5.7E+17	positive	1	Virgin America	JNLPierce	0	@VirginAmerica you know \	#####		Boston   Waltham	Quito
5.7E+17	negative	0.6705	Can't Tell	0.3614 Virgin America	MISSGJ	0	@VirginAmerica why are yo	#####		
5.7E+17	positive	1	Virgin America	DT_Les	0	@VirginAr [40.74804263, -7;	#####			
5.7E+17	positive	1	Virgin America	ElvinaBeck	0	@VirginAmerica I love the h	#####		Los Angeles	Pacific Time (US & Canada)
5.7E+17	neutral	1	Virgin America	rlynch21086	0	@VirginAmerica will you be	#####		Boston, MA	Eastern Time (US & Canada)

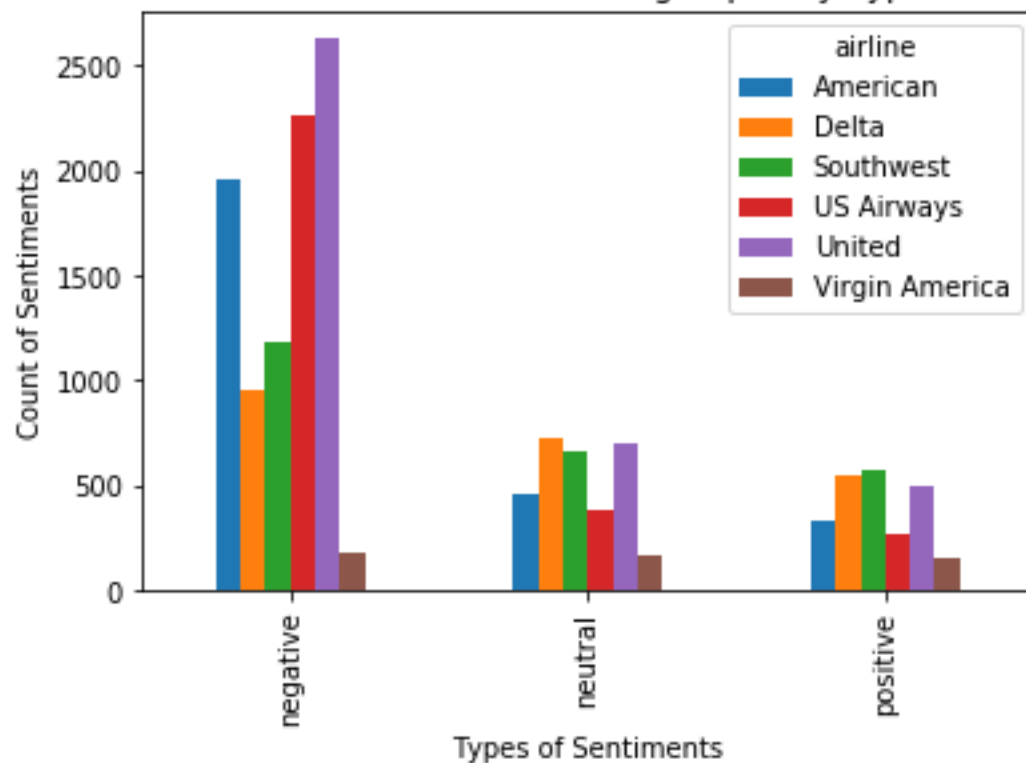
CNN performs pretty well in the sentiment classification task. This is because count vecotrizaton extracts semantic features of words and the convolution neural network can capture some local feature among adjacent while learning useful high-level features through training. Initially, each tweet is represented as a word vector in the n dimensional space. Assuming that d is the dimension of word vectors and l in the length of the tweet (the number of concatenated words in the tweet). Therefore, the dimension of the tweet matrix can be defined as  $l \times d$ . these features are passed to the next layer which acts as a fully connected neural network model. The output of the final hidden layer is applied to the ReLU (rectified linear unit) activation function, which categorizes the respective tweet vector into positive or negative class.

## CHARTS AND GRAPHS

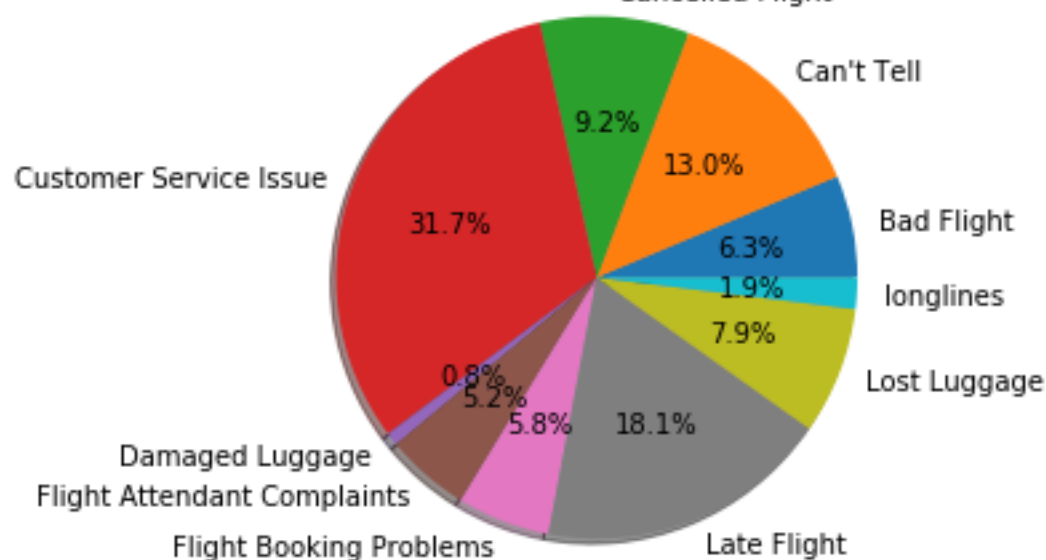




The Count of Sentiments versus Airlines grouped by Types of Sentiments

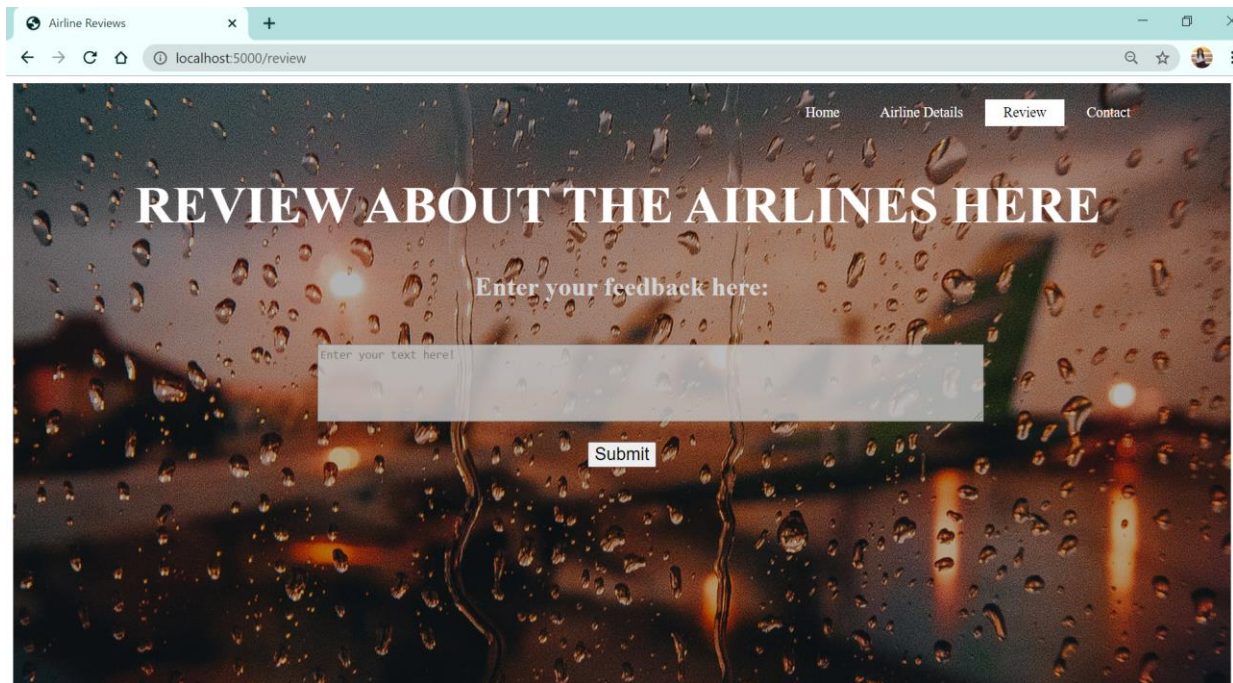
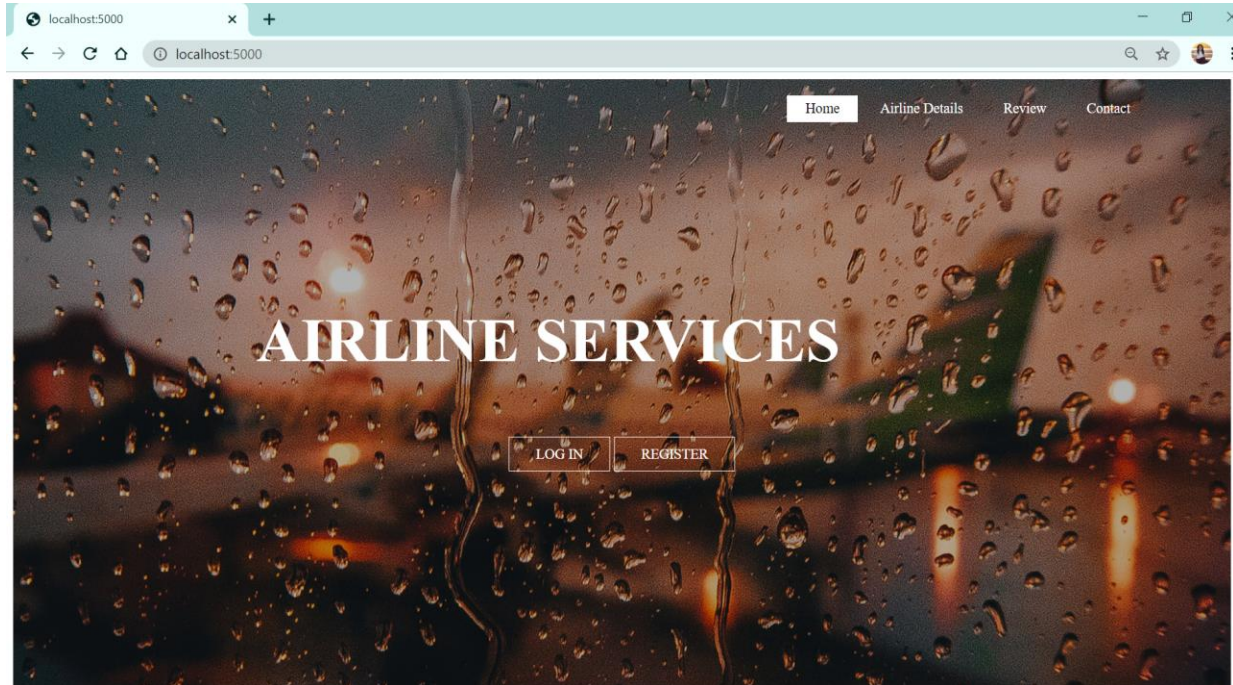


The distribution of Sentiments among All Negative Reasons

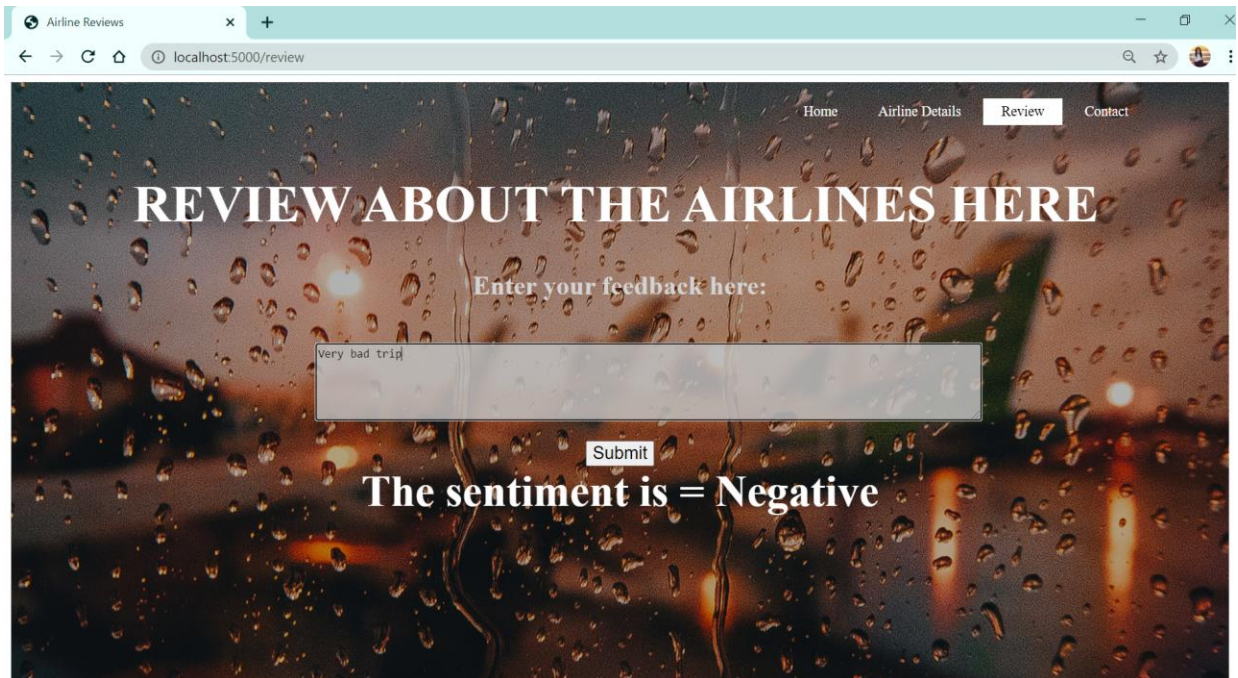
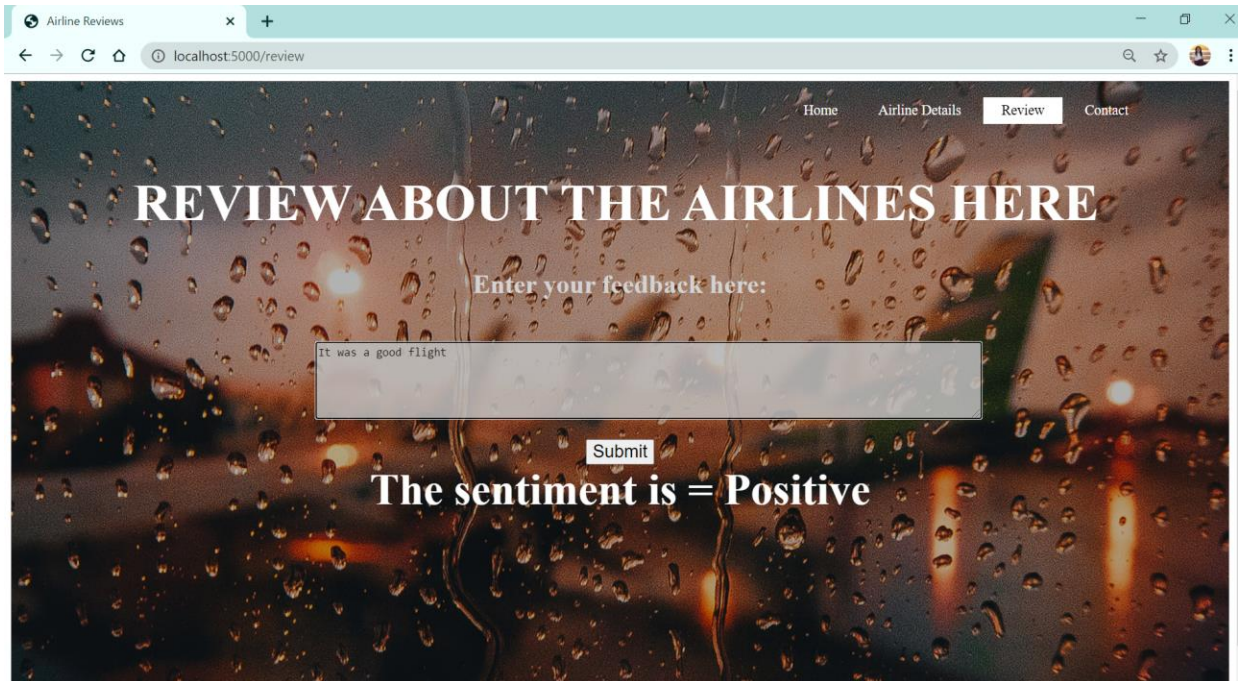


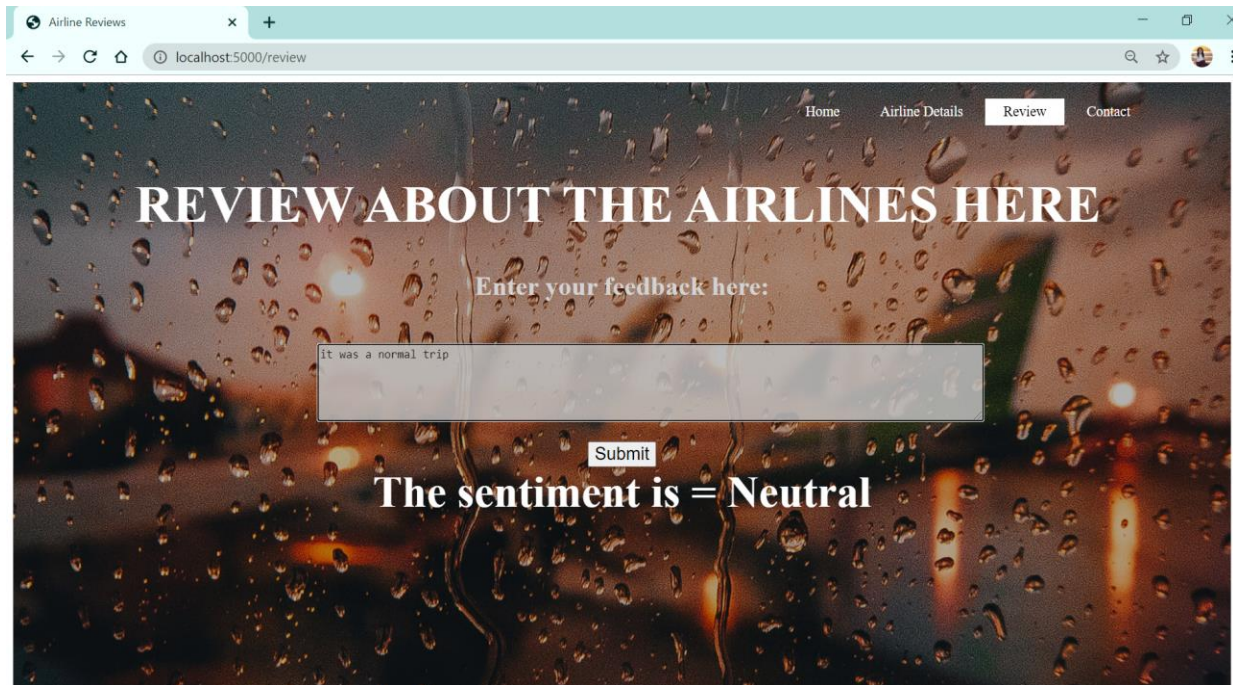
## RESULT

The following images show the screenshots of our application of Sentiment Classification and Opinion Mining on Airline Reviews –









## ADVANTAGES AND DISADVANTAGES

### ADVANTAGES:

- Dictionary is not necessary. Demonstrates the high accuracy of classification.
- Performance accuracy of 91% at the review level and 86% at the sentence level.
- Sentence level sentiment classification performs better than the word level.
- Labeled data and the procedure of learning is not required.

### DISADVANTAGES:

- Classifier trained on the texts in one domain in most cases does not work with other domains.
- Efficiency and accuracy depend the defining rules.
- Requires powerful linguistic resources which are not always available.

## APPLICATIONS

- Consumers can use sentiment analysis to research products or services before booking a flight.
- Using sentiment analysis Airlines companies attract the customers by knowing the requirements of the customers.
- To analyze customer satisfaction.
- Airlines companies can also use this to gather critical feedback about problems in their services.





## CONCLUSION

We explored four learning techniques and successfully applied them on our learning problem. NLP performs best and yields 99.1% accuracy in sentiment task. We also implemented this NLP model with our html page to get a more accurate information, which gives promising results and would be useful if we have a larger labeled dataset for training.



## FUTURE SCOPE

In the future, we consider to combine Recurrent Neural Networks (RNN) and CNN since RNN can 'remember' all previous information of a tweet and CNN can well capture the inter-word information.

## BIBLIOGRAPHY

-  <http://cs229.stanford.edu/proj2016/report/YuanZhongHuang-SentimentClassificationAndOpinionMiningonAirlineReviews-report.pdf>
-  <https://www.semanticscholar.org/paper/Sentiment-Classification-and-Opinion-Mining-on-Yuan/daf1d9de4066eed1d193847cae578389da16c5e8>
-  <https://www.mdpi.com/1099-4300/21/11/1078/htm>
-  [https://www.researchgate.net/publication/329093639\\_Sentiment\\_Analysis\\_for\\_Airlines\\_Services\\_Based\\_on\\_Twitter\\_Dataset](https://www.researchgate.net/publication/329093639_Sentiment_Analysis_for_Airlines_Services_Based_on_Twitter_Dataset)

## APPENDIX

-  **MODEL BUILDING:**
  - Dataset (link)
  - Spyder Notebook (link)
-  **Source Code:**
  - Flask-GitHub
  - HTML-GitHub