Sentiment Analysis of the tweets made by various US airlines' passengers

Sarthak Jain
36065118
School of Computing and Communications
Lancaster University
s.jain22@lancaster.ac.uk

Abstract

Sentiment analysis, sometimes also known as "opinion mining," is the process of understanding the author's opinion on a subject of discussion. It is one of the highly researched fields of Natural Language Processing (NLP), which is a result of the advancement in internet-based applications along with the introduction of a large number of tailored reviews that exist in different forms on several platforms available over the internet, including social media platforms, blogs, discord, and forums. The information in these reviews is helpful for both the consumers of the product or services as well as the provider of the product or services as it provides direct feedback from the consumer. Sentiment analysis helps in automating the process of understanding the emotions behind a particular review and getting aggregate feedback across millions of such reviews.

This paper discusses the various stages of a sentiment analysis task performed over several reviews posted on Twitter by passengers using various US airlines.

1 Introduction

Sentiment analysis is a variant of text classification where the text is classified on the basis of the polarity of the option contained in the text. It is a highly researched topic in Natural Language Processing, which mainly deals with human-computer interaction (Devika et al., 2016).
Sentiment analysis is majorly context-dependent. Generally, a sentiment analysis problem involves:

- an opinion or emotion. An opinion can have a positive, negative, or neutral polarity, while emotion can be either qualitative or quantitative.
- subject or subject of discussion. It clarifies the context of the discussion. At times, the opinion could be on multiple aspects of the same subject. For instance, the subject could be a product like a mobile phone, and the opinion could be about its battery life as well as its camera.
- opinion holder, usually the author of the text about the subject. The sentiment analysis exercise involves predicting how the opinion holder feels about the subject of discussion.

Sentiment analysis has various practical applications, including social media monitoring and brand monitoring. The rapid growth of social media has allowed internet users to express their opinions on various topics. These opinions help the brands to understand how their consumers interact with their products and also help the users to understand the pros and cons of the competitive products in the market. The same is true for airline travel as well.

Social media platforms like Twitter act as a platform to discuss various topics, and these discussions can be accessed faster than any other standard platform (Troussas et al., 2015). When it comes to air travel, multiple options are available to users, and it can be confusing to decide which airline service would best suit the needs of the user. Moreover, by knowing the opinion of the people about their airline services, the service providers can reflect upon the needs of their consumers. Performing sentiment analysis provides a solution for both parties. Hence, in this paper, sentiment analysis of the tweets made by users of various US-based airline travel services is performed. Various text preprocessing steps involved in generating the features from text-based data and extracting new features for performing predictions using various machine learning models are also discussed in detail. The text preprocessing techniques utilized in this paper include the creation of word clouds, tokenization of text data, using n-grams and stemming. Follow-

ing that, text features are converted into a numeric format using Bag-of-words (BOW) and TfIdf. New features are also generated by counting the number of tokens as well as using the valence scores based on polarity and subjectivity. Finally, after preprocessing the data, sentiment polarity prediction is performed on the mentioned dataset using various machine learning algorithms like Logistic regression, Naive Bayes, K-Nearest Neighbours, Multi-Layer Perceptron, and Deep Neural Networks, using 5-fold cross-validation,followed by the results being reported.

2 Related Work

In (Krouska et al., 2016), classification methods and various text preprocessing techniques for Twitter text are compared extensively, establishing that feature selection and repression can have an impact on the performance of sentiment polarity prediction.
In (Yamamoto et al., 2014), emoticons are assigned roles like 'emphasis', 'assuagement', 'conversion', and 'addition' to determine the sentiment of a tweet.
In (Fouzia Sayeedunnissa et al., 2013), a BOW-based approach for sentiment analysis is used for opinion mining. They used various classifiers that ended up giving marginally better results for classifying the positive class than the negative class. Further, the use of information gain and chi-square with Naive Bayes improves the accuracy.
In (Hutto and Gilbert, 2014), Vader, which is a rule-based model for general sentiment analysis, is proposed, where certain lexical features are specifically attuned to sentiment in microblogging sites. These features are then combined with five generic rules based on grammatical and syntactical conventions. The proposed model is then compared against various rule-based and machine learning-based state-of-the-art models.
In (Rill et al., 2014), a system for detecting political topics that come up on Twitter is described. Quick detection based on a few tweets at an early stage of a conversation is emphasized. Furthermore, to detect the polarity of hashtagged topics, an opinion mining component is added to the system. As a result, special Twitter hashtags known as "sentiment hashtags" are used by people to tag their thoughts on politicians or political parties. The idea was to build up relation graphs for emerging political topics enriched with information like context and polarity, which could be used to add a new dimension to an existing web ontology or semantic network.
In (Khan et al., 2014), the authors proposed a hybrid approach for sentiment analysis of Twitter data that leveraged text-preprocessing techniques like slang detection, lemmatization, and stop word removal. To deal with the problem of data sparsity, they used cross-domain techniques. The resulting model was finally compared against the state-of-the-art.

3 Data

The data was picked from the dataset repository of Kaggle, which is an online community for data scientists and machine learning practitioners. The dataset can be downloaded using this link.
The dataset originated from Crowdflower's Data for Everyone library and was made available on Kaggle with a few adjustments. According to the source, the data was web scraped in February 2015 from Twitter to perform sentiment analysis to solve the problems faced by each major US airline service provider. It contains 14640 data points formed out of 15 features consisting of tweets about 6 major US airline service providers.
The data was labelled by human experts, and the web scraped data was classified into neutral, negative, and positive tweets. As the data was web scraped directly from Twitter, it forms an authentic sample to work upon, despite being relatively old. Furthermore, the number of tweets was just enough to conduct the analysis.

4 Methodology

The sentiment analysis task is divided into 4 parts: Exploratory Data Analysis, Data Cleaning, Feature Extraction, and Classification.

4.1 Exploratory Data Analysis

Exploratory data analysis helps in developing an initial familiarity with the data. The following tasks are performed under Exploratory Data Analysis:

- Upon reading the data, the dimensions of the data, number of classes, and datatype of each of the features are observed before calculating the basic summary statistics about the data. The data is found to be imbalanced among the positive, negative, and neutral classes. This is shown in Fig.(1). Further, it is also observed

that the data points are imbalanced between the different airlines as well. The same can be visualized in the Fig.(4)

- Depending on the language of the text in tweets, the steps involved in preprocessing can vary. Hence, the language of the texts is verified and confirmed to be English.

- Next, the tweets are transformed to lower case to avoid issues in the steps where the text is converted into numeric features.

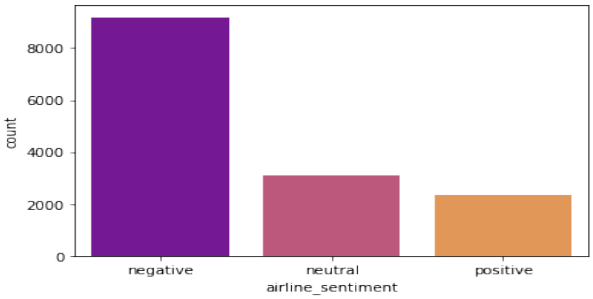


Figure 1: Distribution of data based on the polarity of the sentiments

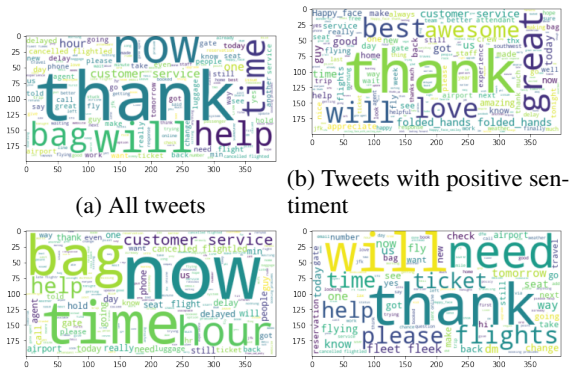
4.2 Data Cleaning

As a next step, the data set is cleaned of missing data and redundant features. Moreover, the data types of the features are changed to ensure they are processed in the intended manner.

Tweets consist of hashtags, mentions, URLs, as well as emojis and emoticons. Creating features out of these directly is difficult. Hence, they are replaced by annotations. Out of these, various emojis and emoticons express different emotions. Thus, instead of using a single annotated text for all of them, a different annotated text is used. Hashtags, URLs, and mentions are identified using regular expressions. However, unlike emoticons, they are replaced with separate annotated texts, respectively. Aside from the tweets, the remaining features were label-encoded before being converted to numeric features. One-hot encoding can be used for categorical features like airline names, but it was avoided to reduce computational complexity and keep the number of features used to a minimum.

4.3 Feature Extraction

More features are extracted to improve the performance of the ML models. The length of a tweet, for example, is used as a feature. Longer tweets



(a) All tweets

(b) Tweets with positive sentiment

(c) Tweets with negative sentiment

(d) Tweets with neutral sentiment

Figure 2: Visualizing tokens of interest using word clouds

are found to be more emotionally charged in general. As a result, making a feature out of it could be beneficial. Similarly, the number of punctuation marks in a tweet can be used to generate a feature. However, because tweets are not written like normal scripts and can include smileys and special characters, processing punctuation as a feature can be difficult. As a result, they are ignored in our application. Lexicon-based sentiment analysis techniques use valence scores. As a result, valence scores may be important features in the ML approach.

A combination of bag-of-words and TfIdf techniques are used to convert text to numeric features. The BOW creates features from the frequency with which certain words appear in tweets. Certain words may be useful in classifying tweets based on their sentiments. Visualization of such words is done using word clouds, which are shown in Fig.(2). BOW exploits this by keeping track of them in a document. BOW, on the other hand, does not consider the length of the tweet. The TfIdf approach gives a higher score to words that are frequently used in a tweet, while words that are frequently used throughout the corpus receive a lower score. This also aids in the reduction or elimination of stop words in the calculation of TfIdf for tweets. It should be noted that when forming features from BOW and TfIdf, unigrams, bigrams, and trigrams are used to account for the language-based context and to make the analysis more granular. Further n-grams are not considered to avoid overfitting. Also, before creating BOW and TfIdf features, the tweets are word-tokenized to perform stemming and lemmatization. Thus,

different forms of the same word are transformed into the same root as they usually contribute the same Usually, with lemmatization, marginally better results are obtained, but it takes more time to lemmatize the tokens than to stem them.

A total of 507 features are then used to classify the text into different sentiments. The data is then standardised and normalised, and the outliers are removed to simplify the training process. To reduce the number of features, Principal Component Analysis is applied. Hence, a total of 178 features explained 80% of the variance. The performance of various ML algorithms, including logistic regression, naive Bayes, knn, mlp, SVM, and DNN, is compared for the classification task. The classifiers using different permutations of hyperparameters are also subjected to 5-fold cross-validation. Metrics including accuracy, precision, recall, and f1 score are used to measure the performance of the models.

5 Results and Findings

It is observed that the sentiments are satisfactorily predicted by all the classifiers. Performance of the models across various metrics is shown in Fig.(3) for the DNN. The performance of all the other classifiers is shown in Fig.(5), which is included in the appendix section. Out of all the models, Logistic Regression, SVMs, MLPs and Deep Neural Networks give relatively better results.

Logistic regression with only L2 regularisation outperforms logistic regression with both L1 and L2 regularisation by a small margin. In comparison to the RBF kernel, the linear kernel gives slightly better results for SVMs. When using the sigmoid activation function with the Adam optimizer and the tanh activation function with stochastic gradient descent for MLPs, the best results are obtained. In light of the models' simplicity, Logistic Regression appears to be the best fit for this application. Fitting the model, on the other hand, necessitates a large number of iterations and thus takes longer. In this application, SVMs outperform Logistic regression in terms of fitting time.

Fitting the logistic regression model by removing non-text based features and valence score based features, BOW and TfIdf, respectively, from the total number of features, is used to perform feature impact analysis. All of the features considered contributed positively to the models' performance, with TfIdf being the most effective feature, followed by BOW and basic features, in that order.

Despite changing the number of features generated by BOW and TfIdf approaches, the models tend to overfit. Further analysis is needed to reduce overfitting in the models to improve performance.

The effects of scaling the data by using standardisation and normalisation techniques are also studied, and it is observed that scaling improves the performance of the models marginally. However, the training time of the models is significantly reduced. Overall, sentiment prediction was performed successfully.

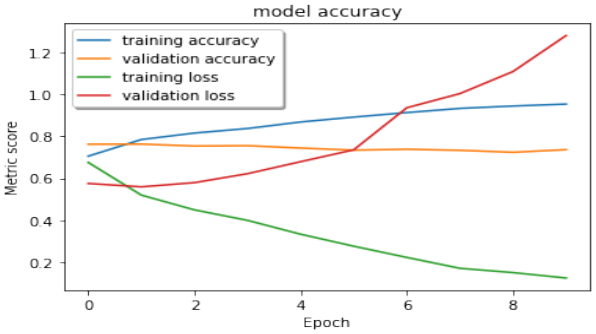


Figure 3: Performance of the DNN in predicting the sentiments in training and validation phase.

6 Conclusion and Future Work

The sentiment analysis of tweets from users of major US airlines was performed in this paper. The passengers voiced their opinions on various aspects of the airline's services. After the tweets were word-tokenized, text preprocessing techniques like BOW and Tfidf were used. The language-based contexts were also captured using the N-grams of the word tokens. URLs, hashtags, and mentions were annotated so that they could be processed properly. Valence scores and tweet-length were also used to create features. To visualise words of interest, word clouds were created. Finally, various machine learning algorithms were used to accurately predict the sentiment of tweets with an accuracy of around 80Further research can be done in the future to reduce overfitting in the predictors and improve performance during the validation phase. For opinion prediction, more features based on punctuation in a tweet could be used. Lexicon and rule-based sentiment analysis techniques could be investigated further and compared to ML-based approaches, as well as the possibility of developing a hybrid approach.

References

M.D. Devika, C. Sunitha, and Amal Ganesh. 2016. Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science*, 87:44–49. Fourth International Conference on Recent Trends in Computer Science Engineering (ICRTCSE 2016).

S Fouzia Sayeedunnissa, Adnan Rashid Hussain, and Mohd Abdul Hameed. 2013. Supervised opinion mining of social network data using a bag-of-words approach on the cloud. In *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, pages 299–309. Springer.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Farhan Hassan Khan, Saba Bashir, and Usman Qamar. 2014. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision support systems*, 57:245–257.

Akrivi Krouska, Christos Troussas, and Maria Virvou. 2016. The effect of preprocessing techniques on twitter sentiment analysis. In *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*, pages 1–5.

Sven Rill, Dirk Reinel, Jörg Scheidt, and Roberto V Zicari. 2014. Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69:24–33.

Christos Troussas, Maria Virvou, and Kurt Junshean Espinosa. 2015. Using visualization algorithms for discovering patterns in groups of users for tutoring multiple languages through social networking. *J. Networks*, 10(12):668–674.

Yuki Yamamoto, Tadahiko Kumamoto, and Akiyo Nadamoto. 2014. Role of emoticons for multidimensional sentiment analysis of twitter. In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services*, pages 107–115.

Appendix A Figures

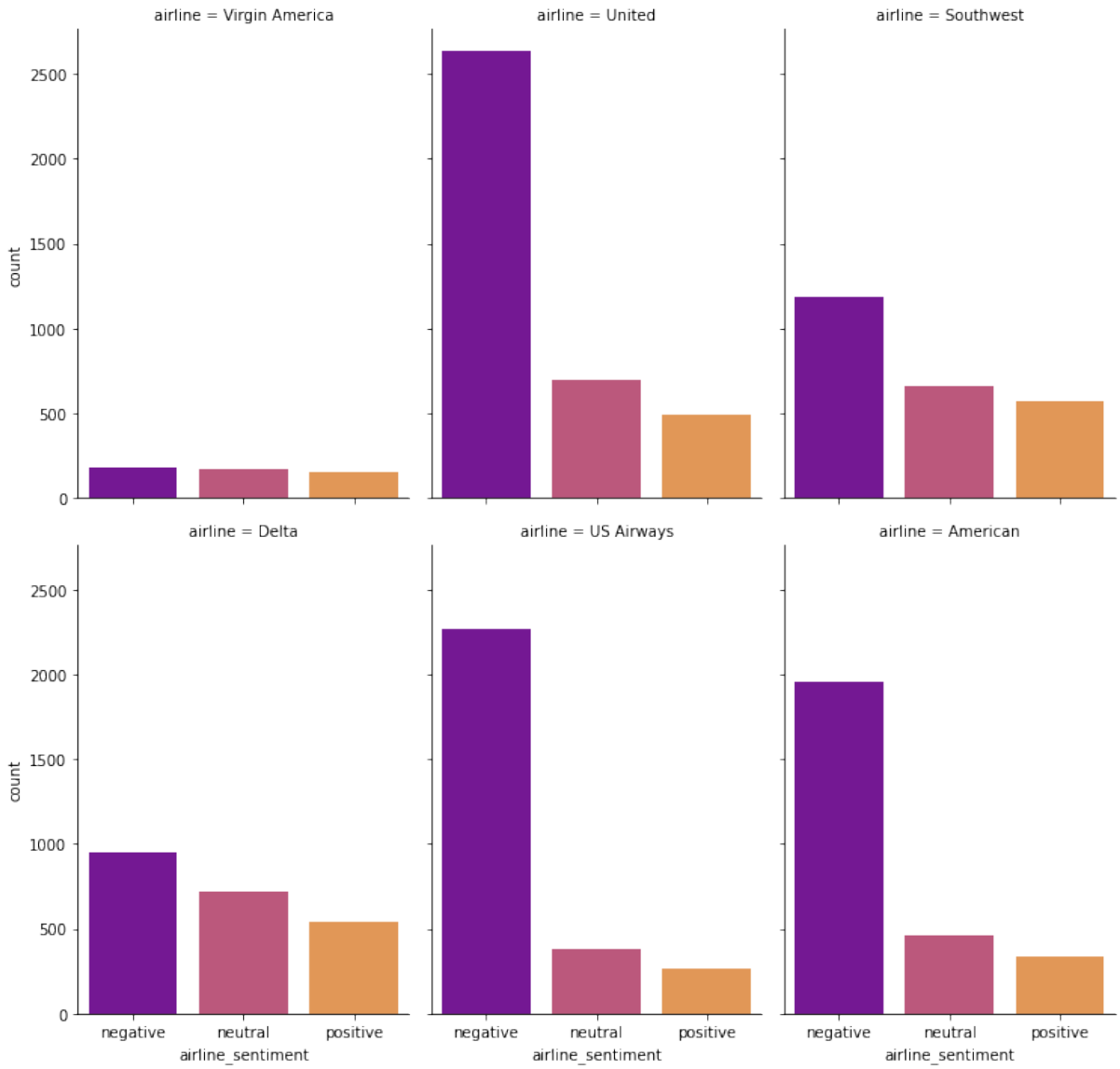


Figure 4: Distribution of data based on the polarity of the sentiments among the different airlines(top-left to bottom-right: Virgin America, United, Southwest, Delta, US Airways, and American)

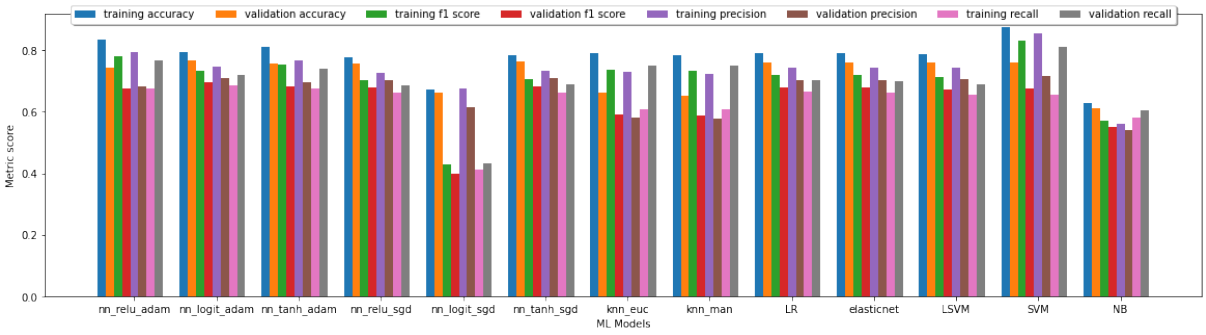


Figure 5: Performance of all the models in sentiment prediction across various metrics