EasyChair Preprint

№ 5228

Predict Sentiment of Airline Tweets Using ML Models

Rana Alqahtani

March 29, 2021

Predict Sentiment of Airline Tweets Using ML Models

Rana Al-Qahtani
Dublin City University
Princess Nourah bint Abdulrahman University
Riyadh
Ranaaqa@gmail.com

Many people like to share their opinion online about everything such as, social events, some products, or services of any industry. Sentiment analysis can be used to understand users' attitude or sentiment through resources with opinion-rich data such as Twitter. This research aims to provide analysis of the content tweets text of users' emotions in US airline company services and investigate the use of *two Machin Learning (ML), and four Deep Learning (DL) methods to prediction of sentiment from US airline tweets.*

Keywords— Sentiment analysis, BERT, CNN, Naïve Bayes, NLP, CNN, Logistic Regression, machinelearning, deep learnning, XLNET, Transform model.

ABBREVIATIONS

The important abbreviation in this paper are, NLP (Natural Language Processing), LR (Logistic Regression), NB (Naïve Bayes), LL (Log Loss), BERT (Pre-training of Deep Bidirectional Transformers), XLNET (Generalized Autoregressive Pre-training), EDA (Exploratory Data Analysis), CNN (Convolutional Neural Network), BOW (Bag of Word), B (Bigram),T (Trigram), TF-IDTF (Term Frequency Inverse Document Frequency), NLTK (Natural Language Toolkit), W2vec (Word2vector), LIME (Locally Interpretable Models and Effects), TP (True Positive), TN (True Negative), FP (False Positive), FN (False), LR (Learning Rate).

I. INTRODUCTION

Regarding the airline industry, sentiment analysis has the potential to help customers decide which airline is best by analyzing other customer opinions from online comments posted on review or micro-blogging sites. Sentiment analysis has been applied in this way across a variety of different domains, such as entertainment, education, automobiles, etc.

The paper is concentrated on the automatic sentiment analysis of US airline customer opinion using Twitter data. Twitter data allows researchers to obtain a wealth of knowledge from their users. Twitter is the most popular social networking website and application, and it is a real-time micro-blogging where news breaks first. Users interact by posting text messages in the form of tweets, which are limited to 140-character length. It is useful to classify tweets in way to improve information retrieval and to enable better decision making. Text mining is one of the featured fields of data mining, which has the potential to extract useful information from raw textual data. Applying sentiment analysis to Twitter is an emerging trend in text mining, with researchers recognizing the specific challenges it brings and its potential applications.

The dissertation consists of six sections, as follows, discusses previous research, which related to this topic of this research. It reviews previous airline Twitter applications using sentiment classification. It provides an overview of previous approaches in the field of data mining. Sentiment classification is also reviewed in this section, along with commonly applied measurements of the performance of such methods, **section III** introduces the analysis of the experiment for this project, clarifying the data, looking at the more comprehensive vision, and studies the potential future relationship, **Section IV** outlines the implementation of the proposed method based on the experiment in section III. It then presents the challenges identified during the process; **Section V** shows the result acquired from the implementation of the experiment in section IV. The detailed evaluations of the outcomes are included in this section, reflecting the objectives of this project, **section VI** summarizes the value of the experiment and suggests future work and the object of future research.

II. RELATED WORK

Previous research has used supervised Machine Learning algorithms and lexicon-based methods to classify tweets for different airline companies. Most researchers have investigated the process of sentiment analysis by detecting emotions found in the text of tweets.

Dutta Das et al.,2017, used 200 tweets directed at Emirates and Jet Airways and analyzed airline Twitter data using the Naïve Bayes algorithm for sentiment analysis. They used R and Rapid Miner tools to improve the classification model and map the tweets into the positive, negative, and neutral categories. They mentioned that the outcomes achieved using Naïve Bayes classifiers were promising for a more significant number of tweets in the analysis.

Hakh et al., 2017, applied the SMOTE method to solve the imbalanced challenge of the datasets and analyze a collection of tweets about six airline companies found in the US using machine learning techniques. They found that the feature selection and over-sampling techniques are equally essential to achieve refined results. Then they applied the sentiment classification (i.e., AdaBoost, Decision Tree, Linear SVM, Naïve Bayes, Random Forest, K- NN, and Kernel SVM).

Rane& Kumar., 2018 used their approach with pre-processing techniques to clean the tweets. These tweets were represented as vectors using a deep learning concept (Doc2vec) to do a phrase-level analysis that considers the word order. They then conducted a comparative study on six US airline companies using a Decision Tree, Random Forest, Gaussian Naïve Bayes, SVM, K-Nearest Neighbors, Logistic Regression, and AdaBoost. 80% of the data was trained by the classifiers, and residual data was used for testing. They

classified the tweets into three categories of sentiments. They mentioned that Logistic Regression, AdaBoost, Random Forest, and SVM performed well with an accuracy of more than 80%. However, they found that the AdaBoost approach is a more robust classifier than the others, according to their results.

Adarsh & Ravikumar., 2018 used tweets relating to Indigo Airlines, Emirates Airlines, and Qatar Airlines from their Twitter and customers who tweeted about these airlines. The approach of detecting sentiments on Twitter was proposed by considering the tweets from three popular Airlines. The definition of positive, negative, neutral sentiments was based on the score computation, which was the difference between the positive and negative words for each tweet. They found that Emirates Airlines had more positive sentiments compared to the other two; Indigo Airlines had more negative sentiments, and Qatar Airlines had more neutral sentiment tweets. The problem of this approach is that the presence of positive and negative words may not give relevant results in the case of sarcastic tweets as the placing of positive and negative words in a sentence gives different conclusions.

Prabhakar et al.,2019 determined their project research to focus on the top ten US airlines, which are America Airlines, Alaska Airlines, Delta Airlines, JetBlue Airlines, Hawaiian Airlines, SkyWest Airlines, Southwest Airlines, United Airlines, Spirit Airlines, and US Airways. The proposed methodology introduces a new, improved AdaBoost approach for sentiment analysis. Various Machine Learning algorithms were employed for identifying the appropriate algorithm for the system. Performance analysis was performed based on the confusion matrix and the accuracy of the algorithms.

Kumar & Zymbler., 2019 used the Glove dictionary and n-gram approach for Word Embedding, then classified the tweets using SVM (Support Vector Machine), and several ANN (Artificial Neural Network) architectures. With developed the classification mode Convolutional Neural Network (CNN) when compared with the most accurate model between SVM and several ANN architectures. They identified interesting associations that helped the airline industries to promote their customers' experience.

Rustam et al., 2019 proposed a voting classifier based on logistic regression and stochastic gradient descent classifier. Soft voting was used to combine the probability of LR and SGDC. Besides, various machine learning-based text classification methods were investigated to perform sentiment analysis. Three feature extraction methods (TF, TF-IDF, and Word2Vec) were investigated to analyze the impact on models' classification accuracy. They found that the feature extraction method TF-IDF is more appropriate for tweet classification. The proposed voting classifier performs better with both feature extraction methods and achieves an accuracy of 78.9 % and 79.1 % with TF and TF-IDF, respectively. Ensemble classifiers show higher accuracy than the non-ensemble classifiers, and the find LSTM does not perform well on the selected dataset.

Alghalibi et al., 2019 used dimensionality reduction-based data mining approaches that they proposed to reduce the data dimensionality in addition to the supervised learning classification approached, such as a Backpropagation Neural Network. The widest algorithms of the dimensionality reduction technique are the Principal Component Analysis (PCA) and Singular-Value Decomposition (SVD). They proposed a system using Backpropagation Neural Network (BPNN) to show that their dimensionality reduction approach has reduced the original data dimension from 2976 to 532 and achieved the highest accuracy. They compared results with other dimensionality reduction algorithms (PCA and SVD) that used the same dimension; they noticed the highest accuracy. Their approach satisfied 94.88% in the testing dataset, while the nearest one was 90.34%, which was achieved by the PCA.

Tiwari & Singh., 2019 the Extra Tree classifier was used, which outperforms all other techniques that were previously applied. The others used three classical algorithms (Naive Bayes, K-neighbor, and Decision Tree), and one is a boosting concept, as AdaBoost. Then there was a comparison of the five algorithms that provide more accuracy than classical algorithms (Decision Tree 63%, K-neighbors 67%, Naïve Bayes 69%, and AdaBoost 74%). Moreover, the Extra Tree Ensemble method outperforms all the others and provides 76% accuracy. This is much greater than the previous classification algorithms accuracy.

Khan & Urolagin., 2018 collected 10,000 tweets for 18 airlines based in four selected regions, which are America, India, Europe, and Australia. To predict consumer loyalty, they used three classifiers, namely, Random Forest, Decision Tree, and Logistic Regression. The model fit used tweet related information such as positive sentiment score, negative sentiment score, mean of retweets, mean of likes, and a number of followers. The two-class prediction performed as either loyal or not loyal. Maximum accuracy of 99.05% was observed for Random Forest on 10-fold cross-validation.

Rathod & Deshmukh.,2016 proposed exclusive sentiment polarity detection approaches. They concluded that using features for training classifier gives the maximum result as compared to without features. Furthermore, SVM gives an excellent result as compared to MaxEnt.

Yuan et al .,2016 explored four learning techniques: Naïve Bayes, Support Vector Machines (SVM), lexicon-based methods, and Convolutional Neural Networks (CNN), to predict the sentiment of tweets (positive, neutral or negative). They applied the same techniques to classify the negative reasons for bad feedback to identify if it was due to a late flight, canceled flight, flight booking problems, or customer service issues. They used N-gram and work2vec as features input to their algorithms. They found that SVM delivers the best accuracy of 79.6%in sentiment task and 64.8% accuracy in negative reason task. They noted that CNN with word2vec features gives promising outcomes and would be useful if they have a larger labeled dataset for training.

Hemakala & Santhoshkumar., 2018 worked on tweets for six major Indian Airlines and started with pre-processing techniques to clean the tweets and then represented these tweets as vectors using deep learning to do a phrase-level analysis. They used seven different classification strategies: Decision Tree, Random Forest, SVM, K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bayes, and AdaBoost. The latter, AdaBoost, gave an accuracy of 84.5%. The accuracies attained by the classifiers were sufficiently high to be used in the airline industry to perform customer satisfaction research.

III. METHODOLOGICAL APPROACH

This section explains the techniques from our proposed methods used in the experiments. focuses on two popular

machine learning classifiers: Naïve Bayes and Logistic Regression. deep learning is introduced, with detail on various models: CNN (Convolution Neural Network), BERT, XLNET, and ALBERT used for US airline tweets.

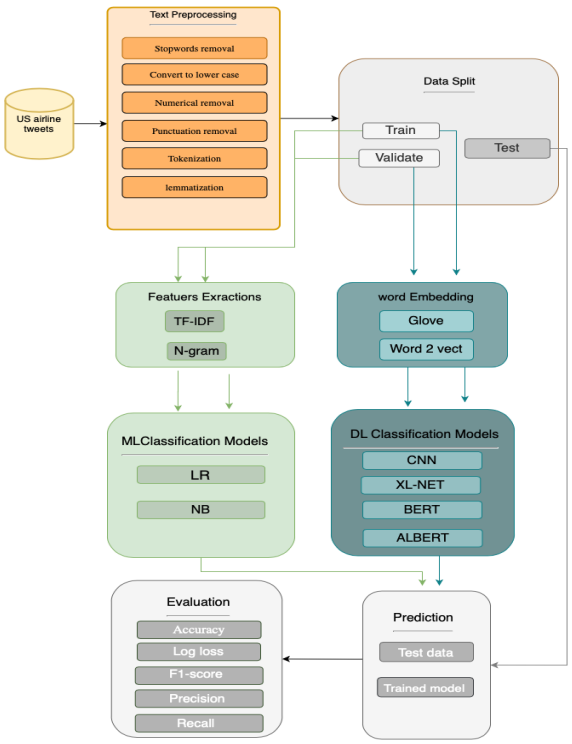


Fig1: The complete outline of our framework.

IV. The Data and Analysis

Data Collection

In this work, the dataset obtained from Kaggle as a CSV file is called "Twitter US Airline Sentiment," which contains different tweets released by CrowdFlower with a total number of 14,640 tweets, and 15 columns [28]. The tweets collected from Twitter, in February 2015, were for six major US Airlines that are: United, US Airways, Southwest, Delta, and Virgin America. The tweets were a mix of positive, negative, and neutral sentiments, and citing the reason for a negative classification as well as a confidence score for the assigned label. The included features are tweet id, sentiment, sentiment confidence score, negative reason, negative reason confidence, airline, sentiment gold, name, retweet count, tweet text, tweet coordinates, time of tweet, date of tweet, tweet location, and user time zone [28].

Exploratory Data Analysis (EDA)

A. Sentiments Distribution

The sentiment analysis in the table was examined by the counts for each sentiment label, it has been observed that there are more negative sentiments than other sentiments. Therefore, the people are more likely to write on Twitter if something goes wrong with their flight, rather than when nothing unexpected happens.

Table 1: The number of tweets in each sentiment

Sentiment	NEGATIVE	NUTRALL	POSETIVE
Count	9178	3099	2363

The distribution of sentiments of the overall tweets was represented by a pie chart diagram (fig1).

Table 2: The rate of Negative, Natural and Positive sentiments per airline

Airline	Negative	Neutral	Positive	Negative rate	Neutral rate	Positive rate
American	1960	463	336	0.710	0.168	0.122
Delta	955	723	544	0.430	0.325	0.245
Southwest	1186	664	570	0.490	0.274	0.235
US Airways	2263	381	269	0.777	0.131	0.092
United	2633	697	492	0.689	0.182	0.129
Virgin America	181	171	152	0.359	0.339	0.301

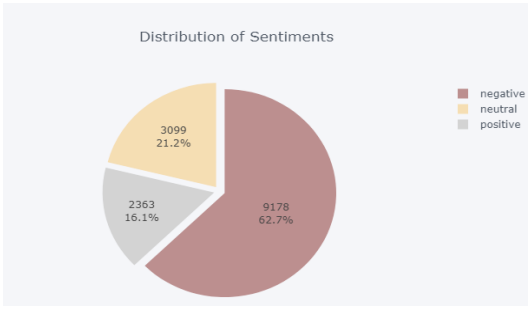


Fig1: The count sentiments numbers as percentages.

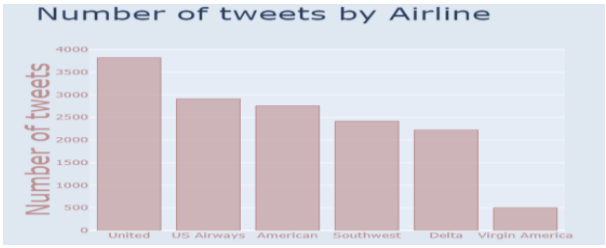


Fig 3: The number of tweets per airline company.

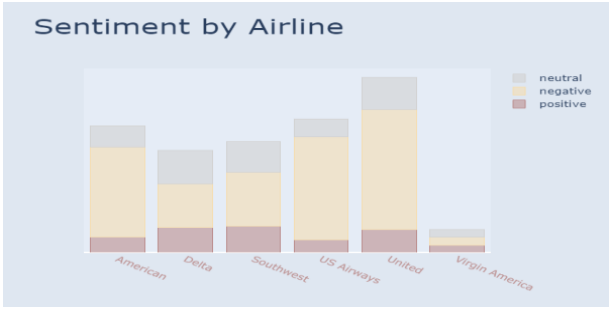


Fig 4 The number of sentiments in each airline.

Additionally, the United had a huge number of tweets about 3900, while US Airways 3000 tweets and American got 2900 tweets. In contrast, Southwest and Delta had a less number around 2000 tweets, though Virgin America had the lowest number of tweets. (fig1). The bar chart had presented the number for each class of sentiments in all six airline companies (fig3). All airlines received a higher number of negative class except Virgin America, which got a similar number compared to other classes.

B. Negative Reasons Distribution

In terms of Negative Reasons, the highest percentage of tweets are about "customer service issues" (around 35%),

which are the most frequent for customer complaints about the airline. (fig 4)

The heat map was used to visualize the percentage of negative reasons in each airline. It illustrates that "service issue "is the most common negative reason across all airlines. (fig 6).

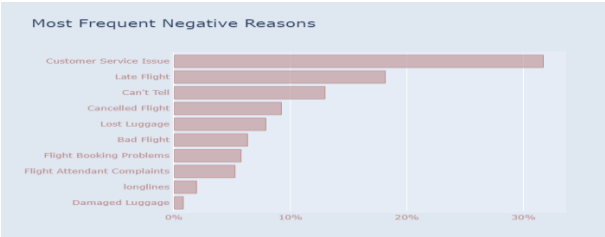


Fig 5: The percentage of repetition of each negative reason in the tweets.

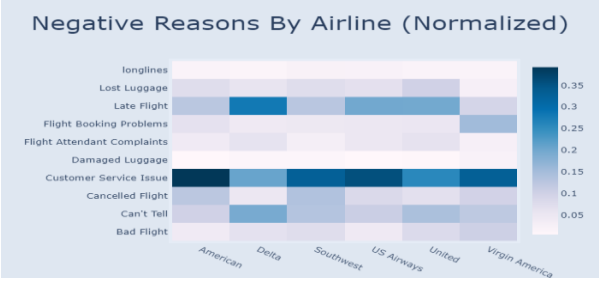


Fig 6: The percentage of each negative reasons by airline.

C. Tweet Length Distribution

The first box plot has been used to show the length of tweet for each sentiment (Fig7), second the tweet length distributions for each airline (Fig 8), third the tweet length distributions for negative reasons (Fig 9).

In (Fig 9) 'Cancelled flights 'is the reason for the longest tweets, but the reason for the shortest tweets cannot be identified. The long text of tweets with negative sentiments makes sense as, when the passengers are unhappy, they write a lot to express their anger. In (Fig.8) The longest tweets are for US airways, and the shortest tweets are for Delta.

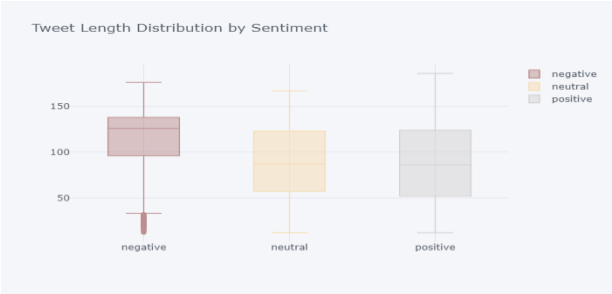


Fig 7: The length of tweets per sentiment.

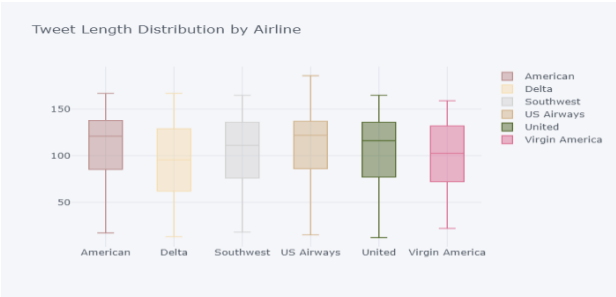


Fig 8: The lengthy of tweets per airline.

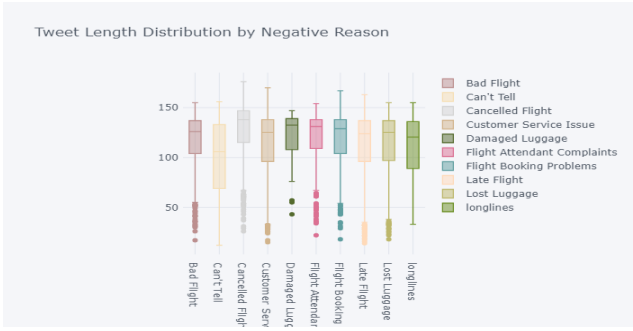


Fig 9: The length of tweets per sentiment per negative reasons.

Data preprocessing

The missed value has been checked in our data. Three columns named 'tweet_coord','airline_sentiment_gold', and 'negativereason_gold' with more than 90% null value were removed. In addition, the useless feature `tweet_id` column, which does not improve our goal, was removed.

V. Experimental results

The implementation of the proposed framework used in the experiments. Our work applied by used python notebook environment 3.7 software (Jupyter version 3.6) for ML models; and for Deep Learning, we used Colab notebook environment because we needed access to GPUs.

Evaluation of ML Models

This section reviews ML models that were used to predict sentiments of US airlines. The data was split into three sets train set, validation set, and test set. A train and test sets were created using a 70% split. Our training data was then further reduced by 15% to create a validation set. We then evaluated the model on both the training and validation sets.

A. Naïve Bayes classifier

This model was applied the same pipeline in all experiments. First, with Bag-of-words, the pipeline contains a Count Vectorizer function that converts a set of text documents to the matrix of token counts, and multi nominal NB function. Second, with the Tweet Tokenizer function, the pipeline contains a Count Vectorizer function. In this function, we applied bigram and trigram in the n-gram range parameter and Multi-nominal NB function. Third, with TF-IDF, the pipeline contains a Count Vectorizer function. In this function, we applied bigram and trigram in the n-gram range parameter, and TF-IDF Transformer and Multi-nominal NB functions. Fourthly, with lemmatization, the pipeline contains Count Vectorizer function. In this function, we applied bigram and trigram in the n-gram range parameter, and Multi-nominal NB function. Furthermore, the noise features were removed by adding two-arguments (max_df=0.6, min_df=2). We removed the frequent words, more than 60% of the document, instead of deleting the stop words in the English language overall (most frequent words) In addition, we removed the words not frequently used in more than two documents (the rarest words). Fourth, with spacy, the pipeline contains a Count Vectorizer function. In this function, we applied bigram and trigram in the n-gram range parameter, and Multi nominal NB function. Fifth, with Word2vec, we applied with n_vectors =1500000, and transform Word2vec features into values between (0, 1).

Table 3: Naïve Bayes result with several tokenizers in Twitter US Airline Sentiment dataset.

Improvements	F1 - score	Recall	Precision	Accuracy	Log Loss
Count Vectorizer	0.778689	0.778689	0.778689	0.778689	0.816139
Tweet Tokenizer	0.778689	0.778689	0.778689	0.778689	0.816139
Tweet Tokenizer & Bigram	0.757286	0.757286	0.757286	0.757286	2.206178
Tweet Tokenizer & Trigram	0.738160	0.738160	0.738160	0.738160	3.209842
Lemma Tokenizer & Bigram	0.820128	0.820128	0.820128	0.820128	1.088646
Lemma Tokenizer & Trigram	0.817851	0.817851	0.817851	0.817851	1.447698
TF-IDF & Bigram	0.699454	0.699454	0.699454	0.699454	0.690528
TF-IDF & Trigram	0.703552	0.703552	0.703552	0.703552	0.730271
Spacy & Bigram	0.820583	0.820583	0.820583	0.820583	1.021812
Spacy &Trigram	0.828324	0.828324	0.828324	0.828324	0.471871
Word2Vec	0.632969	0.632969	0.632969	0.632969	0.837912

B. LOGISTIC REGRESSION

logistic regression provides a better result than a naïve Bayes. The table shows the result of several tokenizers used to improve the performance of our model with avoiding overfitting. We used Count Vectorizer with Tweet Tokenizer, Lemmatization, TF-IDF, Spacy, and Word2Vector. We added some features such as Remove Noisy Features by lemmatization and removal of stop words and Added Important Features Bi-grams and Trigrams. In Logistic regression, the same pipeline was used in Naïve Bayes instead of the model with the Logistic regression model with add regularization strength argument. Further, there was a significant improvement in the Training set, but not on the Validation Set, which hints at overfitting. We removed noisy features before adding the Bigrams.

Table 4: Logistic regression result with several tokenizers in Twitter US Airline Sentiment dataset.

Improvements	F1 -score	Recall	Precision	Accuracy	Log Loss
Count Vectorizer	0.819672	0.819672	0.819672	0.819672	0.489313
Tweet Tokenizer	0.822860	0.822860	0.822860	0.822860	0.472417
Tweet Tokenizer& Bigram	0.827869	0.827869	0.827869	0.827869	0.478758
Tweet Tokenizer& Trigram	0.823770	0.823770	0.823770	0.823770	0.483324
Lemma Tokenizer & Bigram	0.826958	0.826958	0.826958	0.826958	0.495467
Lemma Tokenizer & Trigram	0.822860	0.822860	0.822860	0.822860	0.493465
TF-IDF& Bigram	0.811475	0.811475	0.811475	0.811475	0.512719
TF-IDF &Trigram	0.806011	0.806011	0.806011	0.806011	0.523110
Spacy & Bigram	0.828324	0.828324	0.828324	0.828324	0.471871
Spacy & Trigram	0.830601	0.830601	0.830601	0.830601	0.470926
Word2Vec	0.801913	0.801913	0.801913	0.801913	0.522759

EVALUATION OF DL MODELS

A. CNN

Neural Networks have many hyperparameters that are required to be set before training begins. However, the learning rate should be tuned, which governs the degree to which weights are adjusted during training. In our experiment, we used the maximal learning rate associated with a still-falling loss (prior to the loss diverging). Based on the plot below, we started with a learning rate of 0.001. Second, the number of epochs (epochs) to train. In our experiment, we invoke autofit without supplying the number of epochs. The training will automatically stop when the validation loss fails to improve.

Table 5: CNN Structure

Layer	Layer type	Parameters
1	Input	
2	Embedding	Max features=5000, Embedding dimension=100, Input length= max length=30
3	SpatialDropout1D	0.5
4	Convolutional	filters=32, kernel size=1, activation= Relu
5	Batch Normalization	
6	Dropout	0.5
7	Pooling	GlobalAveragePooling1D
8	Dense	units=32, activation= Relu
9	Batch Normalization	
10	Dropout	0.5
11	Dense	output=3, activation= SoftMax

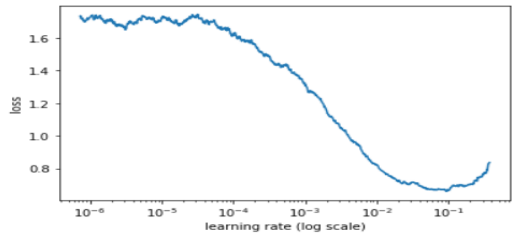


Fig10: CNN Learning Rate1e-3

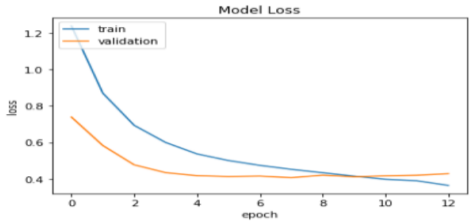


Fig 11: CNN autofit method

B. BERT

Bert-Base-Uncased was used to achieve our goals. The first step preprocessed data and created a Transformer Model. Second, we trained using the autofit method with the Maximum Learning Rate. We tested different values for different parameters, and finally we identified some of the best Values for fine-tuning. Third, we evaluated and inspected the BERT Model.

Table 6: Bert-Base-Uncased parameters

batch size	learning rate	max seq length
32	2e-5	30

.

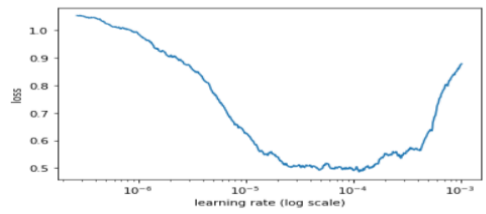


Fig 12: BERT Learning Rate 2e-5

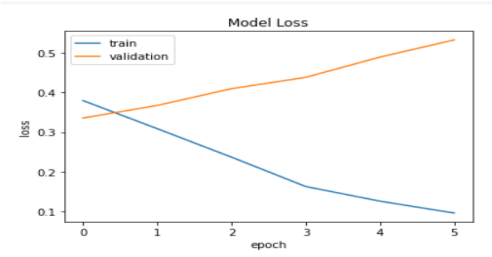


Fig13: BERT autofit method.

C. XLNET

The same parameters had been used in XLNET, as mentioned in BERT. More Transformer models are used in text classification because our dataset is small. In comparing results with the CNN model, we obtained a higher accuracy of 0.8911 in BERT rather than 0.8775 on CNN. In BERT with binary classifications, we found a higher accuracy than the multiple classifications, 0.9741 rather than 0.8911. 0

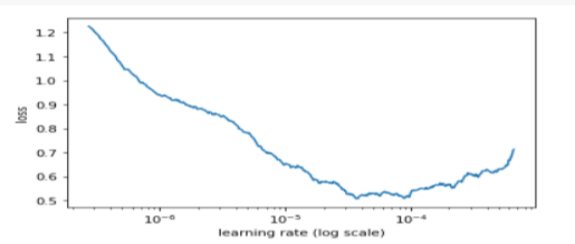


Fig 14: XLNET Learning Rate 2e-5.

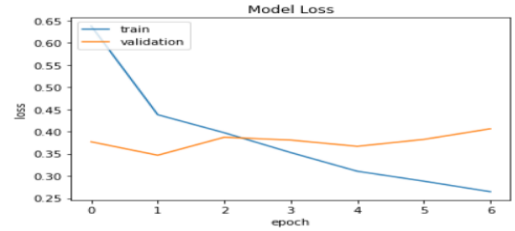


Fig 15: XLNET autofit method.

D. ALBERT

The parameters that had been used in this transform model (ALBERT-base-v2) are the same that had been applied in BERT and XLNET, and we used it with the multiclass and binary class since it gives the comparable result as Bert with little difference.

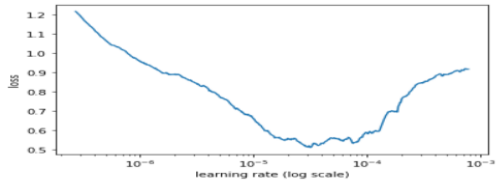


Fig16: ALBERT Learning Rate 2e-5

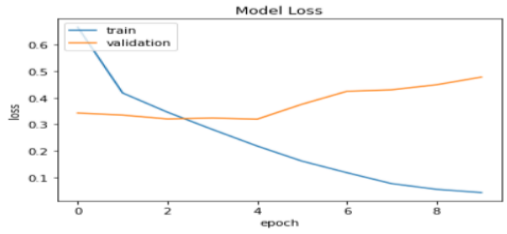


Fig17: ALBERT autofit method.

The results demonstrate that the performance of machine learning approaches is not better than deep learning approaches in text of Twitter data. We show the Confusion Matrix for our six approaches to examine how it was doing in each of the three classes.

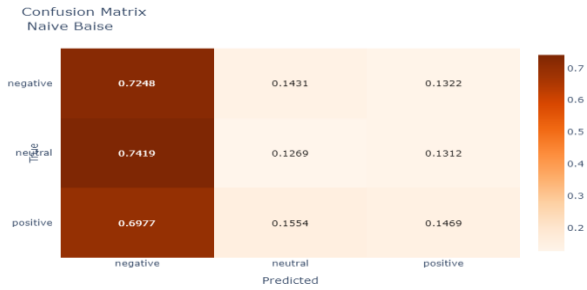


Fig 18: Naïve Bayes & Spacy &Trigram Confusion Matrix.

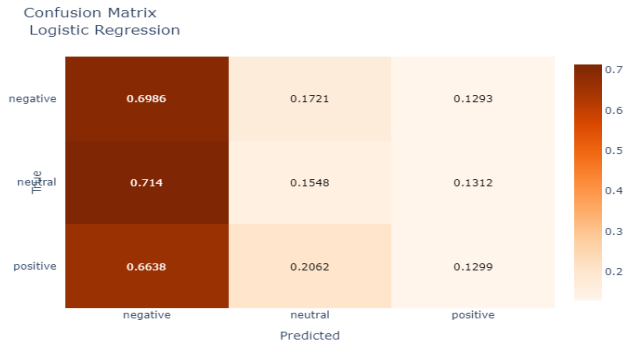


Fig 19: Logistic Regression with Spacy &Trigram Confusion Matrix.

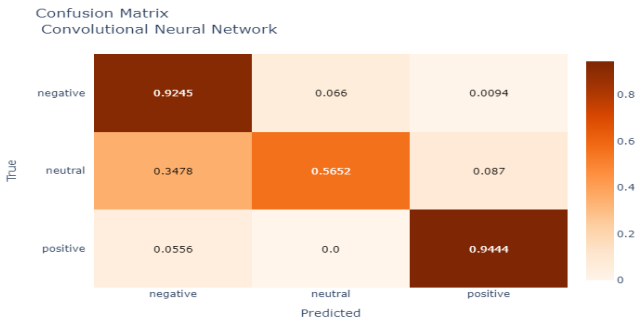
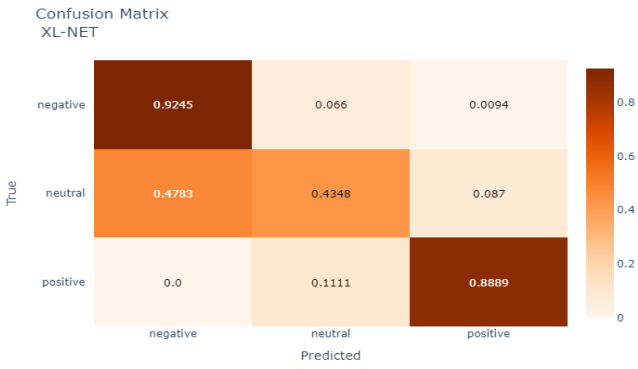


Fig 20 : CNN Confusion Matrix.



Figu21: XLNET Confusion Matrix.

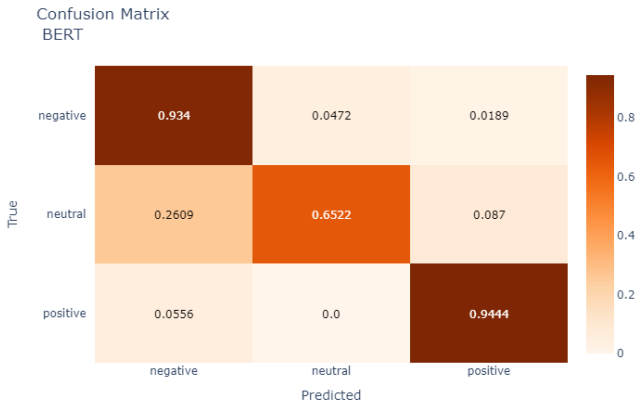


Fig 22: BERT (multi class) Confusion Matrix.

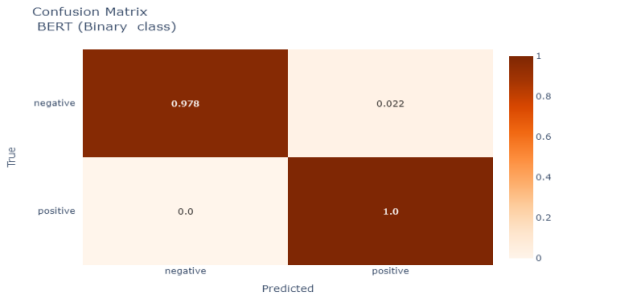
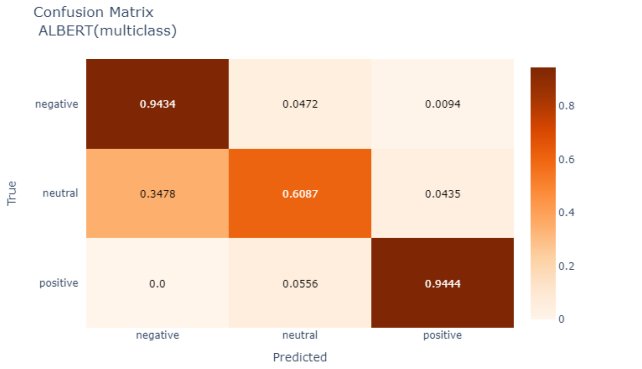


Fig 23: BERT (Binary class) Confusion Matrix.



.Fig 24: ALBERT (multi class) Confusion Matrix.

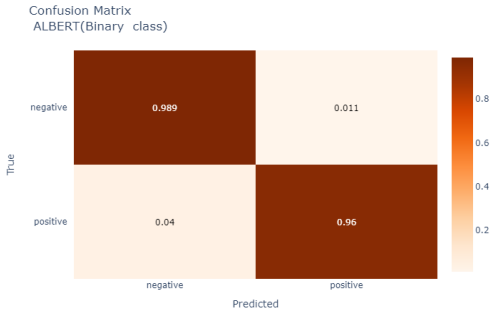


Fig 25: ALBERT (Binary class) Confusion Matrix.

Table 7: Result of each classifier for (MULTI-CLASS) in Twitter US Airline Sentiment dataset

	F1-SCORE	RECALL	PRECISION	ACCURACY	LOG LOSS
NB& Spacy &Trigram	0.8283	0.8283	0.8283	0.8283	0.4719
LR & Spacy &Trigram	0.8306	0.8306	0.8306	0.8306	0.4709
CNN	0.8710	0.8775	0.8697	0.8775	0.4108
XLNET	0.8709	0.8707	0.8713	0.8707	0.3756
BERT	0.8862	0.8911	0.8856	0.8911	0.3349
ALBERT	0.8943	0.8979	0.8930	0.8979	0.3341

Table 8: Result of BERT and ALBERT for (BINARY-CLASS) in Twitter US Airline Sentiment dataset.

	F1-SCORE	RECALL	PRECISION	ACCURACY	LOG LOSS
BERT	0.9745	0.9741	0.9769	0.9741	0.0516
ALBERT	0.9827	0.9827	0.9827	0.9827	0.0685

For Logistic regression and Naïve Bayes models, we obtained the best result from the results shown in tables 5.1 and 5.2. The best overall models are BERT and ALBERT with 0.8911, 0.8979 in MULTI-CLASS and 0.9741, 0.9827 in BINARY-CLASS respectively, and achieving the best performance while in constraint. The lowest performance was seen for Naïve Bayes 0.8283. In general, the results in deep learning models outperform the results in machine learning models in our experiment.

VI. Conclusion and Future Work

In this research, we conducted experiments on the US Airline dataset with six classification methods to predict customer sentiment prediction on tweets. Two Machine Learning algorithms were used, which are Logistic Regression (LR), Naïve Bayes (NB), and four Deep Learning algorithms such as Convolutional Neural Networks (CNN), BERT, XLNET, and ALBERT. We analyzed how model transform, feature extraction, and the number of classes affects classification results. The ML algorithms are applied using different feature extraction approaches such as Bag-of Word (Bigram, Trigram), TF-IDF with (Bigram, Trigram), Spicy with (Bigram, Trigram), Word2Vec. The best results of both models were with Spacy and Trigram. In addition, with DL algorithms BERT and ALBERT are applied with binary classes and multiple classes. The best results of both models were with binary classes. We used different evaluation measurements such as accuracy, precision, recall1, F1 score, and log loss to demonstrate the effectiveness of our model.

all algorithms proposed in this experiment, it can be stated that both BERT and ALBERT methods outperform all other algorithms implemented in this work. Especially with binary sentiment tasks. This novel method works very efficiently on the text data for sentiment analysis.

Potential future research involves developing solutions for handling the imbalance in the current dataset. We hypothesize that this could improve the results. Conducting additional experiments using a new transform model, and using multiple languages, such as Arabic reviews as well as English, could be beneficial, as could using 10-fold cross-validation to evaluate different hyperparameters for the deep neural methods. Moreover, the amount of data available in our study might have, to some extent, affected the accuracy of the deep learning classifiers.

ACKNOWLEDGMENT

I would like to appreciation my supervisor Asst Prof.Yvette Graham, for her continuous support and guidance.

REFERENCES

1. S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," Journal of Big Data, vol. 6, no. 1, 2019.

2. D. D. Das, S. Sharma, S. Natani, N. Khare, and B. Singh, "Sentimental Analysis for Airline Twitter data," in IOP Conf. Series: Materials Science and Engineering, 2017.

3. H. Hakh, I. Aljarah, and B. Al-Shbou, "Online Social Media-based Sentiment Analysis for US Airline companies," 2017.

4. A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 2018.

5. M. J. Adarsh and P. Ravikumar, "An Effective Method of Predicting the Polarity of Airline Tweets using sentimental Analysis," 2018.

6. M. Santhoshb, A. H. Krishnanb, T. Kumar, and R. Sudhakar, "Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach," International Journal of Engineering Research & Technology (IJERT), vol. 7, no. 01, 2019.

7. F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Tweets Classification on the Base of Sentiments for US Airline Companies," Entropy , Nov. 2019.

8. M. Al-Ghalibi, A. Al-Azzawi, and K. Lawonn, "NLP based sentiment analysis for Twitter's opinion mining and visualization."

9. D. Tiwari and N. Singh, "Ensemble Approach for Twitter Sentiment Analysis," I.J. Information Technology and Computer Science, Aug. 2019.

10. R. Khan and S. Urolagin, "Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data," IJACSA) International Journal of Advanced Computer Science and Applications,vol. 09, no. 6, 2018.

11. S. L and S. N. Deshmukh, "Sentiment Analysis Using SVM and Maximum Entropy," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 08.

12. P. Yuan, Y. Zhong, and J. Huang, "Sentiment Classification and Opinion Mining on Airline Reviews," 2016. [Online]. Available: https://www.semanticscholar.org/paper/1-Sentiment-Classification-andOpinionMiningYuan/daf1d9de4066eed1d193847cae578389da16c5e8.

13. T. Hemakala and S. Santhoshkumar, "Advanced Classification Method of Twitter Data using Sentiment Analysis for Airline Service," International Journal of Computer Sciences and Engineering, vol. 6, no. 7, pp. 331–335, 2018.

14. Ashi MM, Siddiqui MA, Nadeem F. Pre-trained word embeddings for Arabic aspect based sentiment analysis of airline tweets. Adv Intell Syst Comput.2019;845:245–51

15. Hakh, H.; Aljarah, I.; Al-Shboul, B. Online social media-based sentiment analysis for us airline companies. In Proceedings of the New Trends in Information Technology, Amman, Jordan, 25–27 April 2017; pp. 176–181.

16. Liu, Y.; Bi, J.W.; Fan, Z.P. Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. Expert Syst. Appl. 2017, 80, 323–339.

17. Catal, C.; Nangir, M. A sentiment classification model based on multiple classifiers. Appl. Soft Comput. 2017, 50, 135–141.

18. Dzisevic, R.; Sesok, D. Text Classification using Different Feature Extraction Approaches. In Proceedings of the IEEE 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 25–25 April 2019.

19. Sternburg F, Pedersen KH, Ryelund NK, Mukkamala RR, Vatrapu R. Analysing customer engagement of Turkish airlines using big social data. In Proc: 2018 IEEE international congress on big data (Big Data Congress), San Francisco, 2–7 July 2018.

20. Yun Wan, Dr.Qigang Gao, An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis, IEEE, 2015.

21. J. Brownlee, "How to Prepare Text Data for Machine Learning with scikit-learn," Machine Learning Mastery, 07-Aug-2019. [Online]. Available: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/ .[Accessed: 23-Apr-2020].

22. D. Karani, "Introduction to Word Embedding and Word2Vec," Medium, 02-Sep-2018. [Online]. Available: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa. [Accessed: 23-Apr-2020].

23. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications. 2013 Apr;6(2):256-61.

24. Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners ' guide to) convolutional neural networks for sentence classification," arXiv preprint arXiv:1510.03820, 2015.

25. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 2018.

26. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding." 2019.

27. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS," in ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS, 2019.

28. F. Eight, "Twitter US Airline Sentiment," Kaggle, 16-Oct-2019. [Online]. Available: https://www.kaggle.com/crowdflower/twitter-airline-sentiment.

29. A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," towards data science.[Online].Available:https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

30. L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay." 2018.