Sentiment Analysis of Customer Reviews on Twitter for US Airlines

Hyeonu(Eric) Kim

Abstract

Nowadays, companies rely heavily on customer reviews and interactions with their customers on social media. Twitter, in particular, has become the most prominent platform for exchanging ideas and thoughts between customers and companies. In this paper, we researched American airline companies and their customer reviews on Twitter by performing sentiment analysis on their customers' tweets. To understand the real thoughts and emotions of the customers towards these airlines, a series of data analyses, text processing (such as special tokens, padding, and attention masks), a Sentiment classifier using BERT, and the Hugging Face Transformers library have been used. The input of the system is the text captured in the tweets, and the output is the classification of these tweets into three classes, positive, negative, or neutral. The model performed at an accuracy of 84%.

1 Introduction

Over the last decade, social media has had an exponential burst of users. It has become a place where it is possible for people to speak out and share their daily ideas and opinions. The social media platform Twitter, where 500 million tweets are tweeted per day [1], has become particularly one of the main platforms for daily opinions.

Nowadays, companies rely heavily on customer reviews and interaction with customers on social media. Companies can take advantage of customers' feedback and utilize the information to adapt and improve their products and services. The airline industry is a major field in the transportation market [2] and relies heavily on customer feedback. Their customers' loyalty and satisfaction are the key components of the airlines' success. With the help of sentiment analysis of customers' tweets and shared feedback, it is possible to understand the thoughts, emotions, and ideas of the airlines' customers. This would significantly help airlines to make informative decisions and changes based on feedback to succeed in this red ocean of business and competition.

Our project will investigate US airlines' customer reviews on Twitter by performing sentiment analysis on a series of tweets. The machine learning model and classifier used is the BERT base with 12 Transformer Encoders, 12 self-attention heads, and a classifier [3]. This model has been chosen based on literature reviews and previous research, showing better performance of this model compared to previous models. With this work, we aim to provide the airlines with an effective and informative solution to understand their consumers' thoughts and emotions to act proactively towards their satisfaction and, in the long run, a chance of winning the market share.

2 Related Work

Sentiment analysis is an exciting growth area of Natural Language Processing (NLP) that focuses on classifying text with either positive, neutral, or negative sentiment. The text can be a document, a paragraph, or a sentence, and the assumption behind sentiment analysis is that the entire text conveys only one polarity. It has recently also become popular both within academic circumstances and industry [4].

The BERT model, or in other words, the Bidirectional Encoding Representation for a Transformer model, has been proven to be effective for feature engineering as it transforms the text into word embeddings [5]. It preserves the context of a word in a way that the meaning of a word depends on the surrounding words, and it feeds all input at once to take care of these dependencies. It has already been proven that using word embedding improves the accuracy of sentiment classification [6].

Other research attempts on the same dataset have been done in the past [7]; where the researcher created a Word2Vec model which performed at an accuracy of 69.4% however, we hypothesized that our model would perform better due to the implementation of self-attention using the BERT model.

3 Experimental setup

Experiments are done on the dataset provided by Kaggle ”Twitter US Airline Sentiment” [8]. The dataset contains a series of customers' tweets scraped from the Twitter accounts of several airlines in February 2015. The dataset has 14,641 tweets with 15 different attributes, including $airline\_sentiment$, $airline\_sentiment\_confidence$, $negativereason$, $airline$, and $text$, which contain more significant important information modeling and classification in our research.

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count
570306133677760513	neutral	1.0			Virgin America		cairdin		0
570301130888122368	positive	0.3486		0.0	Virgin America		jnardino		0
570301083672813571	neutral	0.6837			Virgin America		yvonnalynn		0
570301031407624196	negative	1.0	Bad Flight	0.7033	Virgin America		jnardino		0
570300817074462722	negative	1.0	Can't Tell	1.0	Virgin America		jnardino		0
570300767074181121	negative	1.0	Can't Tell	0.6842	Virgin America		jnardino		0
570300616901320704	positive	0.6745		0.0	Virgin America		cjmcginnis		0
570300248553349120	neutral	0.634			Virgin America		pilot		0

Figure 1: Sample of the dataset

text	tweet_coord	tweet_created	tweet_location	user_timezone
@VirginAmerica What @dhepburn said.		2015-02-24 11:35:52 -0800		Eastern Time (US & Canada)
@VirginAmerica plus you've added commercials to the experience... tacky.		2015-02-24 11:15:59 -0800		Pacific Time (US & Canada)
@VirginAmerica I didn't today... Must mean I need to take another trip!		2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse		2015-02-24 11:15:36 -0800		Pacific Time (US & Canada)
@VirginAmerica and it's a really big bad thing about it		2015-02-24 11:14:45 -0800		Pacific Time (US & Canada)
@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA		2015-02-24 11:14:33 -0800		Pacific Time (US & Canada)
@VirginAmerica yes, nearly every time I fly VX this "ear worm" won't go away :)		2015-02-24 11:13:57 -0800	San Francisco CA	Pacific Time (US & Canada)
@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP		2015-02-24 11:12:29 -0800	Los Angeles	Pacific Time (US & Canada)

Figure 2: Sample of the dataset

The data and text pre-processing were inspired by previous studies and research in the field. The BERT model is a base BERT model with 12 encoders, 12 attention mask heads, and a classifier at the end for the classification approach. The performance is evaluated by the Accuracy Score of how well the model and classifier are classifying the tweets into three classes (negative, positive, and neutral) based on training, validation, and test datasets.

3.1 Data Analysis

The dataset contains 15 attributes such as airline, text, $airline\_sentiment$, $airline\_sentiment\_confidence$, $negativereason$, $airline$, and $text$, etc. There are 14640 data entries with some empty cells in the dataset as no information was provided. The summary of the dataset is shown in the table below.

```
 #   Column                        Non-Null Count
---  ------                        --------------
 0   tweet_id                      14640 non-null
 1   airline_sentiment             14640 non-null
 2   airline_sentiment_confidence  14640 non-null
 3   negativereason                9178 non-null
 4   negativereason_confidence     10522 non-null
 5   airline                       14640 non-null
 6   airline_sentiment_gold        40 non-null
 7   name                          14640 non-null
 8   negativereason_gold           32 non-null
 9   retweet_count                 14640 non-null
 10  text                          14640 non-null
 11  tweet_coord                   1019 non-null
 12  tweet_created                 14640 non-null
 13  tweet_location                9907 non-null
 14  user_timezone                 9820 non-null
```

Figure 3: Statistics Data Attributes

The main player in this dataset is “United” airline”, with 3822 tweets. There are 6 airlines in total receiving tweets from their customers. Based on our analysis, there are three sentiments of the tweets with the distribution shown below. Also, the main reason for negative feedback or tweets is indicated as a “Customer Service Issue” by the users. The sentiment split for each airline is shown in the figure below. United, US Airways, and American have the highest negative sentiment values compared to the rest, and the weight of negative sentiments is much higher compared to neutral and positive sentiments.

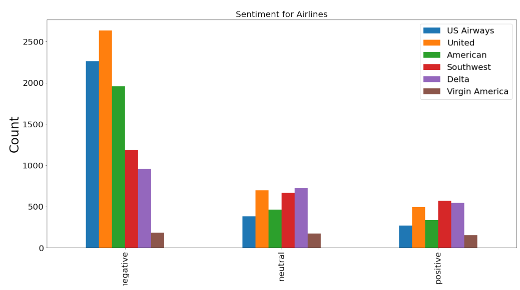

Figure 4: Sentiment Split for Airlines

And when focusing on the reason behind the negative sentiments, it is clear, as expected, “Customer Service” is the reason for the negative sentiments in these airlines.

Our attempts for visualizing WordCloud of text and sentiments have not provided us with much insight into the analysis, as it seems it requires more processing on the text sentiment analysis.

	negativereason
Customer Service Issue	2910
Late Flight	1665
Can't Tell	1190
Cancelled Flight	847
Lost Luggage	724
Bad Flight	580
Flight Booking Problems	529
Flight Attendant Complaints	481
longlines	178
Damaged Luggage	74

Figure 5: Negative reasons

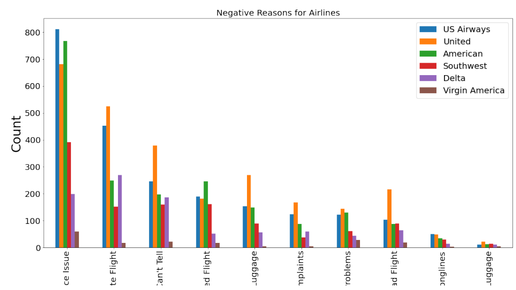

Figure 6: Negative reasons per Airline

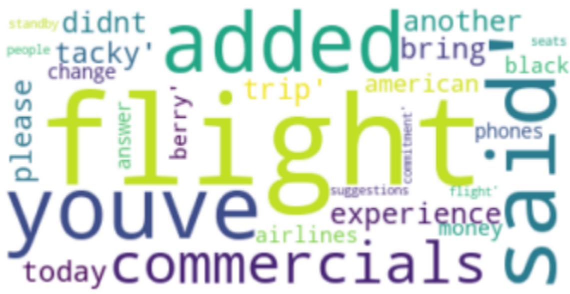

Figure 7: Wordcloud

3.2 Classification Model

We utilized the BERT model [9-15] to classify between 'negative', 'neutral', and 'positive' reviews. First, we separated so that 75% of the data was used for training/validating, and 25% of the data was used for testing. We removed the first word as it is always $@airline\_name$, then we separated each sentence by [CLS] and [SEP] tokens to distinguish the beginning and end of sentences as shown below.

We changed the classes into numeric values so 'negative' = 0, 'neutral' = 1, and 'positive'

Figure 8: Text Pre-processing

= 2. Since all sentences have different lengths, we created a fixed length of 32/64 to either truncate or include padding tokens. Each sequence was vectorized using BertForSequenceClassification.from_pretrained('bert-base-multilingual-cased') since we have tweets from other languages, as shown below.

```
array([  101,   137, 13953, 24769, 85137, 10425, 12976,   137, 11034,
       10410, 35497, 12415,   119,   102,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0])
```

Figure 9: An example of padded input vector

We then created Attention mask by simply if there is no padding = 1, else = 0. Now with our vectorized sentence with padding, and its label, and attention mask we created a DataLoader with help of pytorch. We used ADAM as our optimizer as ADAM does not need advanced fine tuning compared to SGD. We trained and tested with different hyperparameters which resulted with a 84% accuracy on testing set. The outputs from the testing set are shown below.

```
array([[-3.4206972 , -1.9133492 ,  5.0094476 ],
       [ 5.7396345 , -1.7464681 , -3.513607  ],
       [ 5.7170067 , -1.750731  , -3.3972566 ],
       [ 5.73867   , -1.7958874 , -3.4620507 ],
       [-0.19044094,  0.1870742 ,  0.48233527],
       [ 5.7077327 , -1.7117519 , -3.481931  ],
       [ 5.7194304 , -1.5590582 , -3.5067801 ],
       [ 0.98410547,  1.7419723 , -3.4113681 ]], dtype=float32)
```

Figure 10: An example of output from testing set

After concatenating, argmax we got the result of the class which gave '0'(Negative) with text being: 'why load us on the flight if the captain was over the hours he could fly in one consecutive period? unacceptable' which is true.

4 Results

The model performed at an accuracy of 84%. Below is the confusion matrix for our model. It shows the difference between the model prediction and ground truth labels of the tweets.

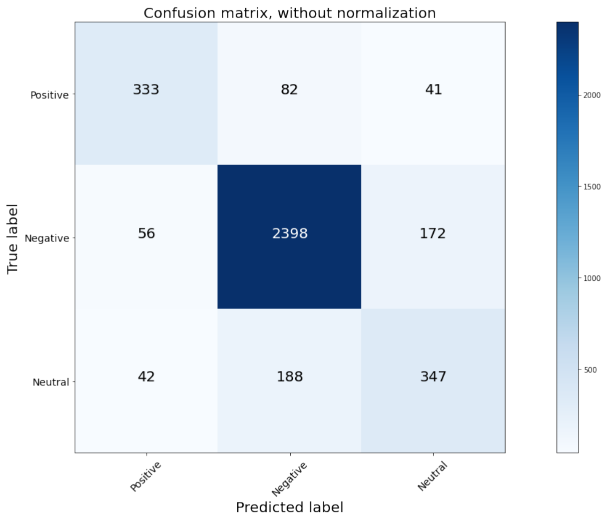

Figure 11: Confusion Matrix of Model Prediction and Ground Truth

As can be seen from the pie chart below, the classification of true negative labels had the highest accuracy, however this may be due to the class imbalance leaning toward negative labels and thus having 'an easier time' to classify tweets as negative rather than positive and neutral.

	predictions	truth	diff	count
Neg-Neu	Negative	Neutral	0	188
Neu-Neg	Neutral	Negative	0	172
Neg-Pos	Negative	Positive	0	82
Pos-Neg	Positive	Negative	0	56
Pos-Neu	Positive	Neutral	0	42
Neu-Pos	Neutral	Positive	0	41
Neg-Neg	Negative	Negative	1	2398
Neu-Neu	Neutral	Neutral	1	347
Pos-Pos	Positive	Positive	1	333

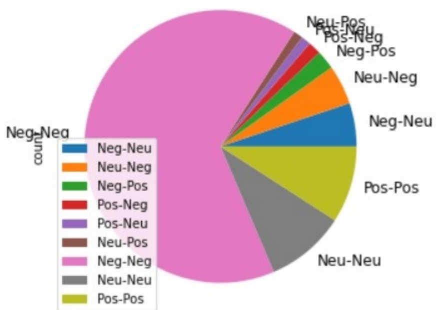

Figure 12: Statistics of Model Prediction and Ground Truth

5 Conclusion & Future Work

To conclude, our model using BERT performed better than previous research with Word2Vec on the same dataset we believe this is due to the utilization of self-attention.

Additionally, we think this is a passable model for airlines to utilise to get a better insight into their customer's sentiment regarding the services which are provided.

For further improvement of the model we would like to take into account cross-validation method for training, class imbalance, emoticons and change abbreviations as we believe this would make the model more robust and able to capture more knowledge regarding the words and context of the tweets.

6 References

[1] Twitter Usage Statistics, https://www.internetlivestats.com/twitter-statistics/

[2] E. Prabhakar, M. Santhosh, A. H. Krishnan, T. Kumar, and R. Sudhakar, 'Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach', International Journal of Engineering Research, vol. 7, no. 01, p. 3, 2019.

[3] C. Sun, X. Qiu, Y. Xu, and X. Huang, 'How to Fine-Tune BERT for Text Classification?', in Chinese Computational Linguistics, vol. 11856, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham: Springer International Publishing, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3$_1$6.

[4] C. Sun, L. Huang, and X. Qiu, 'Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence'. arXiv, Mar. 22, 2019. Accessed: Dec. 07, 2022. [Online]. Available: http://arxiv.org/abs/1903.09588

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. arXiv, May 24, 2019. Accessed: Dec. 07, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] B. Oscar Deho, A. William Agangiba, L. Felix Aryeh, and A. Jeffery Ansah, 'Sentiment Analysis with Word Embedding', in 2018 IEEE 7th International Conference on Adaptive Science Technology (ICAST), Accra, Aug. 2018, pp. 1–4. doi: 10.1109/ICASTECH.2018.8506717.

[7] Airlines Tweet Analysis Trial by Word2Vec and LSTM, 2022, https://www.kaggle.com/code/sasakitetsuya/airlines-tweet-analysis-trial-by-word2vec-and-lstmWord2Vec

[8] Twitter US Airline Sentiment, 2019, https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment?datasetId=17

[9] V. Moshkin, A. Konstantinov, and N. Yarushkina, 'Application of the BERT Language Model for Sentiment Analysis of Social Network Posts', in Artificial Intelligence, vol. 12412, S. O. Kuznetsov, A. I. Panov, and K. S. Yakovlev, Eds. Cham: Springer International Publishing, 2020, pp. 274–283. doi: 10.1007/978-3-030-59535-7_20.

[10] E. Kouloumpis, T. Wilson, and J. Moore, 'Twitter Sentiment Analysis: The Good the Bad and the OMG!', ICWSM, vol. 5, no. 1, pp. 538–541, Aug. 2021, doi: 10.1609/icwsm.v5i1.14185.

[11] M. J. Blosseville, G. Hébrail, M. G. Monteil, and N. Pénot, 'Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together', in Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92, Copenhagen, Denmark, 1992, pp. 51–58. doi: 10.1145/133160.133175.

[12] B. Pang and L. Lee, 'Opinion mining and sentiment analysis', p. 94.

[13] S. Kale and V. Padmadas, 'Sentiment Analysis of Tweets Using Semantic Analysis', in 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), PUNE, India, Aug. 2017, pp. 1–3. doi: 10.1109/ICCUBEA.2017.8464011.

[14] D. Ramachandran and R. Parvathi, 'Analysis of Twitter Specific Preprocessing Technique for Tweets', Procedia Computer Science, vol. 165, pp. 245–251, 2019, doi: 10.1016/j.procs.2020.01.083.

[15] Chris McCormick, BERT fine tuning Tutorial with Pytorch, 2019, https://mccormickml.com/2019/07/22/BERT-fine-tuning/