ECE 537 Data Mining

Final Project Report

Social Media Data Mining to gauge positive or negative responses towards a company

Student names: Muhammad Mustafa

Department name: Computer Information Science

1. Introduction:

Airline customers increasingly use Twitter and other social media to share their experiences during travel. This creates a massive amount of unstructured text that reflects the consumers' views about airline companies. Past research has proven that sentiment analysis is an effective methodology in the extraction of meaningful opinions from such data, which helps organizations understand trends in customer satisfaction, operational issues, and service quality [1, 2, 3, 4]. In this research, the application of sentiment classification techniques in airline related tweets is performed by both traditional machine-learning models, such as Logistic Regression, Naive Bayes, and Support Vector Machine (SVM), and a deep-learning approach based on BERT. The report details steps in data preparation, the model development process, and evaluation results for classical TF-IDF-based models against the transformer-based BERT model to find out which provides the most reliable sentiment prediction for reviews involving airline services.

2. Methods:

This project follows a complete data-mining pipeline that classifies airline related tweets into positive, neutral, or negative sentiment. It presents results from combining a traditional machine learning model with a transformer-based deep-learning model to evaluate differences in performance and behavior.

2.1. Data Mining Technologies:

The dataset was prepared to include only the attributes relevant to sentiment classification, namely the text of the tweet, airline name, sentiment label, and information on negative reasons. Since social media text is often noisy and inconsistent, careful data preparation was necessary to provide clean and reliable input to the models [6]. Once prepared, numerical representations of the text data were created using the Term Frequency Inverse Document Frequency (TF-IDF) technique.

TF-IDF emphasizes important words within the dataset and is widely used in sentiment-analysis and text-classification tasks, particularly with traditional machine-learning models [4], [6]. In addition to the TF-IDF-based methods, this project also uses BERT, a transformer model that produces contextual embeddings for each token. Unlike TF-IDF, which treats words independently, BERT processes entire sentences and can capture more rational patterns in the tweets [3].

2.2. Implementation:

Using the TF-IDF feature matrix, three classical classifiers were implemented:

- **Logistic Regression:** Linear model applied because of its stability and good performance on high-dimensional text data, serving as a strong baseline.
- **Naive Bayes:** a probability model which assumes word independence and is efficient with big documents.

- **SVM:** A linear SVM, selected for its good results in text classification and the ability to separate sentiment classes in a sparse TF-IDF space.

All the models were trained using an 80/20 train-test split. Their outputs were then evaluated using accuracy, precision, recall, F1-scores, and confusion matrices to understand each model's strengths and weaknesses, following standard practices in sentiment-analysis research [2, 4, 5].

- **BERT:** To compare the classical approaches with a modern deep-learning method, a pretrained bert-base-uncased model was fine-tuned for sentiment classification. The tweets were tokenized into input IDs and attention masks and fed into a classification layer. Training was carried out using the API with a fixed sequence length and batch size to balance performance and efficiency, consistent with deep-learning approaches used in prior Twitter sentiment analyses [3].

2.3. Methodological Observations:

The TF-IDF models, during implementation, showed good performance on negative and neutral tweets because the keyword patterns are clear in them. High dimensional feature space is handled effectively by both SVM and Logistic Regression, whereas Naive Bayes struggled with sentimental classes that share similar vocabulary. BERT, with higher computational requirements, was able to capture the context and sensitive wording during training, providing a unique perspective of sentiment classification as opposed to older feature-based models.

3. Experiments:

This dataset consists of airline-related tweets along with their corresponding sentiment labels. Only the essential attributes were kept, which are: tweet text, airline name, sentiment category, and negative-reason data, like prior airline Tweet sentiment datasets [1, 2, 5].

3.1. Experiments Conducted:

Different experiments were designed to compare the performance of traditional machine-learning models with a transformer-based model on the same dataset. Three classical models were adopted: Logistic Regression, Naive Bayes, and Support Vector Machine. Each of them was trained with TF-IDF feature representations. Each model was tested on the test set to observe how well it distinguished between positive, neutral, and negative sentiments.

Running in parallel, a fine-tuned BERT model was trained to see the effect of contextual embeddings on sentiment prediction. In contrast to the TF-IDF models that rely on word frequency patterns, BERT processes entire sentences and captures semantic relationships. The same train-test split was held constant across experiments for consistency.

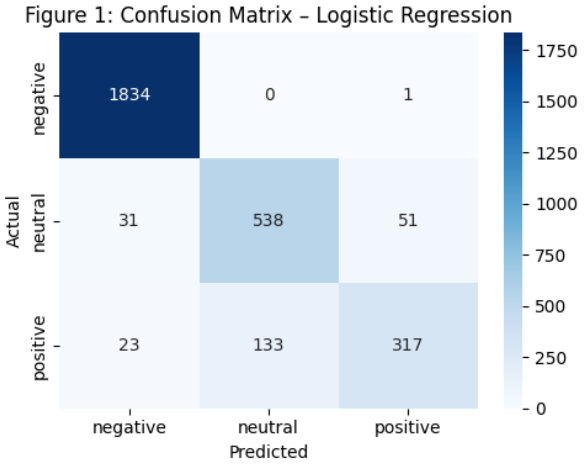
All models are evaluated using accuracy, precision, recall, F1-score, and confusion matrices. These experiments allow for a direct comparison of how traditional linear and probabilistic approaches differ from a modern deep-learning model when applied to real-world airline tweets.

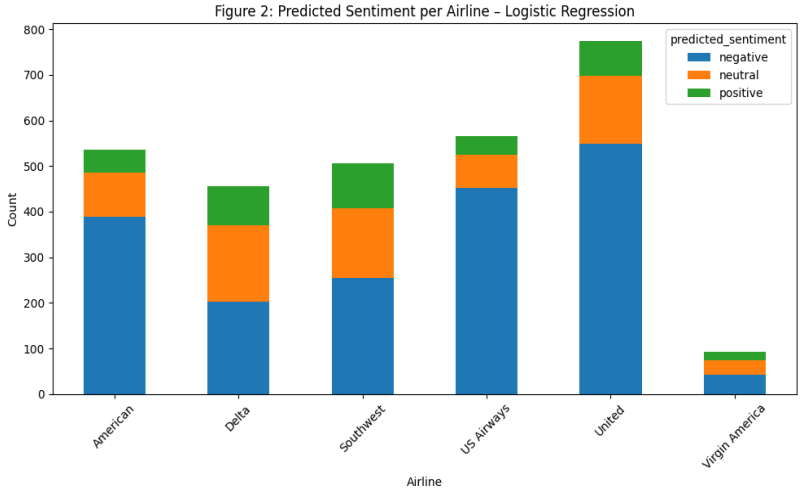
3.2. Results and Discussion:

Logistic Regression:

Logistic Regression achieved an accuracy of 91.8%, having a great F1 performance on negative tweets at 0.99 and quite fair on neutral and positive tweets, having F1 scores of 0.83 and 0.75, respectively. Looking at the confusion matrix (Figure 1), one can see very strong performance for negative sentiment but more confusion

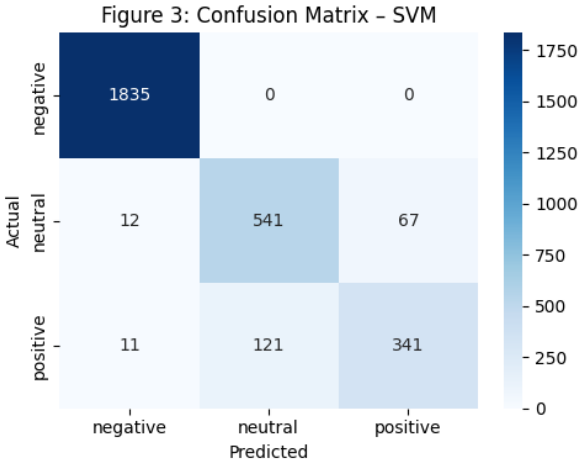
between neutral and positive classes. Similarly, negative sentiment dominates across all carriers for airline-level predictions, as seen in (Figure 2).

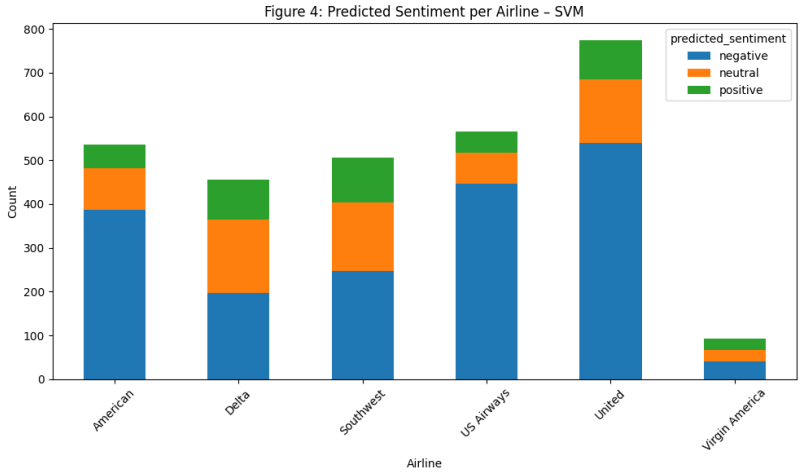




Support Vector Machine (SVM):

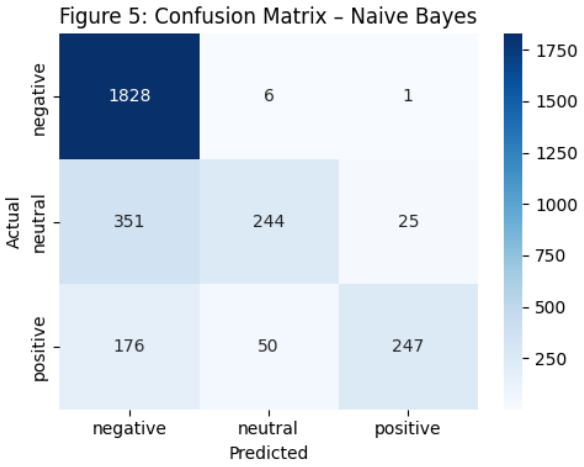
The SVM delivered the strongest performance among the classical models, with an accuracy of 92.8%. The model obtained high F1-scores for negative (0.99), neutral (0.84), and positive tweets (0.77). The confusion matrix (Figure 3) shows fewer misclassifications than Logistic Regression, especially for positive tweets. Airline sentiment distribution (Figure 4) seems comparable across carriers; this reflects the general tendencies of sentiment in the dataset.

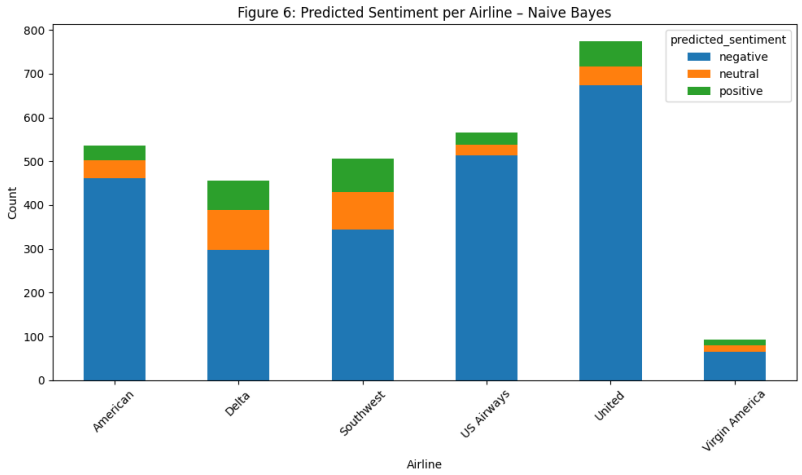




Naive Bayes:

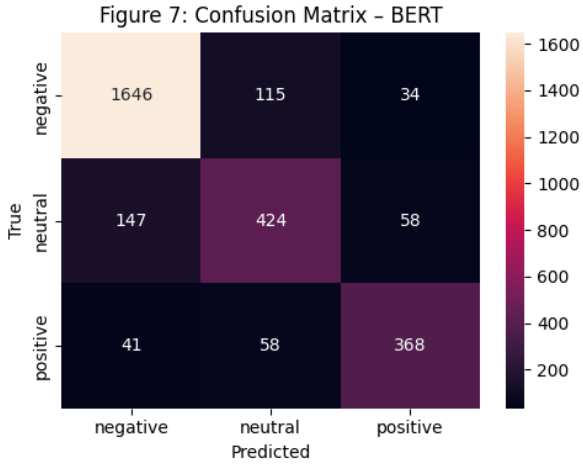
Naive Bayes had the weakest performance with only 79.2% accuracy. Although it perfectly identified negative tweets, recall = 1.00, it had very poor performance on neutral and positive ones with F1 of 0.53 and 0.66, respectively. Its confusion matrix (Figure 5) displays a lot of neutrals and positives misclassified as negatives. The airline sentiment chart (Figure 6) portrays Naïve Bayes predicts substantially more negatives compared to other models.





BERT:

The accuracy of the BERT model was 84%, which is lower than that of the classical TF-IDF models but well balanced across the classes. On the other hand, BERT supported F1-scores of 0.91 for negative, 0.70 for neutral, and 0.79 for positive. The confusion matrix (Figure 7) reflects better positive tweet classification by BERT compared to Naive Bayes and Logistic Regression. From this observation, BERT can capture deeper levels of contextual meaning, although further fine-tuning could increase the performance even further.



Model	Accuracy	Negative F1 Score	Neutral F1 Score	Positive F1 Score
Logistic Regression	**91.8%**	**0.99**	**0.83**	**0.75**
SVM	**92.8%**	**0.99**	**0.84**	**0.77**
Naïve Bayes	**79.2%**	**0.87**	**0.53**	**0.66**
BERT	**84%**	**0.91**	**0.70**	**0.79**

4. Conclusion:

This project applied a full pipeline of sentiment analysis to airline-related tweets, using both traditional machine-learning models and a deep-learning approach using BERT. Logistic Regression, Naive Bayes, and SVM were implemented with TF-IDF features, while BERT was fine-tuned to evaluate the advantage of contextual

understanding. By conducting these experiments, I can compare how each model handles real-world text data and identifies their strengths and weaknesses.

One of the biggest challenges from this project involved working with the deep-learning model BERT. BERT, compared with the other models, required more computational resources, careful preprocessing, and several adjustments before it finally trained. These complications aside, fine-tuning the model proved quite instructive in how transformer architectures pick up on sentiment differently from traditional methods.

With this project, I learned how data preparation, feature extraction, and model selection directly influence the performance of sentiment classification. I also gained practical experience in implementing both classical and modern NLP techniques and evaluated their behavior on an imbalanced dataset.

Overall, SVM tended to provide the highest accuracy, while BERT showed more balanced performance across sentiment types for positive tweets. The Naive Bayes model performed the weakest, because this approach has a problem with overlapping vocabularies. These results point to the strengths of linear models on high-dimensional TF-IDF data and contextual advantages provided by transformer-based architecture.

References:

[1] A. S. Shitole and A. S. Vaidya, "Machine learning based airlines tweets sentiment classification," International Journal of Computer Applications, vol. 185, no. 20, pp. 32-35, Jul. 2023. [Online].

[2] F. Rustam et al, "Tweets Classification on the Base of Sentiments for US Airline Companies," Entropy (Basel, Switzerland), vol. 21, (11), pp. 1078, 2019.

[3] W. Aljedaani et al, "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry," Knowledge-Based Systems, vol. 255, pp. 109780, 2022.

[4] M. D. Devika, C. Sunitha and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," Procedia Computer Science, vol. 87, pp. 44-49, 2016.

[5] G. Ravi Kumar, K. Venkata Sheshanna and G. Anjan Babu, "Sentiment analysis for airline tweets utilizing machine learning techniques," in International Conference on Mobile Computing and Sustainable Informatics, J. S. Raj, Ed. Switzerland: Springer International Publishing AG, 2020, pp. 791-799.

[6] L. Zhang and B. Liu, "Sentiment analysis and opinion mining," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2017, pp. 1152-1161.

[7] A. Hassan and A. Mahmood, "Deep learning for sentence classification," in 2017. DOI: 10.1109/LISAT.2017.8001979.

[8] J. Devlin *et al*, "BERT: Pre-training of deep bidirectional transformers for language understanding," Cornell University Library, arXiv.org, Ithaca, 2019.