Sentiment Analysis of Social Media for Airline Customer Service Improvement

Sujith Chandra Dara[1] and Bhargava Ramani Gorle[2]

[1] *Computer science, Illinois Institute of Technology, A20528384, sdara@hawk.iit.edu*
[1] *Computer science, Illinois Institute of Technology, A20528476, bgorle@hawk.iit.edu*

Abstract—Customer engagement has changed dramatically since the introduction of social media, especially Twitter, especially in service-oriented industries like aviation. Businesses looking to assess client sentiment and make necessary adjustments must comprehend and leverage this abundance of data. Using machine learning methods to extract emotions from textual data, sentiment analysis becomes a crucial tool for interpreting consumer opinions about brands, goods, or services. In the aviation sector, sentiment analysis of Twitter data reveals subtle client sentiments, allowing airlines to evaluate customer happiness, pinpoint areas for development, and proactively handle customer complaints. Using cutting-edge machine learning approaches to handle the complexity of human language and reliably capture sentiment nuances, this research focuses on building strong sentiment analysis models customized for airline-related tweets.

Contents

1. Introduction

Social media platforms, especially Twitter, have completely changed the dynamics of consumer interaction and have emerged as vital spaces where users can instantly share their experiences, voice their thoughts, and get feedback. This shift is particularly important for service-oriented industries like aviation, where customer happiness has a big impact on operational performance and brand reputation. Therefore, it is now essential for businesses to use the enormous amount of social media data if they want to remain aware of the opinions and preferences of their customers.

Sentiment analysis becomes an essential tool in the digital arena for companies trying to glean insights from the barrage of social media posts. Sentiment analysis provides essential insights into client attitudes and perceptions towards companies, products, or services by utilizing computer algorithms to extract feelings, opinions, and emotions hidden in textual data. For airlines, sentiment analysis of Twitter data unveils nuanced customer sentiments, enabling them to assess satisfaction levels, pinpoint areas for enhancement, and proactively address passenger concerns.

To better understand sentiment analysis in the context of the aviation business, this research will specifically use data from Twitter. The research attempts to develop strong models that can reliably classify tweets about airlines into sentiment classes from positive and neutral to negative—by integrating machine learning methods with natural language processing techniques. The initiative intends to allow airlines in an increasingly digitally driven and linked world to make informed decisions, improve customer experiences, and strengthen brand reputation by offering actionable insights generated from social media data. Social networking sites like Twitter have been increasingly important in customer-business relations since the beginning of the digital era, completely changing the way businesses connect with their customers.

Particularly in service-oriented industries like airlines, where customer feedback on social media profoundly impacts brand perception and operational success, understanding and leveraging sentiment analysis tools are paramount. These tools allow businesses to decode customer sentiments, identify trends, and proactively address issues, thereby enhancing overall customer experiences and fortifying brand reputation in an increasingly digital and interconnected environment.

1.1. Background and Context

Social media's pervasive influence has made it a vital medium for customers to share their experiences. In the airline industry, a single tweet can shape public perception significantly. Airlines, therefore, need to monitor and analyze customer sentiment continuously to maintain service standards and improve customer satisfaction.

1.2. Problem Statement

Despite the vast potential offered by sentiment analysis of social media data, analyzing tweets presents unique challenges. The brevity, informality, and linguistic nuances inherent in tweets pose obstacles to accurate sentiment classification. Factors such as sarcasm, slang, and the presence of mixed sentiments within individual messages complicate the task of automated sentiment analysis. Consequently, achieving high accuracy in sentiment classification requires navigating the intricate landscape of human language expressed through tweets.

This project seeks to address these challenges by leveraging advanced machine learning techniques to develop robust sentiment analysis models tailored specifically for airline-related tweets. By overcoming the complexities of natural language and effectively capturing the nuances of sentiment expressed in tweets, the project aims to provide airlines with actionable insights derived from social media data, facilitating informed decision-making and proactive customer engagement strategies.

1.3. Objectives of the Sentiment Analysis

The primary objective of this project is to harness the power of machine learning algorithms to create a robust model capable of precisely categorizing the sentiment conveyed in tweets addressed to airlines. By automating the sentiment analysis process, the project seeks to facilitate swift and informed decision-making, enabling airlines to respond proactively to customer feedback in real-time. Through this endeavor, the project aims to streamline the analysis of sentiments expressed on social media platforms, particularly Twitter, with a focus on enhancing the efficiency of customer service strategies. By developing a reliable and automated sentiment analysis tool, airlines can gain valuable insights into customer perceptions and preferences, ultimately improving overall customer satisfaction levels and bolstering brand reputation.

1.4. Significance of the Study in the Airline Industry

From strategically changing policies to making modifications to frontline services, understanding consumer opinion has a big impact on operations. The study offers valuable insights into consumer sentiment for the airline sector, which can be leveraged to improve customer experience management and facilitate more informed strategic decision-making.

2. Data Description

The dataset forms the backbone of the sentiment analysis project, providing the raw material upon which our models are trained, tested, and evaluated.

2.1. Overview of the Dataset Used

The dataset consists of a collection of tweets directed at various airlines. It encompasses a broad spectrum of customer interactions, from service appreciation to grievances about flight experiences. The dataset's richness and variety offer a realistic challenge for sentiment analysis models aiming to interpret and classify customer sentiments accurately.

2.2. Source of the Data

The data has been sourced from Kaggle, an online community of data scientists and machine learning practitioners providing access to datasets and hosting competitions. The airline tweet dataset can be found at this URL: Tweets Dataset on Kaggle
.

2.3. Explanation of Data Attributes

The key attributes of the dataset that are particularly salient for our analysis include:

- tweet_id: A unique identifier for each tweet, ensuring each entry is distinguishable and verifiable.
- airline_sentiment: The core attribute of interest, this is the annotated sentiment of the tweet categorized as positive, neutral, or negative.
- text: The actual text content of the tweet, which serves as the primary input for our analysis.
- airline: This attribute specifies the airline to which the tweet is directed, offering context and specificity to the sentiment expressed.
- retweet_count: This feature provides an indication of the tweet's reach and, potentially, the intensity of sentiment within the network.

The data attributes are curated to shed light on customer sentiment, with each tweet acting as a data point reflecting an individual's experience or opinion. The comprehensive nature of the dataset, with 14,640 entries, offers a robust platform for training and validating sentiment analysis models.

3. Methodology

The methodology section elaborates on the systematic approach taken to prepare the data for analysis and the steps involved in developing the sentiment analysis models.

3.1. Data Preprocessing

Effective data preprocessing is the linchpin of successful machine learning outcomes, particularly in the domain of natural language processing. The preprocessing steps undertaken for this project are as follows:

3.2. Data Cleaning and Normalization

The project commenced with data cleaning and normalization, which encompassed eliminating extraneous characters such as symbols and numbers, standardizing text cases, and rectifying typographical errors. This initial phase was crucial for reducing dataset noise and ensuring consistency across entries. Following the loading of the dataset to comprehend its structure, we proceeded with cleaning by removing irrelevant information and standardizing the remaining data, a process that forms the foundation for subsequent preprocessing steps.

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0

Figure 1. The initial structure of the dataset upon loading.

3.3. Handling Missing Values

In addressing missing values within the dataset, we adopted strategies to maintain the reliability of our machine learning models. Notably, we encountered missing data in columns like 'negativereason_confidence' for non-negative tweets, which we rectified by filling these gaps with zeros, ensuring the dataset's integrity remained intact for analysis purposes.

Additionally, non-essential columns were eliminated to streamline the dataset, focusing solely on relevant data conducive to sentiment analysis. Moreover, to ensure consistency in temporal data, we normalized the 'tweet_created' timestamps, enhancing the dataset's coherence and facilitating more accurate analysis.

3.4. Sentiment Distribution

To understand the overall sentiment bias in the dataset, we visualized the distribution of sentiments across the tweets.

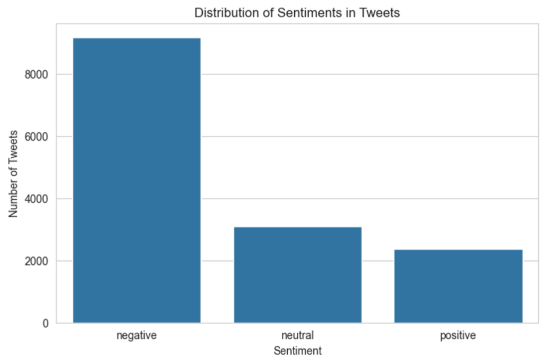

Figure 2. Bar chart illustrating the distribution of sentiments within the dataset.

3.5. Tweet Length Distribution by Sentiment

To investigate possible relationships with sentiment, we looked at the duration of tweets in many sentiment categories.

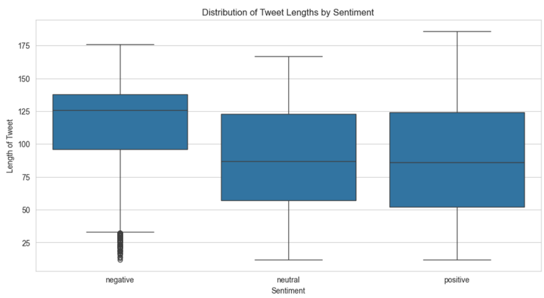

Figure 3. Boxplot showing the variation in tweet lengths across different sentiment categories

3.6. Text Vectorization Techniques Employed

The Term Frequency-Inverse Document Frequency (TF-IDF) approach was used to transform text data into a format that could be processed by machine learning algorithms. By converting the text into a numerical representation, TF-IDF highlights the significance of words in relation to how frequently they appear in documents in this case, individual tweets.

3.7. Feature Selection and Engineering

The primary feature utilized for the sentiment analysis was the tweet text itself. However, additional features were also considered, such as the length of tweets. These features were carefully selected to bolster the models' ability to discern and accurately classify sentiments.

3.8. Serialization of Preprocessed Data

The preprocessed dataset was serialized using pickle, facilitating easy access and reproducibility for future analysis.

4. Model Development and Training

4.1. Overview of the Models Selected

Three different models that were specifically designed for sentiment analysis of tweets on airline services were to be developed and tested for the project. These models were chosen in light of their individual merits and capacities for managing the intricacies involved in textual data analysis:

1. **Logistic Regression:** Designed to be a comparative standard for more sophisticated models, the baseline model was chosen due to its ease of use and effectiveness in managing linear relationships. When it comes to sentiment analysis jobs where class distributions may differ greatly, Logistic Regression is well-known for its effectiveness in managing imbalanced datasets in addition to its interpretability and scalability. Furthermore, it enables sophisticated study of sentiment probabilities due to its probabilistic output capabilities.
2. **Support Vector Machine (SVM):** SVM was chosen because of its proficiency in managing the intricate borders separating sentiment classes and because of its resilience in high-dimensional environments, such as text data. SVM ensures reliable sentiment classification even in difficult situations since, in addition to its robustness and versatility, it works incredibly well in handling noisy data and outliers. By transforming data into higher dimensions, its kernel technique improves the accuracy of capturing complicated relationships in text data.
3. **Long Short-Term Memory (LSTM):** Text data analysis is a particularly good application for Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs), which are known for their ability to capture order dependence in sequence prediction tasks. Sentiment analysis applications including lengthy tweets or reviews can benefit from the superior sequence handling and memory preservation of LSTM, in addition to its proficiency with contextual comprehension. The model's precision in interpreting the evolution of sentiment is improved by its capacity to capture temporal relationships over long time spans.

With the help of these several models, the study sought to offer a thorough examination of the sentiment conveyed in tweets pertaining to airlines, utilizing the special advantages of each model to produce precise and complex sentiment classification.

4.2. Training Process

Each model underwent training on a dataset that was preprocessed and vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) methodology. This transformation of the text data into a numerical format was crucial as it emphasized the significance of each word within the dataset, enabling the models to effectively process and analyze the sentiment. The training process for each model involved specific steps:

Logistic Regression and SVM: The vectorized text data was used to train these models, with an emphasis on modifying model parameters to maximize performance. These models' precision in sentiment classification can be attributed to their ability to represent the significance of words through TF-IDF vectorization.

Long Short-Term Memory (LSTM): The LSTM's capacity to capture temporal relationships and its intricate architecture need additional preprocessing steps throughout the training phase. In order to ensure that the model could accurately analyze the context and word order in sentiment expression, texts were transformed into sequences in order to maintain their sequential nature. The LSTM model was then trained using these sequences, which helped it identify and precisely capture the complex patterns found in the text data.

4.3. Hyperparameter Tuning Approach

Using GridSearchCV and RandomizedSearchCV, a methodical approach to hyperparameter tuning was used to improve the models' performance, especially for Logistic Regression and SVM. This procedure consisted of:

- Defining a grid of hyperparameter ranges.
- Using cross-validation to evaluate each combination.
- Selecting the combination that produced the best results based on predefined scoring metrics (e.g., accuracy).

For LSTM, hyperparameters such as the number of units, dropout rates, and the number of training epochs were tuned to find the optimal settings that balance training efficiency with prediction accuracy.

4.4. Model Evaluation and Comparison

Each model's performance was assessed based on accuracy, precision, recall, and F1-scores:

- The accuracy of 79.51% offered by logistic regression served as a reliable foundation for assessing more intricate models.
- Performance of SVM was found to be similar or somewhat better; hyperparameter tweaking helped to further improve it.
- Deep learning techniques have potential for natural language processing applications, as demonstrated by the performance of LSTM, which demonstrated encouraging results, particularly in capturing the subtleties of sequential data inherent in text.

A thorough analysis was conducted to examine the models in more depth, taking into account not only accuracy but also each model's handling of the various sentiment classes, with a special emphasis on striking a balance between precision and recall.

4.5. Visualizations

Several visualizations were produced to illustrate the training dynamics and performance of each model:

1. Confusion Matrix for SVM: This visualization provides a detailed breakdown of true vs. predicted classifications, highlighting the model's performance across different sentiment classes.

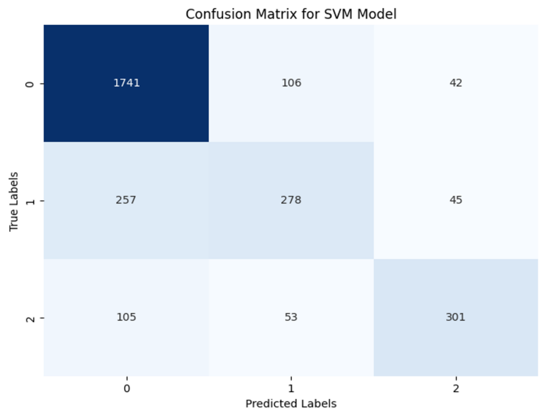

Figure 4. Confusion matrix visualization for the SVM model, showing the number of correct and incorrect classifications per class..

2. Learning Curves for SVM: These curves, which show the model's performance with time for both training and validation sets, are helpful in determining if a model is under- or overfitted.

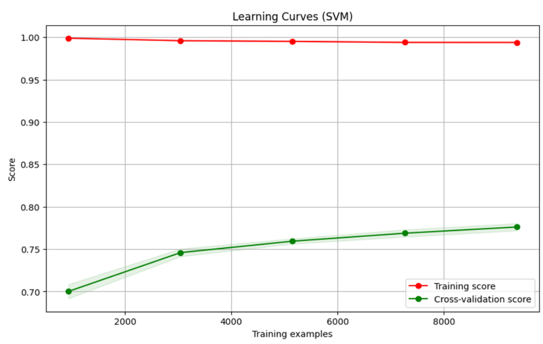

Figure 5. Learning curves for the SVM model, illustrating changes in training and validation accuracy and loss over epochs.

3. Classification Report for SVM: Comprehensive report that provides a more nuanced picture of the model's performance than just accuracy by displaying precision, recall, and F1-score for every class.

Classification Report for SVM Model:

	precision	recall	f1-score	support
negative	0.83	0.92	0.87	1889
neutral	0.64	0.48	0.55	580
positive	0.78	0.66	0.71	459
accuracy			0.79	2928
macro avg	0.75	0.69	0.71	2928
weighted avg	0.78	0.79	0.78	2928

Figure 6. Detailed classification report for the SVM model, highlighting performance metrics for each sentiment category.

5. Results and Discussion

Each model's performance was evaluated in-depth using several measures, such as F1-score, accuracy, precision, and recall. These measures shed light on how well the algorithms balance properly predicting each class with accurately classifying sentiments.

5.1. Logistic Regression Model Results:

- Accuracy: 79.51
- Precision, Recall, and F1-Score:
- Negative Sentiments: Precision: 82
- Neutral Sentiments: Precision: 65
- Positive Sentiments: Precision: 79

The logistic regression model served as a solid baseline, performing particularly well with negative sentiments but showing room for improvement in accurately classifying neutral sentiments.

5.2. SVM Model Results:

- Accuracy: 79.23
- Precision, Recall, and F1-Score:
- Negative Sentiments: Precision: 83
- Neutral Sentiments: Precision: 64
- Positive Sentiments: Precision: 78

With a little increase in the categorization of positive sentiments, the SVM model outperformed the logistic regression model. The optimal hyperparameter settings were 'kernel': 'rbf', 'gamma': 1, 'C': 100, which improved the model's performance in managing high-dimensional spaces.

5.3. LSTM Model Results:

- Test Accuracy: 77.12
- Epoch-wise Performance: The model showed consistent improvements over epochs, with the highest accuracy reached at 95.36
- Precision, Recall, and F1-Score:
- Negative Sentiments: Precision: 86
- Neutral Sentiments: Precision: 54
- Positive Sentiments: Precision: 72

The LSTM model demonstrated its efficiency in handling sequential data, although exhibiting considerable overfitting as seen by a gradual decline in validation accuracy. It performed better than the other models in the difficult task of managing neutral sentiments.

5.4. Discussion of Results from Each Model

Analyzing the outcomes from each model offers insights into the nuances between traditional machine learning methods and sophisticated deep learning techniques:

- **Performance Overview:** The results showcase strong performance across all models in identifying negative sentiments, likely attributed to their prevalence in the dataset. However, differentiating between neutral and positive sentiments proved challenging, underscoring the intricacies of sentiment analysis within concise, informal texts like tweets.

- **Model Applicability:** Logistic Regression and SVM emerge as favorable choices for rapid deployment and scenarios where interpretability holds paramount importance. Conversely, LSTM's proficiency in capturing word sequence dynamics grants it a superior edge in nuanced sentiment comprehension, albeit requiring higher computational resources and longer training periods.
- **Optimization and Avoiding Overfitting:** While fine-tuning hyperparameters optimized SVM's performance, the training trajectory of LSTM suggests potential overfitting, necessitating further adjustments such as integrating dropout layers or regularization techniques to enhance model robustness.

5.5. Model Comparisons

5.5.1. Comparison of Logistic Regression, SVM, and LSTM Results

The analysis involves a comparative review of the three models: Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks. Each model has its strengths and weaknesses, which become evident through their performance metrics and the way they handle different sentiment classifications.

Accuracy Overview:

1. Logistic Regression showed robust performance with an overall accuracy of approximately 79.51%.
2. SVM displayed slightly lower accuracy at 79.23% but improved handling of positive sentiments.
3. LSTM had a test accuracy of 77.12%, with higher precision and recall for neutral sentiments compared to the other models.

Handling of Sentiment Classes:

- **Negative Sentiments:** Logistic Regression and SVM both performed exceptionally well, with both models achieving high precision and recall. LSTM also performed well but slightly lower compared to Logistic Regression and SVM.
- **Neutral Sentiments:** This was challenging for all models, with LSTM showing a slight advantage in handling these, indicating its effectiveness in capturing nuances through sequence modeling.
- **Positive Sentiments:** SVM and LSTM demonstrated similar capabilities, outperforming Logistic Regression in terms of recall.

5.6. Visualizations:

To visually compare the performance across these models, several key graphs were generated:

1. Confusion Matrices: Each model's confusion matrix helps visualize the correct and incorrect classifications across the sentiment categories. These matrices highlight how well each model performs in distinguishing among negative, neutral, and positive tweets.

	precision	recall	f1-score	support
negative	0.82	0.93	0.87	1889
neutral	0.65	0.49	0.56	580
positive	0.79	0.63	0.70	459
accuracy			0.80	2928
macro avg	0.76	0.68	0.71	2928
weighted avg	0.79	0.80	0.78	2928

Figure 7. Confusion matrix for the Logistic Regression model showing detailed class-wise performance.

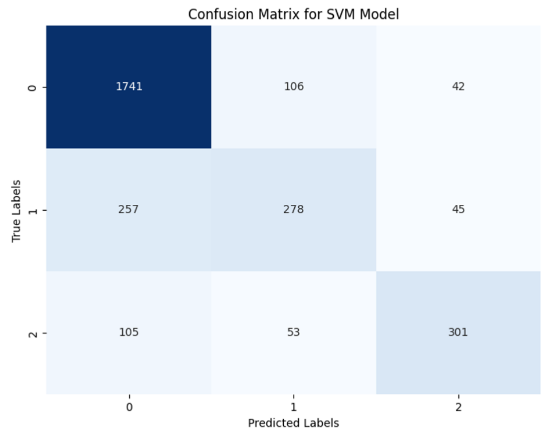

Figure 8. Confusion matrix for the SVM model, indicating class-wise accuracy and misclassifications.

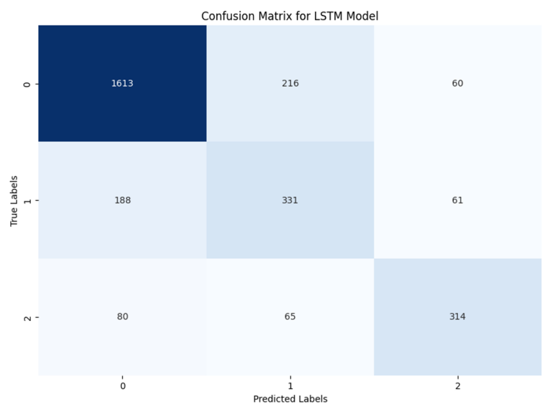

Figure 9. Confusion matrix for the LSTM model, highlighting its proficiency in handling sequential data and sentiment nuances.

2. Learning Curves: The learning curves for each model provide insights into their learning behavior over epochs (for LSTM) or iterations (for Logistic Regression and SVM). These curves are instrumental in identifying overfitting or underfitting at various stages of the training process.

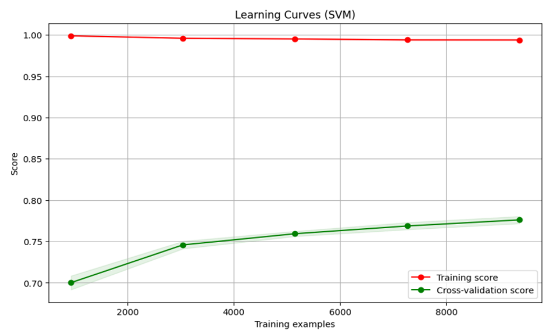

Figure 10. Learning curves for the SVM model, illustrating training and validation accuracy over time.

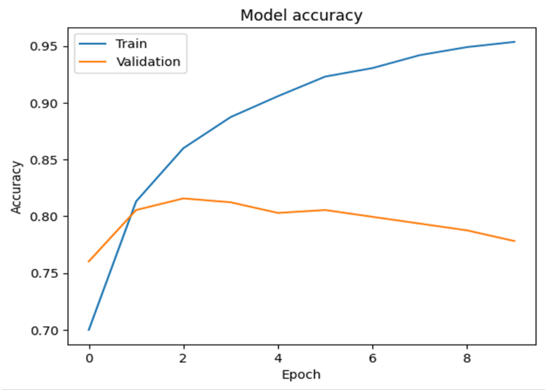

Figure 11. Learning curves for the LSTM model, showing changes in training and validation accuracy, which help in assessing the model's loss

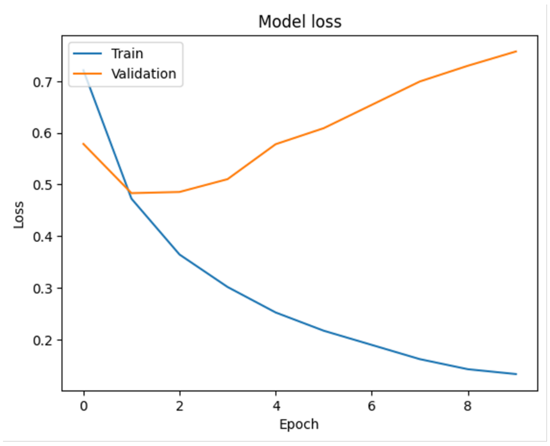

Figure 12. Learning curves for the LSTM model, showing changes in training and validation accuracy, which help in assessing the model's efficiency and potential overfitting issues.

6. Conclusions

6.1. Summary of Findings

In summarizing the findings of this study on sentiment analysis of airline tweets using Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) models, several key observations emerged:

- **Logistic Regression** established a robust baseline, showcasing notable proficiency in accurately categorizing negative sentiments within the tweets.
- **SVM** demonstrated comparable performance to Logistic Regression, showcasing slight enhancements in discerning positive sentiments, largely due to meticulous hyperparameter tuning.
- **LSTM** exhibited its prowess in capturing subtleties within text sequences, particularly in classifying neutral sentiments, despite showing indications of potential overfitting during later training epochs.

6.2. Implications of the Results

The outcomes of this analysis carry substantial implications for the airline industry, offering valuable insights into enhancing customer experiences and operational strategies. The findings enable airlines to:

- Swiftly respond to customer feedback by pinpointing both complaints and positive feedback, fostering improved customer service and satisfaction.
- Refine marketing initiatives and operational tactics based on the real-time sentiment analysis, allowing for more targeted and effective campaigns.
- Strengthen customer engagement efforts by delivering personalized and timely responses, thereby fostering stronger relationships and loyalty.
- Utilize sentiment analysis to identify emerging trends and sentiments, enabling proactive decision-making and strategic adjustments to meet evolving customer needs.
- Leverage sentiment insights to enhance brand perception, reputation management, and overall competitive advantage within the airline industry.

6.3. Limitations of the Study

While the study provides valuable insights, there are several limitations:

- **Data Imbalance:** The dataset had a disproportionate number of negative tweets, which might have biased the models towards predicting negative sentiments more accurately than neutral or positive.
- **Text Complexity:** Tweets often contain slang, abbreviations, and emojis, which pose challenges in text processing and interpretation that were not fully addressed in this study.
- **Model Generalizability:** The models were trained and tested on a specific dataset from the airline industry, which may limit their applicability to other domains or more diverse linguistic datasets.

6.4. Suggestions for Future Work

Looking ahead, future endeavors in this domain could consider the following avenues for further exploration and improvement:

1. **Augmenting Dataset Diversity:** Enriching the dataset with a broader range of sentiments and perspectives can enhance model robustness and reduce bias, ensuring a more comprehensive sentiment analysis framework.
2. **Advanced Text Processing Methods:** Investigating advanced preprocessing techniques tailored for social media language nuances, including slang, emojis, and informal expressions, can refine sentiment analysis accuracy.
3. **Sector-Generalized Models:** Extending model applicability beyond airlines to diverse industries can broaden insights and utility, fostering cross-domain sentiment analysis capabilities.
4. **Deep Learning Advancements:** Experimenting with intricate deep learning architectures like transformer models such as BERT holds promise in refining accuracy and handling nuanced textual nuances effectively.
5. **Real-time Analysis Framework:** Creating a framework for real-time sentiment analysis can empower businesses with instantaneous insights, enabling prompt and targeted actions to enhance customer experiences and brand reputation.

7. Repository and Dataset Links

1. Repository Link: Sentiment-Analysis-of-Social-Media-for-Airline-Customer-Service-Improvement
.
2. Dataset Link : Tweets Dataset on Kaggle
.

8. References

1. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc. Retrieved from O'Reilly Media.
2. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830. Retrieved from Journal of Machine Learning Research.
3. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121-167. Retrieved from Springer.

4. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. Retrieved from Nature.
5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. Retrieved from MIT Press.
6. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. Retrieved from ACM Digital Library.
7. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. Retrieved from arXiv.
8. Tang, D., Qin, B., & Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Retrieved from ACL Anthology.
9. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. Retrieved from Now Publishers.
10. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. Retrieved from Morgan & Claypool Publishers.