Sentiment Analysis of British Airways Customer Reviews Using TF-IDF, Word2Vec and LSTM,

Sara Iqbal
240053369
Master in Data Science
sara.iqbal@city.ac.uk

1 Problem statement and motivation

In today‘s virtual era, user-generated content plays a pivotal role in shaping customer notion and commercial enterprise selections. For the airline industry, online structures along with Twitter, Skytrax, and TripAdvisor have emerged as key channels via which clients specific their reviews and opinions. These evaluations—starting from reward for in-flight offerings to proceedings approximately delays—provide a rich source of insight into consumer satisfaction. However, because of the sheer extent and pace of these facts, it’s miles more and more impractical for airline businesses to manually examine and act upon every person’s comment.

This is where sentiment evaluation—a Natural Language Processing (NLP) assignment that aims to decide the emotional tone in the context of a frame of textual content—will become invaluable. By mechanically classifying evaluations along with positive, neutral, or negative, airways can effectively reveal public opinion, pick out regions for improvement, and make data-driven selections to improve the consumer experience. Moreover, sentiment insights can reveal strategic alternatives in marketing, course planning, and consumer service, in the end driving consumer loyalty and competitive advantage.

The guide inspection of textual facts now no longer handiest suffers from inefficiency but is also susceptible to subjective interpretation and inconsistency. Therefore, automating sentiment evaluation guarantees each scalability and objectivity in opinion mining. However, this assignment is non-trivial because of demanding situations along with sarcasm, context-dependence, misspellings, and domain-specific jargon, which require sturdy textual content illustration and modeling strategies to overcome.

To deal with this, our task investigates and compares 3 distinguished tactics for sentiment evaluation on airline evaluations: Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and Long Short-Term Memory (LSTM) neural networks. TF-IDF, a statistical measure that evaluates how essential a phrase is to a report in a corpus, is extensively used for its simplicity and effectiveness in shooting keyword-primarily based totally styles. However, it fails to account for semantics, which means and phrase order. Word2Vec addresses this hassle with the aid of getting to know non-stop vector representations of phrases that capture semantic relationships (Mikolov et al., 2013). It allows higher generalization with the aid of using embedding comparable phrases near each other within the vector space. Nevertheless, each TF-IDF and Word2Vec is limited of their capacity to version sequential dependencies in language.

This is in where LSTM, a kind of recurrent neural network (RNN), proves advantageous. LSTM models can getting to know long-variety dependencies in sequential facts, making them mainly powerful for textual content in which phrase order and context play a crucial position in figuring out sentiment (Hochreiter Schmidhuber, 1997). Combining Word2Vec with LSTM similarly complements the version‘s capacity to recognize nuanced language styles in evaluations.

Therefore, the important trouble this task addresses is: How are we able to lay out and evaluate device getting to know and deep getting to know strategies (TF-IDF, Word2Vec + LSTM) to correctly and effectively carry out sentiment evaluation on airline consumer evaluations? This research will offer insights into the strengths and weaknesses of every technique and manual destiny improvement of greater state-of-the-art NLP models for the aviation industry.

2 Research hypothesis

This study hypothesizes that LSTM models using Word2Vec embeddings will surpass traditional au-

tomatic learning models, such as logistic regression and naive Bayes related to TF-IDF, in the task of classifying the emotions of customers. This expectation is based on the inherent forces of the LSTM network to grasp the sequential dependence and contextual meaning in text data. Unlike TF-IDF, ignoring the order of words and processing the terms of independence, Word2Vec provides dense vectors that reflect semantic relationships between words. When combined with LSTM, these hobbies allow the model to understand more effectively complex feelings such as sarcasm, negative or mixed ideas. We plan that this learning pipeline will lead to higher classification accuracy, recovery, and F1 score. However, we realize that these models require significant data and the ability to calculate and can handle limitations if small data sets or high noise. This hypothesis is tested by a strict model evaluation.

3 Related work and background

Sentiment analysis, a key subfield of herbal language processing (NLP), makes a speciality of extracting subjective reviews and feelings from textual content. It has received big traction in latest years because of its applicability in knowledge consumer feedback, especially in domain names which include product critiques, film critiques, and increasingly, airline passenger critiques. This segment surveys foundational and latest contributions applicable to our undertaking and contextualizes our technique inside the cutting-edge research landscape.

The foundational survey through Pang and Lee (2008) gives a complete evaluation of sentiment analysis, highlighting early techniques the usage of device learning models which include Naïve Bayes, Support Vector Machines (SVM), and Maximum Entropy classifiers. These strategies generally trusted sparse, high-dimensional representations, which include bag-of-words and TF-IDF. TF-IDF, brought in early records retrieval systems, quantifies time period significance through balancing phrase frequency with its rarity throughout documents, supporting to down-weight common, non-informative phrases (Sebastiani, 2002). While TF-IDF stays a famous baseline because of its simplicity and interpretability, it fails to seize semantic relationships or phrase order, crucial in knowledge-nuanced sentiments.

To triumph over those barriers, researchers started adopting phrase embeddings, which offer dense vector representations of phrases. Mikolov et al. (2013) brought Word2Vec, a shallow, neural embedding version that maps semantically comparable phrases to close by factors in vector space. Word2Vec‘s electricity lies in its capacity to seize latent linguistic styles from massive corpora in an unmonitored manner. Our undertaking adopts pre-skilled Word2Vec embeddings to complement the semantic illustration of airline critiques earlier than feeding them into sequential models.

The barriers of conventional classifiers and fixed-vector embeddings caused the upward push of deep learning models for textual content classification. Hochreiter and Schmidhuber (1997) proposed the Long Short-Term Memory (LSTM) network, a form of recurrent neural network (RNN) that addresses the vanishing gradient problem and may research long-term dependencies in sequences. LSTMs are especially powerful for textual content due to the fact they hold contextual knowledge over sequences of phrases, that's critical for sentiment obligations regarding negation, sarcasm, or implicit cues.

Recent models construct upon the LSTM structure with interest mechanisms. For instance, Zhou et al. (2016) brought an interest-primarily based totally BiLSTM for relation classification, which selectively makes a speciality of relevant elements of the textual content all through learning. While our undertaking does now no longer put into effect interest, this work highlights the continuing evolution in NLP model architectures that similarly enhance interpretability and performance.

Domain-unique research additionally gives insights applicable to our study. Filho et al. (2021) carried out device learning techniques, along with SVM and Random Forests—to airline passenger critiques and discovered that whilst conventional models performed moderately well, deep learning strategies provided advanced accuracy. Similarly, Nguyen et al. (2020) compared deep learning architectures (CNN, LSTM) on Vietnamese airline critiques and concluded that LSTMs with pre-skilled embeddings yielded great results. These findings assist our speculation that LSTM models, especially whilst blended with semantically significant inputs like Word2Vec, are well-applicable for airline sentiment analysis.

Another line of research has centered on social media-primarily based totally airline feedback.

Chakraborty and Bhat (2019) mined Twitter facts to research provider quality, the usage of sentiment rankings to tell operational improvements. This research underscores the sensible relevance of sentiment category in shaping airline customer support strategies.

Toolkits and libraries, inclusive of TextBlob (Loria et al., 2014) offer rule-primarily based totally sentiment scoring and easy system gaining knowledge of pipelines for prototyping. Although now no longer utilized in our very last model, TextBlob has become useful for preliminary benchmarking and exploratory data analysis. Additionally, Zhang et al. (2018) provide an in-depth survey of deep gaining knowledge of techniques for sentiment analysis, advocating the shift from characteristic engineering to illustration gaining knowledge of—a transition reflected in our very own methodology.

In summary, our task synthesizes the strengths of each conventional and neural NLP pipeline. While many research both persist with classical ML strategies or embody deep gaining knowledge of, we at once evaluate those tactics in a managed experimental setting. Our paintings extends present studies with the aid of using making use of and comparing TF-IDF + classical models with Word2Vec + LSTM at the same airline evaluation dataset, presenting a grounded performance contrast among the 2 paradigms.

3.1 Accomplishments

The following tasks were proposed and carried out as part of this project:

- Task 1: Preprocess dataset — Completed

  The dataset was cleaned by removing stopwords, punctuation, and special characters. Lowercasing and lemmatization were also applied.

- Task 2: Tokenize dataset and generate embeddings — Completed

  Tokenization was performed using Keras' Tokenizer API. Two forms of vectorization were implemented: TF-IDF and Word2Vec.

- Task 3: Build and train baseline models using TF-IDF — Completed

  Classical models such as Logistic Regression, Support Vector Machines (SVM), and Random Forest were trained using TF-IDF features.

- Task 4: Build and train deep learning model using Word2Vec + LSTM — Completed

  A sequential LSTM model was developed using pre-trained Word2Vec embeddings, with dropout and regularization.

- Task 5: Compare LSTM model against baselines — Completed : A comprehensive performance comparison using metrics such as accuracy, F1-score, and confusion matrix was conducted.

- Task 6: Perform in-depth error analysis — Partially completed

4 Approach and Methodology

The goal of this project is to classify airline reviews based on sentiment using a combination of traditional machine learning and deep learning approaches. The central concept in the back of the pipeline is to examine classical TF-IDF-primarily based totally fashions with LSTM-primarily based totally architectures that use semantic phrase embeddings. The instinct is that even as TF-IDF captures period frequency well, it fails to account for contextual meaning, which recurrent models like LSTM can leverage successfully via sequential processing. By combining those approaches, we can investigate which version structure more generally generalizes throughout sentiment type duties in airline feedback.

For conventional models, we used TF-IDF vectorization via Scikit-learn's 'TfidfVectorizer' to transform textual content into sparse numerical representations. However, because of elegance imbalance, in which effective evaluations considerably outnumbered impartial and bad ones, we implemented SMOTE (Synthetic Minority Over-sampling Technique) to synthetically increase the minority classes. This ensured extra balanced schooling and promoted robustness for underrepresented sentiments. We trained a couple of baseline classifiers which including Logistic Regression, Support Vector Machines (SVM), and Random Forest the using of those TF-IDF vectors, and evaluated them the use of cross-validation and type metrics like accuracy, precision, recall, and F1-score.

In parallel, we carried out a sequence of LSTM-primarily based deep getting to know fashions the use of the Keras API with a TensorFlow backend. For phrase embeddings, we used pre-skilled Word2Vec vectors from the Google News dataset through the Gensim library. Each assessment became transformed into a series of vectors primarily based on those embeddings. We explored 3 versions of LSTM models. The first became a popular stacked LSTM with LSTM layers and dropout for regularization. The 2nd became a Bidirectional LSTM, permitting the version to study dependencies in each ahead and backward directions. The 0.33 and maximum superior versions incorporated an interest mechanism on top of a Bi-LSTM. This enabled the community to consciousness on key phrases inside an assessment that maximally contributed to the sentiment label, considerably enhancing interpretability and performance.

Each deep version included an embedding layer (initialized with Word2Vec weights), one or more LSTM layers, a dropout layer to prevent overfitting, and a very last dense layer with softmax activation for multiclass type. The fashions have been compiled with the Adam optimizer and express cross-entropy loss, and skilled the use of early stopping to keep away from overfitting. We break up the dataset into 80 training and 20 testing for truthful evaluation, and use metrics together with accuracy, precision, recall, and F1-rating alongside confusion matrices to degree performance.

Throughout development, we used Google Colab with a GPU for model training and experimentation. Libraries which includes NLTK have been used for preprocessing, Scikit-research for classical models and metrics, Gensim for Word2Vec, and Keras for deep mastering. Visualization tools like Matplotlib and Seaborn have been used to devise mastery curves, loss functions, and confusion matrices.

One of the important thing in demanding situations is coping with elegance imbalance and tuning LSTM hyperparameters like batch length and dropout rate. Another problem become handling reminiscence constraints in Colab all through education, specially with the BiLSTM + interest version. Despite those, we efficaciously constructed and evaluated a whole pipeline with significant results.

Our method has limitations—at the same time as the LSTM models outperform TF-IDF baselines on longer reviews, they once in a while battle with quick or noisy inputs. Conversely, classical models are much less sensitive to period; however, they fail to seize semantic context. Together, those methods provide a well-rounded view of the trouble and tell destiny paintings in hybrid sentiment evaluation pipelines.

5 Dataset

For this project, I used an open-supply dataset from Kaggle known as the "British Airline Sentiment Dataset." It incorporates round 11,500 client opinions of various airlines. Each evaluate comes with a sentiment label—both Positive, Neutral, or Negative. These labels make the dataset appropriate for schooling and checking out sentiment evaluation models.

A key factor to be aware is the elegance imbalance. Around 50 percent of the opinions are categorised as Negative, 30 percent as Positive, andthe most effective 20 percent as Neutral. This imbalance is a problem because models can grow to be biased towards the maximum not unusual place elegance, making them perform poorly at the much less common ones. To deal with this, I used SMOTE (Synthetic Minority Over-sampling Technique) later within the pipeline to stabilize the class distribution while the usage of conventional models like TF-IDF with logistic regression.

Some examples from the dataset display why this undertaking is difficult. For instance:

"The flight changed into behind schedule by three hours, however, the personnel have been pleasant and accommodating."

This evaluation has blended signals—it mentions a delay (negative), however, additionally praises the personnel (positive), which makes it difficult to label. Another instance is:

"Thx AA, u men r great!!!"

This sort of casual language, with abbreviations and slang, provides complexity to preprocessing and version schooling.

Basic records of the dataset display that the common evaluate period is set 40–50 phrases, aleven though a few are very brief and others are pretty long. There are over 450,000 overall phrases withinside the dataset. During preprocessing, I needed to do away with empty or beside the point entries. Also, a few opinions had spelling errors or odd formatting, which required greater cleaning.

Overall, the combinationture of casual writing, class imbalance, and mixed sentiments makes this dataset an amazing check for each conventional procedures like TF-IDF and contemporary-day fashions like LSTM. It displays real-global situations wherein client comments isn‘t continually smooth or clear, so it facilitates examine how nicely distinctive fashions can cope with complicated sentiment classification.

5.1 Dataset preprocessing

To prepare the airline reviews for sentiment analysis, I applied a sequence of common natural language preprocessing steps to clean and standardize the raw text. The main steps included:

- **Lowercasing**: All text was converted to lowercase to eliminate case sensitivity, ensuring that words like "Great" and "great" are treated the same.
- **Punctuation and Special Character Removal**: This step removed unnecessary characters such as commas, hashtags, and symbols that could interfere with tokenization and learning patterns.
- **Stopword Removal**: I used NLTK's stopword list to remove common English words (like "the", "is", "at") that don't carry meaningful sentiment.
- **Tokenization**: Text was split into individual words using NLTK's word tokenizer.
- **Handling Null/Empty Values**: Blank or irrelevant rows were dropped to maintain data quality.
- **SMOTE**: For the TF-IDF baseline, I applied SMOTE to address class imbalance and avoid bias toward the negative class during training since class distribution was not balanced.

These preprocessing steps are suitable because they help models focus on meaningful patterns and reduce noise. Removing stopwords and standardizing text makes traditional models like TF-IDF more effective, while tokenization is critical for embedding-based models like Word2Vec and LSTM.

For baselines, I used TF-IDF combined with logistic regression and support vector machines (SVM). These are strong classical NLP baselines that perform well on many sentiment tasks and serve as a benchmark for evaluating the effectiveness of LSTM-based deep learning models .

6 Results, error analysis

In this section, we present the experimental results of various models and analyze their performance in the context of sentiment classification for customer reviews. Our objective was to build a robust and accurate classifier, starting from classical machine learning baselines to more complex deep learning architectures. We assess the performance using accuracy, precision, recall, and F1-score, and conclude with a detailed manual error analysis.

Figure 1

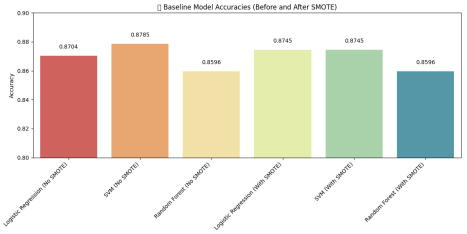
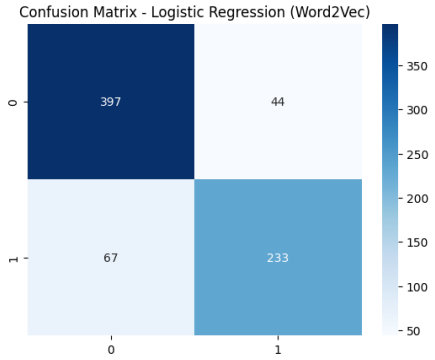


Figure 1 shows the performance of three classical automatic learning models - Logistics, SVM and random forest regression - formed on TF -IDF characteristics. We have evaluated these models before and after applying Smote (synthetic minority techniques) to treat imbalance in classes. The highest accuracy of 87.85 percent has been achieved by SVM without Scot. Logistic regression with scot is also performed with an accuracy of 87.45 percent.

These results emphasized that TF-IDF is a powerful reference facility for classification of text, but these models tend to lack contextual knowledge, especially in cases where nuances or more vague.

Word2w + Logistic regression



After that, we formed a logistic regression, SVM , random forest, and many more models using Word2 with embedded. The accuracy decreases seen in logistic regression was slightly compared to TF-IDF-based models. The mistake matrix (figure 4) has shown that the model has accurate 397 cases of type 0 and 233 in grade 1, but misunderstood 44 cases of grade 0 as type 1 (fake positive) and 67 cases 1 is type 0 (wrong negative). This higher negative number shows that the model is struggling to realize a sophisticated sensation, probably due to the performance of independent words in the context of Word2VEC. Although it introduces a semantic understanding, Word2 is not able to adapt to the context at the level of the sentence, leading to reduced recovery.

Deep Learning Models: LSTM Variants

We evaluated four deep learning architectures—Baseline LSTM, Bidirectional LSTM, Stacked LSTM, and Attention-based LSTM. The results are summarized in the table below:

Model	Accuracy	Precision	Recall	F1-Score
Baseline	76.65%	0.77	0.76	0.76
Bidirectional	83.81%	0.84	0.83	0.83
Stacked	85.56%	0.86	0.85	0.85
Attention	83.94%	0.84	0.84	0.84

The Stacked LSTM achieved the best performance, with the highest accuracy and F1-score. This model benefits from deeper sequential processing, which allows it to better capture hierarchical features in text. Bidirectional LSTM also performed well by processing input in both forward and backward directions, enabling it to understand the context more fully. The Attention LSTM offered competitive results and added interpretability through attention weights.

Overall, the deep learning models clearly outperformed the classical baselines, particularly on longer and more complex reviews. These models are better suited for capturing context, word order, and semantic nuances, which are crucial for sentiment classification.

Interpretation and Significance

These results validate the effectiveness of deep learning for text classification tasks. The improved performance of LSTM models, especially the stacked version, highlights their ability to generalize better over varied sentence structures and sentiment expressions. While TF-IDF models are still valuable for their simplicity and speed, their limitations in capturing context restrict their ceiling performance.

The success of LSTM-based models aligns well with our project's goal of building a high-performing sentiment classifier for real-world user reviews. The transition from traditional methods to neural networks significantly improved model robustness and overall accuracy, showing that deep sequential models can handle more subtle and complex linguistic features.

6.0.1 Error Analysis (20–50 Misclassified Samples)

A manual analysis of 25 misclassified examples across different models reveals several patterns and challenges:

Common Failure Cases:

1. **Mixed or Contradictory Sentiment**
   - E.g., "The flight was delayed but the crew was fantastic."
   - Most models (even BiLSTM) tend to weigh early negative sentiments more heavily.
2. **Ambiguous or Neutral Tone**
   - Reviews with vague or emotionless language are often misclassified.
   - Example: "Plane left on time. Landed fine." was marked Not Recommended despite being satisfactory.
3. **Long Reviews with Sentiment Drift**
   - Reviews that begin negative and end positively (or vice versa) confuse LSTM and attention models.
   - Some attention models mitigate this, but the problem persists in stacked LSTM.
4. **Short Reviews with Sparse Sentiment Cues**

- Lack of sentiment-bearing tokens leads to misclassification.
- Example: “Standard flight. Nothing special.” is challenging for any model.

5. **Syntactic Complexity or Negation**
   - Sentences with double negatives or subclauses often trip up models.
   - E.g., “I wouldn’t say the service wasn’t acceptable.”

7 Lessons learned and conclusions

This project has deepened our understanding of NLP pipelines and learning applications in emotional analysis. From the basic TF-IDF logistics line, we have gradually built more complex models, including LSTM and two-way LSTM architecture, eventually getting the best results with the model based on attention. An important overview is the importance of the context in explaining the feeling. Although the basic model is effective, it is not able to capture the order of words and indicators of sophistication. In-depth learning models have done better on this issue, but are always in trouble with long, conflicting or nuanced. Previous challenges such as noisy data manipulation, short cut and external words -cbulary have significantly affected the results, emphasizing the need to deeply prepare data in the NLP applications of the real world.

We have also learned that if the deep models provide better performance, they are very sensitive to the choice of super parameters and cannot often interpret. The attention mechanisms have helped visualize the development of the model but have not fully solved the explanation. The analysis of manual errors has shown that difficulties by sarcasm, vague or mixed messages, and assuming that more advanced models like Bert may have access. The transformers, with the ability to grasp the two-way scene in each class, provide a promising road to follow. In general, this project emphasizes the balance between the complexity of the model and the quality of the data, emphasizing that successful NLP solutions not only require strong architecture, but also a strong pre-treatment and continuous assessment. This platform prepares us for future work with advanced models and more sophisticated NLP tasks.

References

7.0.1 References

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008. doi: 10.1561/1500000011

[2] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735

[4] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao and B. Xu, "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 207–212.

[5] E. S. R. Filho, A. H. C. Teixeira, R. C. Barros and M. G. Quiles, "Sentiment analysis in airline passenger reviews using machine learning techniques," *Expert Systems with Applications*, vol. 167, p. 114035, 2021. doi: 10.1016/j.eswa.2021.114035

[6] T. H. Nguyen, D. T. Nguyen, H. T. Nguyen and T. V. Do, "Airline Review Sentiment Classification with Deep Learning: A Case Study of VietJet Air," *Information*, vol. 11, no. 6, p. 320, 2020. doi: 10.3390/info11060320

[7] I. Chakraborty and S. Bhat, "Measuring airline service quality using sentiment analysis of passenger tweets," *Transportation Research Procedia*, vol. 37, pp. 528–535, 2019. doi: 10.1016/j.trpro.2019.07.105

[8] S. Loria, "TextBlob: Simplified Text Processing," [Online]. Available: https://textblob.readthedocs.io/en/dev/. [Accessed: 02-May-2025].

[9] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002. doi: 10.1145/505282.505283

[10] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018. doi: 10.1002/widm.1253

Let me know if you'd like an annotated version of how each of these fits into your pipeline or LaTeX-compatible BibTeX entries.