US Airline Twitter – Sentiment Analysis

Authors:
Prabhu Shankar
Rajivni Tiruveedhula

OBJECTIVE

Our goal is to conduct sentiment analysis on tweets related to U.S. airlines by utilizing data and dividing it into training and testing sets to ensure comprehensive insights. By applying natural language processing techniques, this project aims to uncover trends and patterns in customer sentiments expressed through tweets. The findings will provide valuable insights to help companies make informed decisions while considering customer opinions.

DATASET DESCRIPTION

The dataset, sourced from Kaggle, offers insights into customer tweets about various airlines. It includes sentiment labels, categorized as either positive or negative, which are used for natural language processing (NLP). The dataset comprises approximately 16,000 records, with each entry representing a tweet about an airline.

EXPERIMENT PROCESSING

Text Processing:
The Process Document from Files operator was employed to import text files containing sentiment-related words categorized as positive, negative, or neutral. The following steps were then performed:

1. Tokenization: Words were split into individual units called tokens, and a function was applied to remove stop words that did not contribute to the analysis.
2. Case Transformation: Words were standardized to a consistent case for uniformity across the dataset.
3. Stemming: Words were reduced to their root forms to generalize and simplify the analysis.

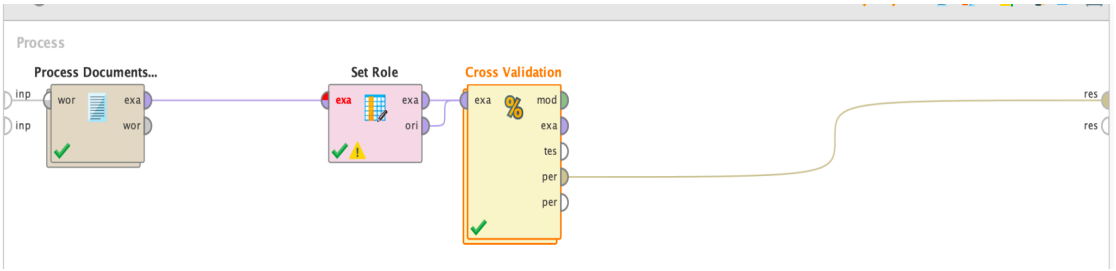
Data Partition: R was implemented to partition the dataset into training and testing datasets with 80% and 20% weightage respectively.

Model Construction:
Text Vectorization: The text data was transformed into numerical representations, making it suitable for processing by the model.
Machine Learning Algorithm: A supervised learning algorithm was employed to train the sentiment analysis model effectively.
Hyperparameter Tuning: The model's parameters were fine-tuned to achieve optimal performance and accurate results.

Model Evaluation:
Cross-Validation: K-fold cross-validation was applied to evaluate the model's performance and ensure its ability to generalize effectively across different subsets of the data.
Confusion Matrix: Used to compare predicted sentiment labels with actual labels, offering insights into the model's classification accuracy and potential areas of improvement.

Evaluation Metrics: Key metrics such as precision, recall, F1 score, and accuracy were calculated to comprehensively assess the model's performance.
Fine-tuning: The model was refined based on the evaluation results to improve its accuracy and overall effectiveness.

EXPERIMENT PROCESS:



RESULTS:

KNN METHOD FOR DIFFERENT VALUES OF K

K = 5

accuracy: 84.22% +/- 16.65% (micro average: 84.26%)

	true neg	true pos	class precision
pred. neg	2255	635	78.03%
pred. pos	109	1728	94.07%
class recall	95.39%	73.13%	

K = 200

accuracy: 79.21% +/- 17.82% (micro average: 79.20%)

	true neg	true pos	class precision
pred. neg	2184	803	73.12%
pred. pos	180	1560	89.66%
class recall	92.39%	66.02%	

K = 1000

accuracy: 83.16% +/- 16.19% (micro average: 83.18%)

	true neg	true pos	class precision
pred. neg	2314	745	75.65%
pred. pos	50	1618	97.00%
class recall	97.88%	68.47%	

pred. pos

ASSESSING ACCURACY USING KNN METHOD FOR DIFFERENT VALUES OF NUMERICAL MEASURES WITH K = 200

Folds = 10

accuracy: 71.61% +/- 1.48% (micro average: 71.61%)

	true neg	true pos	class precision
pred. neg	2265	1243	64.57%
pred. pos	99	1120	91.88%
class recall	95.81%	47.40%	

Folds=100

accuracy: 72.76% +/- 4.70% (micro average: 72.75%)

	true neg	true pos	class precision
pred. neg	2266	1190	65.57%
pred. pos	98	1173	92.29%
class recall	95.85%	49.64%	

Folds = 1000

accuracy: 72.89% +/- 18.41% (micro average: 72.88%)

	true neg	true pos	class precision
pred. neg	2267	1185	65.67%
pred. pos	97	1178	92.39%
class recall	95.90%	49.85%	

ASSESSING ACCURACY USING K-NN METHOD FOR DIFFERENT NUMBER OF FOLDS WITH K = 200

Cosine similarity:

accuracy: 85.43% +/- 16.53% (micro average: 85.45%)

	true neg	true pos	class precision
pred. neg	2086	410	83.57%
pred. pos	278	1953	87.54%
class recall	88.24%	82.65%	

Manhattan Distance:

accuracy: 72.89% +/- 18.41% (micro average: 72.88%)

	true neg	true pos	class precision
pred. neg	2267	1185	65.67%
pred. pos	97	1178	92.39%
class recall	95.90%	49.85%	

pred. pos

Euclidean Distance:

accuracy: 79.21% +/- 17.82% (micro average: 79.20%)

	true neg	true pos	class precision
pred. neg	2184	803	73.12%
pred. pos	180	1560	89.66%
class recall	92.39%	66.02%	

CONCLUSION AND INTERPRETATION:

As we increase the value of K the model will become more accurate for local data. If we try smaller K values the accuracy decreases, especially among the classes and the model will become more prone to overfitting challenges if the data size increases. With greater K values the predictions become more accurate, and the importance of individual data points diminishes making the predictions more stable and better for handling unseen data also increasing the value of k smoothens the decision boundaries. If the k is too large the model might oversimplify the decision boundaries and might not be able to capture the subtle differences of different sentiments.

Numerical Measures:
The following measures were utilised in this project:
• Cosine Similarity
• Manhattan distance
• Euclidean distance

How Businesses Benefit from Sentiment Analysis

1. **Customer Insights**: Understand customer satisfaction, preferences, and pain points.
2. **Brand Monitoring**: Track public perception and respond to trends or crises.
3. **Customer Support**: Address negative sentiments quickly to enhance service.
4. **Market Research**: Gauge consumer reactions to products or campaigns.
5. **Competitive Analysis**: Identify competitors' strengths and weaknesses.
6. **Personalized Marketing**: Target satisfied customers for upselling or advocacy.
7. **Product Improvement**: Highlight areas for enhancement and innovation.
8. **Predictive Insights**: Forecast customer behavior and market trends.