University of Stavanger

Artificial Intelligence for Engineers

DAT305-1 24H

Course: DAT305	Autumn semester, 2024 Open/ Restricted access
Writer: Group number: 5	..... (Writer's signature)
Subject supervisor:	Mina Farmanbar
Assignment 1:	Artificial Intelligence for Engineers.
Credits (ECTS): 5.	
Keywords: Supervised Learning, NLP,RNN.	Stavanger, 26 November 2024

Contents

Sentiment Analysis of US Airline Tweets

GODSPOWER OGAGA UTI*, University of Stavanger, Norway

The goal of this project is to classify tweets related to several major US airlines into positive, neutral, or negative sentiments. The dataset used is the US Airline Sentiment Dataset, containing 14,640 tweets labeled as positive, neutral, or negative. Tweets are from a period of ten days in February 2015. A sentiment analysis of airline tweets may yield insights on customer satisfaction, operational issues and brand perception.

CCS Concepts: • **Computing methodologies** → **Supervised learning by classification**; **Information extraction**.

Godspower Ogaga Uti. 2024. Sentiment Analysis of US Airline Tweets. 1, 1 (November 2024), 5 pages. https://github.com/elyte5star/Text-Classification

1 Introduction

As of 2024, the U.S. airline industry is dominated by four major carriers: American Air- lines, Delta Air Lines, United Airlines and Southwest Airlines. These airlines collectively dominate the U.S. market, offering extensive domestic and international networks. American Airlines is the largest and offer thousands of flights daily to more than 350 destinations in more than 60 countries. American has hubs in Charlotte, Chicago, Dallas-Fort Worth, Los Angeles, Miami, New York, Philadelphia, Phoenix and Washington, D.C.[Group 2024]

2 Data Wrangling

The following tools were used to conduct data analysis in this project:

- `Pandas`: It supports operations like sub-setting of large datasets, handling of missing data, reshaping of a dataset, reading and writing data in various formats like CSV.
- `NumPy`: It supports fast numeric computation and the manipulation of complex tensors.
- `Matplotlib`: This library supports visualization of the dataset.
- `Seaborn`:It is built on top of Matplotlib. It provides uni-variate and bi-variate graph in this project.

2.1 Data Analysis

The dataset contains `14640` tweets about six major air carriers, namely:

- `Virgin America`
- `United`
- `Delta`
- `US Airways`
- `American`
- `Southwest`

Furthermore, it contains 15 columns, namely:

- `tweet_id`
- `airline_sentiment`
- `airline_sentiment_confidence`
- `negativereason`
- `negativereason_confidence`
- `airline`
- `airline_sentiment_gold`
- `name`
- `negativereason_gold`
- `retweetcount`
- `text`
- `tweet_coord`
- `tweet_created`
- `tweet_location`
- `user_timezone`

Table 1. Frequency of Tweets

Airline Name	Negative sentiment freq.	Neutral sentiment freq	Positive sentiment freq.
American	1960	463	336
Delta	995	723	544
Southwest	1186	664	570
US Airways	2263	381	269
United	2633	697	492
Virgin A.	181	171	152
Sum	9178	3099	2363

Table 1 shows the distribution of the tweets grouped by sentiments for the various airlines. Statistics show an unbalanced distribution of the tweets, with negative sentiments about three times higher than positive and neutral sentiments.
Virgin America has the least negative sentiments, while United Airlines and US Airways have the highest. Further research shows sentiment confidence of 0.65 for 25% of the dataset. Customer service issues and late flights account for the majority of reasons for complaints amongst the airlines. The two top authors are "`JetBlueNew`" and "`kbosspotter`", and their tweets were regarding Delta Airlines, which means they are frequent flyers. Figure 1 shows the top 10 tweet authors in the dataset.
Figure 2 shows the top 10 reasons for customer complaints in the dataset.

2.2 Data Preprocessing

The tweets went through the following stages;

- `Cleaning`: Removal of links, tags, and emoticons with the help of regex and creating features from them. Furthermore,

Author's Contact Information: Godspower Ogaga Uti, go.uti@stud.uis.no, University of Stavanger, Stavanger, Norway.

ACM XXXX-XXXX/2024/11-ART
https://github.com/elyte5star/Text-Classification

Top 10 Authors

	top_tweet_authors	count
0	JetBlueNews	63
1	kbosspotter	32
2	_mhertz	29
3	otisday	28
4	throthra	27
5	rossj987	23
6	weezerandburnie	23
7	MeeestarCoke	22
8	GREATNESSEOA	22
9	scoobydoo9749	21

Fig. 1. Top 10 tweet authors

Top 10 complain

	top_complain_reason	count
0	Customer Service Issue	2910
1	Late Flight	1665
2	Can't Tell	1190
3	Cancelled Flight	847
4	Lost Luggage	724
5	Bad Flight	580
6	Flight Booking Problems	529
7	Flight Attendant Complaints	481
8	longlines	178
9	Damaged Luggage	74

Fig. 2. Top reasons for complain

the date column `tweet_created` was decomposed into weekly day, day, and hour. Month and year were not taken into consideration since all the tweets are from the 16th of February 2015 to the 24th of February 2015.

- `Tokenization`: Splitting tweets into individual words.
- `Stop word removal`: Removing common words with less importance to sentences.
- `Stemming`: Reduce a word to its root.

3 Model selection and implementation

Columns with a lot of missing data were dropped and an extra column called `target` was created for the dataset. The `target` column is an integer encoding of the sentiment classes. With negative, positive and neutral classes mapped to 0, 1, and 2 respectively. Furthermore, the dataset is split into training and test sets in a ratio of 80 to 20 and the duplicates in the training set were dropped.

3.1 Bag of Words (BoW)

3.1.1 Baseline Model. The baseline model is a simple Linear Support vector classifier that works well with sparse text matrices. The algorithm is provided by `scikitlearn`. Since it is an unbalanced classification task, the weight parameter was set to 'balanced'. This tells the algorithm to calculate class weights and attach more weight to the minority classes (positive and neutral) automatically. The text was transformed into Bags of Words (Bow) using `TfidfVectorizer` from `scikitlearn`. Putting words in bags makes the words lose their order relationship. As a fix, the **n-gram** parameter is specified. N-grams associate adjacent tokens together. Finally, grid search is used to fine-tune the hyper-parameters for the support vector machine. The baseline model achieved a weighted f1 score of **0.785**.

3.2 Word Embeddings

Word embeddings use semantic similarity to convert words into values in a matrix.

3.2.1 Model 1 Word2vec. A word vector is created by converting the sparse matrix created from the `TfidfVectorizer` into a dense one. The longest sequence of words is taken into consideration. The length of the longest sequence is used to pad the others by using the `pad_sequence` utility from the `Keras` library. For this model, a randomly chosen embedding vector length of 32 is used. Finally, it is fed into a Recurrent Neural Network. Since it is a multi-classification task, the `softmax` activation function is deployed on the final layer of the network. The loss function is Categorical cross entropy, since

the target (y) is one-hot encoded. The optimizer is `Adam` which is generally good for text analysis. The confusion matrix shows that the model learned more of the majority class (negative) but struggled to predict the minority classes. The f1 score dropped.

3.2.2 Model 2 Word2vec + Class Weights. This model is similar to the first model, but class weights were introduced. The classifier adds more weights to the less-represented classes.

3.2.3 Model 3 Pre-Trained GloVe Embedding + Class Weights. GloVe (Global Vectors for Word Representation) is a popular pretrained technique that uses matrix factorization to transform a sparse matrix of word-to-word co-occurrence into a dense matrix [Jeffrey Pennington 2015]. It is developed as an open source project at Stanford University. The selected embedded dimension is 100.

3.3 Final Model (Model 3 + Many Features)

The previous models with class weights did not learn much, because increasing the weights does not increase the information from the dataset but adds more bias to the output. To help the algorithm learn more, new features were introduced. This model takes two inputs for X, the embedded and the new features. The results show improvements from the previous RNN models.

3.4 Experiments

3.4.1 Upsampling of the miniority classes using numpy. This technique involves oversampling of minority classes (positive and neutral) by replicating its examples until minority classes have the same examples as the majority class (negative).

3.5 Model comparison and Evaluation

	classifier-name	f1_score	precision	recall	accuracy	Average-method
0	LinearSVC + Weights	0.785	0.783	0.789	0.789	weighted
1	RNN-Word2Vec	0.585	0.647	0.557	0.713	macro
2	RNN-Word2Vec + Weights	0.649	0.657	0.643	0.643	weighted
3	RNN-Glove + Weights	0.735	0.764	0.723	0.723	weighted
4	RNN-Glove + Weights + other features	0.769	0.787	0.761	0.761	weighted
5	RNN-Glove + Upsampling	0.674	0.657	0.701	0.735	macro

Fig. 3. Top 10 tweet authors

Figure 3 shows the metrics calculated from all models. The baseline model outperformed the ANN models due to small training samples. A better approach would be the introduction of pre-trained networks, for example, the Google Bidirectional Encoder Representations from Transformers(BERT). However, this project is focused on getting as much information as possible from the provided dataset. Furthermore, the addition of new features proved to be more productive than adding weights and oversampling the minority classes. The implementation of this project can be found on `GitHub`.

4 Reflections and Conclusion

The results from the data analysis can help US airlines solve some of their customer issues. Firstly, they would need to address their customer service department, then their late flights. Finally, the statistics from the dataset show more complaints on Sundays and Mondays. Delta Airlines should provide a better program for their frequent flyers.

Acknowledgments

To Shaima Ahmad Freja for the mentorship.

References

American Airlines Group. 2024. *American Airlines Group Customer service.* Retrieved November 2, 2024 from https://www.aa.com/i18n/customer-service/about-us/american-airlines-group.jsp

Christopher D. Manning Jeffrey Pennington, Richard Socher. 2015. *Global Vectors for Word Representation.* Retrieved November 25, 2024 from https://nlp.stanford.edu/projects/glove/

Received 26 November 2024