US Airline Tweets Sentiment Analysis

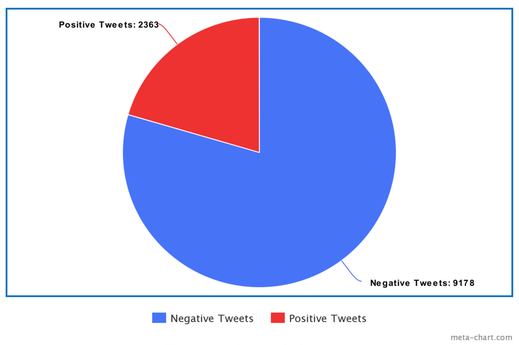
Efecan Demirkıran, Fatma Tuğçe Akgül, Sena Nur Yılmaz, Oğuzhan Derebaşı, Hasan Güzelmansur
Machine Learning, Türkisch-Deutsche Universität

Abstrakt

Sentimentsanalyse von Tweets, die von einer Fluggesellschaft (Virgin America) empfangen werden. Tweets werden als positiv, negativ und neutral klassifiziert. Wir trainieren ein Modell für positive und negative Klassifikation. Das Modell, das wir für die positive und negative Klassifikation verwenden, ist Multinominal Bayes.

Über die Daten

Der ausgewählte Datensatz enthält eine Sentimentanalyse der Probleme jeder großen US-Fluggesellschaft. Es ist ein Datensatz auf Kaggle, der Tweets enthält, die im Februar 2015 mit dem Namen "Twitter US Airline Sentiment" veröffentlicht wurden. Tweetsdatensatz besteht aus insgesamt 15 Spalten, wie Tweet_Id, Airline_sentiment, text usw. Die Gesamtzahl der positiven, neutralen und negativen Tweets beträgt 14640. Da beim Modelltraining positive und negative Tweets verwendet werden, wurden neutrale Tweets entfernt, sodass 9178 negative und 2363 positive Tweets übrig blieben. Dann werden mit der Bibliothek "nltk.corpus" Funktionen geschrieben, die Wortarten (Substantive, Verben usw.) und Satzendemarkierungen erkennen.



Datenvisualisierung

Schritten

1. Datensammlung
2. Datenvorbereitung
3. Datenbereinigung
4. Modell-Training & Hyperparameter Einstellung
5. Ergebnisse Interpretieren

Datensammlung

"Positiv", "negativ", "neutral" gekennzeichnete Tweets aus der Twitter-Konto der Fluggesellschaft *Virgin America*.
https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Datenvorbereitung & Datenbereinigung

Tokenisierung
Lemmatisierung
Bereinigung von Stoppwörtern und Satzzeichen

Modell-Training: Multinomial Naive Bayes

MNB versucht, das Label des angegebenen Datensatzes zu finden. In diesem Beispiel besteht das Ziel darin, die negativen und positiven Tweets zu klassifizieren. Es berechnet die Zugehörigkeit jedes Textes, nämlich Tweets, zu diesen beiden Klassen über die Wahrscheinlichkeit und ordnet ihn der Klasse mit der höchsten Wahrscheinlichkeit zu.

Für Hyperparameter-Tuning:
1. die verschiedenen Alpha-Werte ausprobieren und das beste Alpha finden
2. Ausprobierung der verschiedenen Vektorisierungen

Unter den getesteten Alpha-Werten wird derjenige mit der größten Fläche unter der ROC-Kurve ausgewählt.

```
: mulnb(bow_train,bow_cv,Y_tr,Y_cv) # mod
  # en iyi roc_auc değerleri 1 ve 0.005 a
100  ------>  0.8119084861986697
50  ------>  0.8268131856556752
10  ------>  0.8584027218193213
5  ------>  0.8723251857820754
1  ------>  0.9040911556672092
0.5  ------>  0.9170182174400783
0.1  ------>  0.9373238954459032
0.05  ------>  0.9395411669905674
0.01  ------>  0.9349933903187921
0.005  ------>  0.9311724152459803
```

Beste Alpha-Parameter: 0.05

Ergebnisse

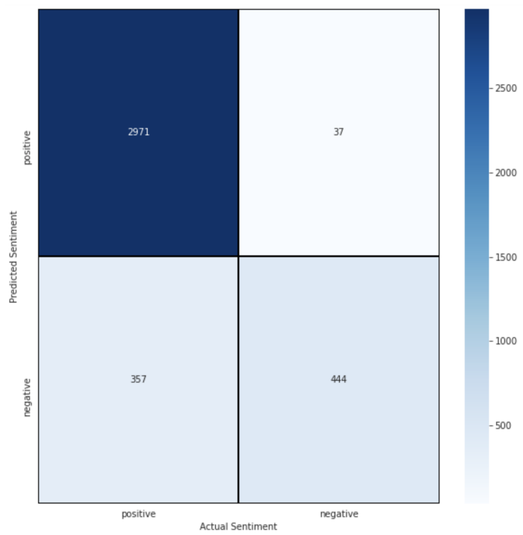
BOW- und TF-IDF-Vektorisierungsverfahren wurden ausprobiert und ähnliche Ergebnisse wurden erhalten. Die Genauigkeit ist für beide 90%.


```
f1 score for data : 0.6926677067082684
hamming loss for data: 0.1034392228931478
Precision recall report for data:
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      3008
           1       0.92      0.55      0.69       801

    accuracy                           0.90      3809
   macro avg       0.91      0.77      0.82      3809
weighted avg       0.90      0.90      0.89      3809
```

Die Ergebnisse aus BOW und TF-IDF Vektorisierungen



Die Konfusionsmatrix aus BOW und TF-IDF Vektorisierungen: 394 falsche Klassifizierungen aus 3809 Datensätze.

Bei den meisten Tweets handelte es sich gemäß den TF-IDF-Ergebnissen um Erstattungsanträge für ihre Flugerfahrungen:

	feature	tfidf
0	asked reimbursement	0.272420
1	reimbursement something	0.272420
2	added account	0.272420
3	like mile	0.272420
4	mile added	0.272420
5	account told	0.260198
6	usairways asked	0.251525
7	something like	0.251525
8	reimbursement	0.244798
9	added	0.223903

Die am häufigsten genannten Themen in Tweets mit TF-IDF-Vektorisierung.

Fazit

	Category	AUC Score	F1-Score
0	BOW Vectorization	0.95	0.69
1	TF-IDF Vectorization	0.95	0.65

Die Vektorisierungsmethoden liefern die gleichen AUC-Werte, aber basierend auf den F1-Werten kann man die BOW-Vektorisierung bevorzugen. Denn der Klassifikator erhält nur dann einen hohen F1-Wert, wenn sowohl Recall als auch Präzision hoch sind.

https://github.com/roroquentin/MLProjekt_INF502/blob/main/ML_Sentiment_Analysis%20(1).ipynb