Twitter-based Opinion Mining for Flight Service Utilizing Machine Learning

Prayag Tiwari
Department of Information Engineering, University of Padova, Italy
E-mail: prayagforms@gmail.com

Hari Mohan Pandey
Department of Computer Science, Edge Hill University, Ormskirk, UK
E-mail: pandeyh@edgehill.ac.uk

Aditya Khamparia
School of Computer Science and Engineering, Lovely Professional University, Phagwara, India
E-mail: aditya.khamparia88@gmail.com

Sachin Kumar
Department of System Programming, South Ural State University, Chelyabinsk, Russia
E-mail: sachinagnihotri16@gmail.com

Keywords: sentiment analysis, random forest, logistic regression

Received: December 12, 2018

Twitter is one of the most prominent social networking platforms so far. Millions of users utilize Twitter to share their thoughts and views on various topics of interest every day resulting a huge amount of data. This data could be considered to have a rich source of useful information hidden inside. Using machine learning to this data may give rise to effective recommender frameworks for individuals to manage their lives in a much more convenient way. In this paper, we propose a machine learning approach to classify the passenger's tweets regarding the airplane services to understand the pattern of emotions. We adopt Random Forest (RF) and Logistic Regression (LR) to classify each tweet into positive, negative and neutral sentiment. The evaluation of the collected real data demonstrates that these two methods are able to achieve an accuracy ≈80%.

Povzetek: Z metodami strojnega učenja so analiziranji tviti (čivki) letalskih potnikov o letalskih storitvah.

1 Introduction

At present, large scale companies are investing plenty of time, resources and energy to enhance the consumer's loyalty. It may explore more opportunities for the interaction between companies and consumers to get their feedback and suggestion about the products and services with an aspect of customer satisfaction and product quality improvement. This may increase the both the economic and social development of the company. A crucial but challenging step is to automatically analyze the customer feedback by extracting useful information from the huge data of customer feedbacks [1]. Customer feedback data is very important in addressing several issues and sentiment and opinion analysis is one of the important issue among them. Extracted patterns from the data may be utilized by company experts to understand the polarity of the opinion towards different products and services. In general the polarity of opinion may be positive, negative or neutral. Companies may use these polarity of opinions in order to improve their quality of products and/or services.

Sentiment analysis/opinion mining assists in answering different question about products and services by understanding the emotions in the feedbacks [2]. Present world is utilizing the natural language processing (NLP) and text classification techniques to map the sentiments within the text into positive, negative and neutral classes [3].

The sentiments can be seen as an indirect publicity of a company's products and services in the world that provide a direct impact on other customer's. For travelers, the most popular and convenient platform for sharing their opinion is Twitter [4]. Each travel journey on different carriers may bring different comfort levels i.e. good, average or poor level of comfort. These comfort levels are conveyed to the social media i.e. Twitter etc. by the travelers in terms of tweets. If a traveler enjoyed the trip, the respective tweet would demonstrate the happiness or positive emotions towards the travel carrier otherwise negative emotions may be conveyed. Figure 1 depicts a furious tweet by a passenger on British Airways flight. As a result, the company considered it very urgent and important and settled the issue at the earliest. In another scenario (Figure 2), a sarcasm tweet for Indigo Airways

was fired because the baggage of passenger was transferred to a different location (Hyderabad) other than the traveler (Calcutta). The tweet in figure 2 seem to be negative from a human perspective whereas it is difficult to put this into negative class for the machine because of the complex words used in the tweet. Also, tweets on tweets may not contain more than 140 characters at once. Therefore, it is useless to expect the detailed information inside the tweet. However, a general understanding about polarity of emotions can be developed using machine learning methods. Further, tweets in the categories may be analyzed to get insights or possible reasons for these sentiments [5].

Every day more than a million of people are travelling around the world and tweeting their views with respect to the journey. It results in a huge amount of data available for analysis every day. Hence, machine learning techniques can be considered as a solution for such analysis. Machine learning techniques are efficient to handle huge data with large dimensions [6, 7].

Figure 1: A negative tweet illustrating loss of luggage.

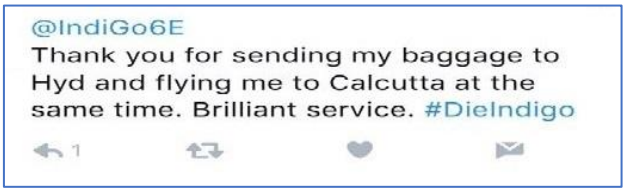

Figure 2: A tweet illustrating wrong transfer of luggage in sarcastic way.

The main motivation behind this work is to provide a better analysis for classification of sentiment from the tweet data in order to assist the airline companies to improve customer satisfaction and improve the quality of service. The organization of the paper is as follows: Section 2 provides the state of art literature review. In section 3, proposed work is discussed. Section 4 presents the experimental results and discussion which is followed by conclusion in section 5.

2 Literature review

Kusen et al. [8] analyzed a twitter data set consisting of 343645 tweets about 2016 Austrian presidential election. This analysis amalgamated approaches from sentiment analysis, network science, and bot detection. It was shown that the immediate relationship between the winners of the 2016 Austrian presidential races was more famous and had a high impact on Twitter than other rivals.

Ahmed et al. [9] have demonstrated how the first time twitter utilized as a campaign tool in the Indian election 2014 by different parties. They demonstrated computer-aided and multi-level manual analysis of 98363 tweet messages by 11 parties during the campaign. It had a high impact on twitter of winning party than other parties.

Stigleitz et al. [10] examined whether opinion persisting in online networking content is related to a client's data sharing coordination. They conducted an examination with regards to political correspondence on Twitter. On the basis of two dataset collections of about 165,000 tweets altogether, they found out that candidly charged Twitter messages had a tendency to be retweeted all the more regularly and more immediately contrasted with biased ones. As a general suggestion, organizations should give careful consideration to the examination of opinion identified with their brands and items in social networking correspondence, in addition to planning promoting content that triggers emotions.

Gunarathne et al. [11] investigated the objection resolution experience of passengers of U.S. aircraft, by utilizing an interesting data collection amalgamating both customers– brand cooperation's on Twitter and how clients felt toward the end of these associations. They located that objection Customer who is more dominant in online networking communities will probably be fulfilled. Customers who have beforehand objection to the brand via social networking media and customers who grumble about process-related instead of result related issues are less inclined to feel better at last. To the best of our insight, this examination is the first to recognize the key factors that shape client sentiments toward their brand– client communications via social networking media. Their outcomes give useful direction to effectively settling clients' objection using social networking field that expects exponential development in the coming decade.

Seunghyun et al. [12] showed social networking examination utilizing Twitter data alluding to cruise travel. This examination likewise incorporated an inside and out an investigation on tweets by three kinds of group users: private, commercial and blogs. The outcomes demonstrated that not exclusively were words identified with industry, travel, emotions, and destination most often utilized as a part of organizing tweets, but also proficient bloggers, cruise lines, celebrities and travel organizations really drove significant subgroups on cruise themes on Twitter. On the basis of such outcomes, this examination gives attainable marketing approach.

3 Proposed work

In this section, our proposed model consists of several steps like preprocessing, feature extraction etc. in order to train the model and use the test dataset to check the evaluation metric on the test dataset. Precision, F1-measure, and Recall are used as an evaluation metric.

3.1 System architecture

Proposed architecture can be seen in the figure no. 3 that how flow started of our model from the dataset, text pre-processing, feature extraction, a division of dataset into training and testing set, the trained model then tested on the test dataset.

3.2 Text preprocessing

As a pre-processing step, we do a basic statistical analysis on the collected data. The statistics include the number of words (denoted as word_counts), the number of hashtags (denoted as hashtag_counts), and counts for other punctuation marks.

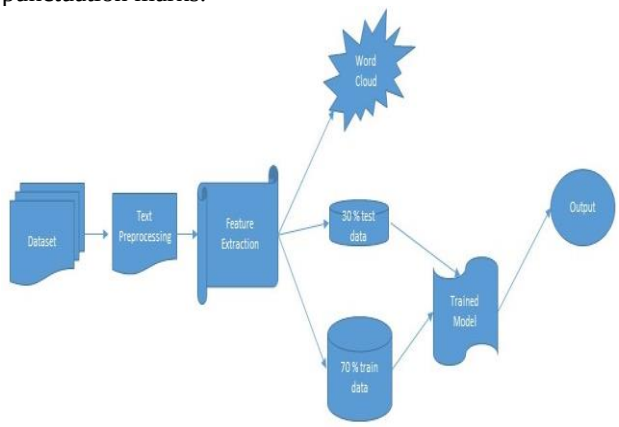

Figure 3: Architecture of Proposed Sentiment Analysis Model.

The distribution of those textual variables over the three sentiment classes is shown in Fig 4.

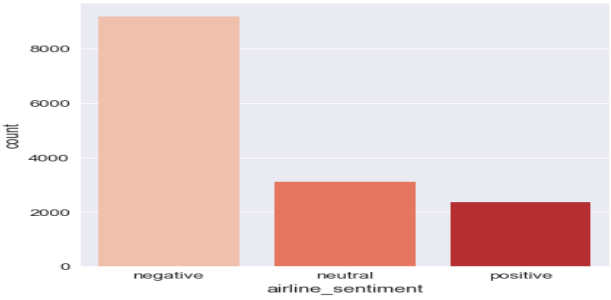

Figure 4: Distribution of Class Labels

We then remove the hashtags, mentions, URLs etc= to make text data more clean for further analysis. We also removed punctuations, stop words and digits. Finally, we stem words and convert them to lowercase. This is the standard procedure for pre-processing textual data. The examples of tweets after pre-processing can be seen in Figure 5.

```
13892    pilot told us would release bags well offer hotel vouchers neither happened
8817     always happy help
6270     wanted southwest know think great used anymore nothing look
420      helping step game tindertips tinderchamp
7170     no come
7529     landing usual great flight wiyh great crew hello sunny west palm beach jetbluerocks
10131    please thank mellie cae tammy baggage claim clt excellent customer service day complaint
609      worstunitedflightsever ua iad las mechanical problems switched aircraft delayed hours
12166    sounds like date
4232     next flights miami another airline
Name: text, dtype: object
```

Figure 5: A sample of preprocessed tweets.

3.3 Random forest

Decision Trees are the most widely used machine learning methods. Random Forest provides an effective way of averaging several decision trees, trained in different segments of the same training dataset with the aim to deteriorate the variance and provide a stable and accurate prediction. Random forest could be an ensemble learning procedure for regression, classification, and elective undertakings, which is achieved by building a large group of decision trees at training phase and provoking the classes which are the model for the mean prediction (regression) or classifications (classes) of the distinctive trees. In a distinct computation, classification is implemented recursively until every leaf is pure. The aim is to dynamically predict the best decision tree until it catches up the adaptability, precision, and balance. There are three measures to split the node are shown in Eq. 1-3.

$$Entropy = \sum_j P_j \log_2 P_j \quad (1)$$

$$Gini = 1 - \sum_j P_j^2 \quad (2)$$

$$Classification\ Error = 1 - max\ P_j \quad (3)$$

Where P_j is the probability of class j.

The algorithms starts as follows: we pick a bootstrap observation from the S in which $S^{(i)}$ represents the i^{th} bootstraps for every tree in the given forest. Then train the decision tree utilizing a revised decision tree algorithm. The revised decision tree algorithms as follows: in contrast of analyzing all feasible feature split, some random features f ⊆ F, at every node of the tree where F is the feature sets. The given node split on the top features in f comparably than selecting F. In this, f is much more compact and smaller than F. The most challenging task is to choosing on which feature to split in the decision tree learning that is why making narrow the feature set makes faster learning. The pseudocode is given as follows:

Algorithm Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

3.4 Logistic regression

Logistic Regression is a statistical method for investigating a dataset in which there are at least one or more than one independent variables that decide a result. The result is estimated with a dichotomous variable (in which there are just two conceivable results). The objective of logistic regression is to locate the best fitting model to depict the connection between the dichotomous feature and the set of independent factors. Our Hypothesis function can be written like as given below,

$$Y = W^T X \quad (4)$$

A sigmoid function is implemented across the notable hypothesis function to keep into the range of (0, 1). The sigmoid function can be described as,

$$sg(y) = 1/(1+e^{-y}) \quad (5)$$

So our new hypothesis is

$$sg(y) = sg(W^T X) = 1/(1+e^{-WTX}) \quad (6)$$

Boundary Estimation:

Our new hypothesis function provides us the values in between 0 and 1so it can be clarified probability of y would be 1 for given X and this can be written in this form,

$$sg(y) = P(y = 1|x, W) \quad (7)$$

Cost Function:

Taking a square error function does not work from the transformed hypothesis function so we make a new form of cost function which is as follows:

$$E(sg(W, x), y) = -log(1-sg(W, x)) \text{ if } y = 0$$
$$E(sg(W, x), y) = -log(sg(W, x)) \text{ if } y = 1$$

Therefore, the mean of cost function will be as follows,

$$H(W) = \frac{1}{m}\sum_{i=1}^{m} E(sg(W, x\,i), yi) \quad (8)$$

Parameter Estimation:

We utilize an iterative approach known as Gradient Descent to enhance the parameters across every step and reduce the cost function to the most feasible value. Gradient Descent requires a convex cost function to avoid getting stuck in a local minimum at the optimization stage. We begin with irregular parameter values and update their values at every stage to reduce the cost function to some extent until we reach the lowest point or equivalently there are not any changes to the value of the target function. The gradient descent step is as follows,

$$\beta_{(i+1)} = \beta_i - p\,\frac{\delta H(W)}{\delta \beta i} \quad (9)$$

For every i =1, 2, 3…, n and p is the learning rate controlling the speed that it moves across the slope on the curve to reduce the cost function.

Above process can be shown in the pseudocode for logistic regression with L1 regularization. The procedure starts with providing input dataset D with corresponding labels and iteration numbers. In this, w_h is the temporary variable. Our algorithm start working as mentioned in the pseudocode.

Algorithm Logistic Regression with L1 regularization

```
procedure StochasticGradientDescent(D, Labels, Iter)
Input: Dataset D, Labels of Dataset, Iteration num
Output: optimal weight of logistic regression
    w ← [1, 1, ..., 1]
    Initialize q_i with zero for all i
    for k = 1 → Iter do
        chooseData = D
        for i = 1 → m do
            γ ← Learning Rate
            λ ← Regularization Lambda
            u = u + γλ
            Select a index of chooseData idx randomly
            x ← chooseData[idx]
            del chooseData[idx]
            for i ∈ featuresinsamplex do
                w_i = w_i − γ ∂loss(w,x_i)/∂w
                wh ← w
                if w_i > 0 then
                    w_i ← max(0, w_i − (u + q_i))
                else if w_i < 0 then
                    w_i ← min(0, w_i + (u − q_i))
                end if
                q_i ← q_i + (w_i − wh)
            end for
        end for
    end for
```

3.5 Evaluation metric

In order to measure the accuracy of classification [13], we used different parameters such as Recall, Precision, and F-measure [12]. Recall can be regarded as the measure of completeness whereas Precision can be seen as a measure of exactness. Formally, precision can be defined as the ratio of correctly classified instances of one class and a total number of instances classified in the same class, whereas recall is the ratio of correctly classified instances of one class and overall instances of the same class. Both precision and Recall can be calculated using the confusion matrix. Confusion matrix represents the number of correctly classified and incorrectly classified instances of all classes. Using the confusion matrix, all performance evaluation measures can be calculated. For a twitter dataset with a binary classification problem, if the total 600 tweets are classified to one class, among which 500 of them are correctly classified, and the total number of tweets in this class are 700. Then, the precision of the classifier is 500/600= 83.3%, and the recall of the classifier is 500/700=71.4%. The Recall and Precision are integrated to develop a new measure known as F-measure or F-score. The formula to calculate F-measure is given in Equation 12.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (10)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (11)$$

$$\text{F-measure}=2\left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (12)$$

Where TP is True Positive, TN is True Negative, FN is False Negative and FP is False Positive.

4 Experiments

4.1 Data preparation

In this study, we experiment on the US Airlines 2016, which contains 14500 passenger tweets. Since the number of original features is too large, we manually select the textual based features, because are easily accessed by passengers. As can be seen from Figure 6, the class labels are highly unbalanced. The dataset is available for public use [14]. After the preprocessing step, we identified the top 30 frequent words in the dataset, which is shown in Figure 6.

4.2 Experimental analysis

For further evaluation, it is necessary to have test data that could be helpful to evaluate several measures of our model. Data was divided into 70 percent train and 30 percent test set Text count variable has been combined with cleaned data to create a data frame.

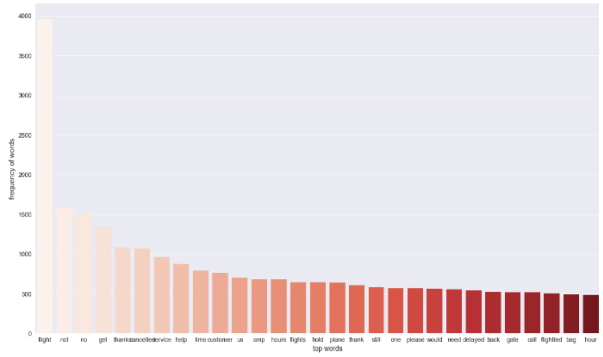

Figure 6: Top 30 most frequent words.

Figure 7: Distribution of Text Variables.

For opting better parameters, it is needed to assess on a different validation from training. By utilizing just a single validation set one might not deliver reliable validation result. To get a more precise estimation, cross-validation is performed.

In this study, we conduct k-fold validation on the data at hand and utilize GridSearchCV to search for the best-performed parameter combination. We select precision as the metric for optimization for both logistic regression and Random Forest classifiers. In order for bag-of-word features to be properly fed into classifiers, we use CountVectorizer to transform words into vectors. The word cloud in Figure 10 gives a decent visual depiction of the word recurrence for each kind of opinion, in which the left one corresponds to the positive opinion and the right one the negative. The span of the word relates to its recurrence across all tweets.

This figure gives us a rough idea of what passengers are discussing. For instance, for negative opinion, passengers appear to gripe about delayed of flight, cancellation of flights, the low-quality of the flight service, the hours holding up and etc. Be that as it may, for positive opinion, passengers are thankful and they discuss extraordinary administration/flight. A cloud of the word has been mentioned in Figure 10 to visualize those positive and negative tweets more properly.

Several other approaches have been used but Logistic Regression and Random Forest gave better result on train and test dataset. The main advantage of using Random Forest for text classification is that it ensemble multiple and different kinds of decision trees and utilize an assortment of the different trees to improve the result of the model.

4.3 Results and discussion

Our proposed model provided this result on the test dataset. As it can be seen that in the case of positive, negative or neutral categories, our proposed model can classify with high precision, recall and f-measures. After applying logistic regression and random forest on the dataset, the performance values are recorded in table 1 and table 2.

As from above tables, it can be seen that both classifiers performed very well, but Random Forest works better as compared to logistic regression, with a consistent higher value in Precision, Recall, and F-score than logistic regression. The 82 % accuracy value on the test data is superior to our predefined target, which is to the maximum value we can achieve by setting the prediction labels for all samples to be the dominant class. The precision is also high for all the three classes and the recall rate is relatively low for the neutral classes.

For better illustrating the effectiveness of our proposed models, we also present examples of some negative and positive tweets classified by our proposed approaches.

Model-predicted accurately like Negative, Negative in the first column and Positive, Positive for the second column based on the test set.

Sentiment Class	Precision	Recall	F1-Score
Positive	0.80	0.74	0.77
Negative	0.73	0.53	0.62
Neutral	0.83	0.93	0.88

Table 1: Evaluation Metric of Logistic Regression.

Sentiment Class	Precision	Recall	F1-Score
Positive	0.82	0.74	0.78
Negative	0.75	0.60	0.65
Neutral	0.84	0.95	0.90

Table 2: Evaluation Metric of Random Forest.

Negative Tweets	Positive Tweets
" @united It's a shame choosing #United may be the difference between reuniting with aging friends and never seeing them again #PoorService",	"@united Big thanks to Ms. Winston for assisting me over the phone with a baggage claim issue today. She really went the extra mile!"
"@united flight attendant doesn,Äôt understand not understanding English doesn,Äôt mean they are deaf. Stop yelling English slowly at them"	"@United THANK U! Secured room for the night Thx to VERY helpful customer service rep N. Dorns. I thanked her. Can u 2? #goodenoughmother"

Table 3: Sample of the classified data into positive and negative tweets.

5 Conclusion and future scope

This study tackles the sentiment classification problem by utilizing two machine learning models. On the collected data, we achieve an accuracy of 82%. This study has impacts on the aviation industry in that it provides an effective and efficient way to monitor the passengers' sentiments for aviation companies to improve their service. For future work, we would like to conduct a deeper analysis of the data and extract more useful information for providing recommendations for several airplane organization and passengers. It would be also used to use a bigger dataset than the used dataset because a larger dataset may provide some better result than used one. The author would like to use also deep learning models and especially focus on how to identify the sarcasm because there are several sentences seems positive but their meaning is negative so this is a really big issue to sort out and at present, existing models are not efficient to sort it out effectively.

5.1 Acknowledgement

Prayag Tiwari has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 721321.

Sachin Kumar has received financial support by the Ministry of Education and Science of Russian Federation (Government Order 2.7905.2017/8.9).

6 References

[1] Kumar S and M Nezhurina. An ensemble classification approach for prediction of user's next location based on Twitter data. *Journal of Ambient Intelligence and Humanized Computing. 2018.* https://doi.org/10.1007/s12652-018-1134-3.

[2] Kumar S and M Zymbler. A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1):62, 2019. DOI: 10.1186/s40537-019-0224-1

[3] Yee L and P Tan. Gaining customer knowledge in low cost airlines through text mining. *Industrial Management & Data Systems*, 114(9): 1344-1359, 2014. https://doi.org/10.1108/IMDS-07-2014-0225

[4] Twitter: www.twitter.com access on 11.02.2019

[5] Zhang L, Y Sun and T Luo. A framework for evaluating customer satisfaction. *In Proc: International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, IEEE, Chengdu, China, 15-17 December, 2016. https://doi.org/10.1007/s11263-007-0056-x.

[6] Jaiswal AK, P Tiwari, S Kumar, D Gupta, A Khanna, JJPC Rodrigues. Identifying pneumonia in chest x-rays: a deep learning approach. Measurement, 145: 511-518, 2019. https://doi.org/10.1016/j.measurement.2019.05.076

[7] Tiwari P and M Melucci. Towards a Quantum Inspired Binary Classifier. IEEE Access, 7:42354-42372, 2019. DOI: 10.1109/ACCESS.2019.2904624

[8] Kusen E and M Strembech. An analysis of tweeter discussion on the 2016 Austrian presidential election, arXiv preprint arXiv: 1707.09939, 2017.

[9] Ahmed S, K Jaidka and J Cho. The 2014 Indian elections on Twitter: a comparison of campaign strategies of political parties. *Telematics and Informatics*, 33 (4):1071-1087, 2016. https://doi.org/10.1016/j.tele.2016.03.002

[10] Stieglitz S and L Dang-Xuan. Emotions and information diffusion in social media - sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29 (4):217-248, 2013. https://doi.org/10.2753/MIS0742-1222290408

[11] Gunarathne P, H Rui and A Seidmann. Whose and what social media complaints have happier resolutions? Evidence from Twitter. *Journal of Management Information Systems* 34 (2):314-340, 2017. https://doi.org/10.1080/07421222.2017.1334465

[12] Seunghyun BP, C Ok, B Chae. Using Twitter Data for Cruise Tourism Marketing and Research. *Journal of Travel & Tourism Marketing*, 33(6):885-898, 2016. https://doi.org/10.1080/10548408.2015.1071688

[13] Gräbner D, M Zanker, G Fliedl and M Fuchs. Classification of customer reviews based on sentiment analysis. In: Fuchs M, Ricci F, Cantoni L (eds) Information and Communication Technologies in Tourism 2012, Springer, Vienna, pp. 460-470, 2012. Doi: 10.1007/978-3-7091-1142-0_40

[14] Data-set:https://data.world/crowdflower/airline-twitter-sentiment accessed on 12.11.2018.