Categorization and Classification of Uber Reviews

Mugdha Sharma, Daksh Aggarwal and Divyasha Pahuja

Abstract This paper presents a technique for the categorization of reviews of the brand, Uber. This paper contains a classification algorithm which takes textual data as input. This algorithm takes many reviews and concepts (say, cost, safety, etc.). This algorithm is performed on online conversations happening on social media, taking into account the considered concepts. Further, the categorized reviews are classified according to the sentiment (that is, positive and negative). This paper also performs a comparison of different algorithms—Naïve Bayes, KNN, decision tree and random forest. The results of comparison of these different models are then represented using certain identified parameters.

Keywords Opinion mining · Classification · Categorization · Sentiment analysis · Natural language processing · Twitter · Uber reviews

1 Introduction

The unremitting digitalization is bringing a whole lot of different types of data into action. The data is present on different social platforms in sundry forms. Some of the data include complaints and feedbacks for various brands. These grumbles and feedbacks in form of comments, tweets, etc., can be used by brands to improve the services they provide or to work on some specific areas. So, the overall business can be improved if meaningful insights are drawn from the social media platforms for specific brands. The unprejudiced set for the study was to categorize the reviews of Uber present on Twitter and Uber application into different categories and concepts (say cost, safety, service, etc.) and to classify them as either positive or negative. The

M. Sharma (✉) · D. Aggarwal · D. Pahuja
Bhagwan Parshuram Institute of Technology, Rohini, India
e-mail: mugdha.sharma145@gmail.com

D. Aggarwal
e-mail: dakshaggarwal.official1996@gmail.com

D. Pahuja
e-mail: divyashapahuja@gmail.com

H. Sharma et al. (eds.), *Advances in Computing and Intelligent Systems*,
Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-15-0222-4_31

data retrieved is then presented in a graphical format for easy understanding. The methodology part explains how the classification and categorization are achieved through a proper process. First, the data is extracted and polished. A data dictionary is created containing all the synonyms of the label categories that are required (say safety, price, etc.). After this, an algorithm is applied that provides the multiclass categorization followed by the polarity judgment using the Naïve Bayes classification. The data extracted is presented in bar graph form for easy understanding.

The proposed approach helps the organization to recognize the reaction of thousands of people using their services in various sectors, and at the same time to indicate which areas of business lag behind and have scope of improvement. The algorithm created is simple and efficient which gives big organizations a platform to analyze people's introspection of their services. This approach is better than others as it gives an idea of the peoples' sentiment and reaction for each aspect whereas other approaches only provide with the sentiment overall. The paper also discusses briefly Uber and the reason for picking up the brand for study and analyses. The final section of the paper discusses the conclusion and the future scope of the paper.

2 Literature Review

The approaches considered for mining the customer reviews have been consulted from the following: Paper [1] performs sentiment analysis of reviews and feedbacks of Uber sourced from Facebook. This paper presents the technique of sentiment analysis, as well as the methods and stages it involves and discloses how Uber is perceived by its customers. Similar approaches for performing sentiment analysis on different datasets, in particular: Twitter Messages [2], Parents Feedback of Institutes [3] and Facebook Group [4], and in [5]. In addition to the explanation of the working of opinion mining using Rapid Miner, paper [6] performs parametric study of different classification algorithms. The comparison of different algorithms used in sentiment analysis is also done in [7, 8]. Saif et al. [9] use a lexicon-based approach. This paper demonstrates the value of contextual semantics of words for gathering sentiments of tweets from three different datasets. It used Thelwall-Lexicon as a case study and SentiCircle approach to gather the contextual semantics, while Ding et al. [10] propose a holistic lexicon-based approach that deals with excerpts that affect the opinion based on their linguistic patterns. In paper [11], the model created for opinion mining is used to get results for higher order n-grams. Naïve Bayes and SVM classifiers are proposed. Feature-based extraction and opinion mining are performed in [12]. Lexicon-based, Statistical and Intelligent Feature selection approaches are discussed for defining the polarity of opinion word. Mishra et al. [13] use dictionary-based approach to accomplish the same. In the paper [14], Twitter Vigilance platform is designed, which is a tool for gathering and analyzing Twitter data, presenting sentiment analysis of textual content. Paper [15] presents an algorithm for mining of sentiments on different platforms. It has three propositions: sentiment classification, feature-based sentiment analysis, and comparative sentences and relation analysis. Munjal et al. [16–18] proposed some new strategies for sentiment analysis and opinion dynamics using natural phenomenon.

3 Why Choose UBER

At the time of Uber's foundation, the need of the market was affordable drivers to have easy contact with their passengers for a smooth sailing taxi service. Uber's idea took the form of a mobile application named uberPOP. Through this application, passengers could receive fast transportation at a budget-friendly fare. The application's purpose was to match drivers with their passengers. Uber now is one of the world's largest global transportation service providers.

This business model was successful due to the presence of some key elements. These elements served to provide a clear real-time connection between both app users, the passengers and the drivers. The reason why people could get inexpensive transportation is the display of fare upfront as the passenger is booking the taxi. This promoted the idea of a clear expenditure estimate in the mind of the customer. Furthermore, traffic control and pollution reduction increased as private cars were used efficiently.

Uber was growing as a startup and becoming more and more diverse. The growth led to a lot of problems and the need for a mechanism for the brand to recognize the problems and the categories they fall into. Keeping the above factors this research paper chooses Uber for the study.

4 Data Sources

Data plays an important role in any application. So the selection of data is important for the application to run with high accuracy. The data sources used for this paper are.

4.1 Twitter Tweets

A lot of opinions are represented using tweets on Twitter. So the data from Twitter is extracted using Twitter API and then data cleaning and further processes are applied to make it usable. Tweepy is used to extract the tweets posted on Twitter. Tweepy is a python library which provides access to the Twitter API so we can get hold of any functionality that Twitter API offers. Tweepy gives access to Twitter through OAuth, which is the authentication system used by Twitter. The tweets are monitored using user_timeline() function.

4.2 Uber App Reviews

The official app reviews form the most essential part of data. Uber app is the foremost place where reviews related to any problem are posted. Kaggle is used to cater to the need. It provides with the dataset from the Uber app. Kaggle provides the data extracted from Uber app using web scrapers. The data is represented in.csv format which is ideal for analysis.

5 Methodology

This paper defines a step-by-step approach for achieving the desired results. Figure 1 shows the proposed workflow. First, the categories are decided, i.e., the classes on which the positive and the negative sentiments are to be found. This paper defines five such classes, namely, cancel, payment, price, safety, and service.

5.1 Data Loading and Data Cleaning

Data loading means to load the unclean data onto the software on which the analysis is to be performed. Preprocessing of data needs to be done. First preprocess task is to clean the data, i.e., to remove all the redundancies and inconsistencies present in the original data.

Cleaning of data is done by:

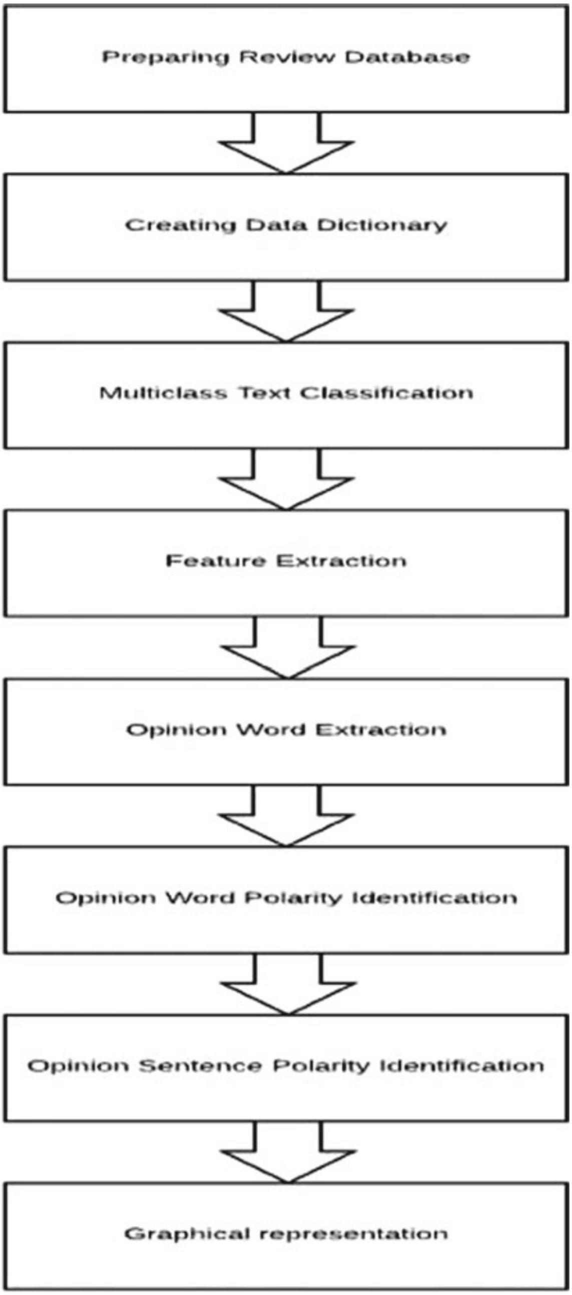
- Representing the review in lower case.
- Removal of all symbols, digits, and everything except alphabets.
- Removal of extra spaces and truncation of sentences.
- Stemming of words.

The data is stemmed (or lemmatized in this case), so that the word is transformed to its root form which makes it easier to evaluate the data. The data cleaning process is done using regular expressions. The re library in python is used to do so. And the nltk library consists of the functions for stemming and lemmatizing of words. After data preprocessing, the main data mining techniques are to be applied.

5.2 Applying Data Mining Techniques

For the main process, a data dictionary is created. The data dictionary contains synonyms for the words of the specified categories, i.e., cancel, payment, price, safety, and service. The synonyms for the determined concepts are gathered using WordNet corpus of nltk library.

Fig. 1 Implementation process


Preparing Review Database
Creating Data Dictionary
Multiclass Text Classification
Feature Extraction
Opinion Word Extraction
Opinion Word Polarity Identification
Opinion Sentence Polarity Identification
Graphical representation

After the dictionary creation, an algorithm to group all the reviews together in their respective categories is applied. This is done using a complex comparison of all the words of the review with all the category synonyms and if a match occurs then the review is added to that category. A review can belong to one or more than one concept. Now each category consists of a set of reviews which are then classified according to sentiments.

The main process of sentiment analysis, i.e., the judgment of the polarity of the review is performed using the Naïve Bayes classification algorithm which is a probabilistic approach for review classification. A bag of words model is created for the features. The Naïve Bayes classification classifies the reviews belonging to a particular category as positive or negative. The model is trained on the available dataset. The predicted sentiments are now compared with the attached data of sentiments. This forms the test set for the model.

5.3 Classifier Selection

They are used to show the classifier type the algorithm is based on. Probabilistic, linear, decision-based classifiers are some of the classifiers available.

The classification models tested in this paper are:

- Naïve Bayes,
- KNN,
- Decision tree, and
- Random forest.

From the comparison shown in Table 1, it can be concluded that Naïve Bayes algorithm is plain sailing and easier to understand and apply as compared to others. But it can be seen that random forest classification model has the best accuracy on our chosen dataset. At the end, the accuracy of the system depends on factors like the chosen domain, sources of data, and even the cleaning methods involved for data preprocessing. Other factors like Precision and Recall are also considered for a good comparison and analysis of the different models of classification.

Table 1 Comparison of classifiers

Algorithm	Naïve Bayes	KNN	Decision tree	Random forest
Understanding complexity	Less	Very less	Moderate	Moderate
Theoretical accuracy	Moderate	Very less	Moderate	High
Theoretical training speed	High	Very less	Low	Low
Performance with small observations	High	Low	Low	Low
Accuracy	90.28	50	94.02	95.5

5.4 Evaluation Parameters

The dataset which was used was segmented into positive and negative classes. The four cases for the given reviews and classifiers are: true positive, true negative, false negative, and false positive.

If the review is labeled positive and is classified as positive it is taken into account as true positive (TP) else if it is classified as negative it is counted false negative (FN).

Similarly, if a document is labeled negative and is classified as negative it is counted as true negative (TN) else if it is classified as positive it is considered as false positive (FP). Based on these outcomes, a two by two confusion matrix can be drawn for a given test set.

The following metrics can be derived from the confusion matrix (Table 2):

- Accuracy,
- Precision,
- True positive rate/Recall, and
- F-measure.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Table 2 Confusion matrix

#	Predicted positives	Predicted negatives
Actual positive cases	Number of true positive cases (TP)	Number of false negative cases (FN)
Actual negative cases	Number of false positive cases (FP)	Number of true negative cases (TN)

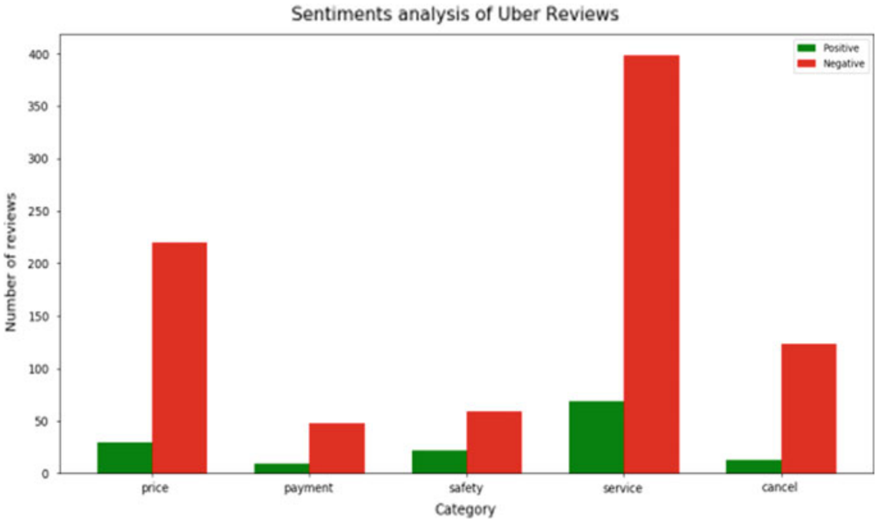


Fig. 2 Positive and negative reviews in each category

5.5 *Knowledge Presentation*

The main process is now completed, i.e., the knowledge has been extracted and just needs to be presented in a suitable format for easy use and understanding. The final results are shown graphically. The bar graphs are used to display the positive and the negative comments for a particular category. Figure 2 shows the results obtained and its graphical representation.

6 Conclusion and Future Work

In the paper, the comprehensive process for reviews and extraction of the meaning of these reviews has been discussed. Second, the paper shows a survey of diverse classifiers which shows that Naïve Bayes classifier best suits the working conditions.

Moreover, this paper presents a comparison of the various techniques on the basis of our recognized parameters.

The project has a huge and achievable future scope keeping in mind the time as well as economic constraints. The future scope includes using the technique of boosting to get more advanced and accurate results. Moreover, the project could be developed to support multilingualism or the reviews from different languages. Also, the various forms of writing need to be considered to handle the reviews with utmost accuracy, for example, the model should be able to detect sarcasm and the tone or context of the review.

References

1. Baj-Rogowska, A. (2017). Sentiment analysis of Facebook posts: The Uber case. In *8th IEEE International Conference on Intelligent Computing and Information Systems (ICICIS)* (pp. 2–5).
2. Kapil,P., & Sahoo, S. K. (2017). *Message classification for Twitter data* (pp. 2–7). IIT Patna, Bihar.
3. Patel, T., Undavia, J., & Patel, A. (2015). Sentiment analysis of parents feedback for educational institutes. *International Journal of Innovative and Emerging Research in Engineering, 2*(3), 2–3.
4. Akter, S., & Aziz, M. T. (2016). Sentiment analysis on Facebook group using lexicon based approach. In *ICEEICT* (pp. 1–3).
5. Bhuta, S., Doshi, A., Doshi, U., & Narvekar, M. (2014). A review of techniques for sentiment analysis of Twitter data. In *Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 583–591).
6. Dhivya Bino, D. V., & Saravanan, A. M. (2016). Opinion Mining from student feedback data using supervised learning algorithms. In *3rd MEC International Conference* (pp. 3–5).
7. Desai, M., & Mehta, M. A. (2016). Techniques for sentiment analysis of Twitter data: A comprehensive survey. In *International Conference on Computing and Automation (ICCCA)* (pp. 1–5).
8. Jain, A. P., & Katkar, V. D. (2015). Sentiments analysis of Twitter data using data mining. In *International Conference on Information Processing (ICIP)* (pp. 1–4).
9. Saif, H., He, Y., Fernandez, M., & Alani, H. (2014). *Adapting sentiment lexicons using contextual semantics for sentiment analysis of Twitter* (pp. 3–7). Knowledge Media Institute, The Open University, United Kingdom.
10. Ding, X., Liu, B., Yu, P. S. (2008). *A holistic lexicon-based approach to opinion mining* (pp. 4–7). Department of Computer Science University of Illinois at Chicago, Morgan Street Chicago.
11. Tripathi, G., & Naganna, S. (2015, June). Feature selection and classification approach for sentiment analysis. *Machine Learning and Applications: An International Journal (MLAIJ), 2*(2), 5–11.
12. Tribhuvan, P. P., Bhirud, S. G., Tribhuvan, A. P. (2014). A peer review of feature based opinion mining and summarization. *International Journal of Computer Science and Information Technologies (IJCSIT), 5*(1), 247–250.
13. Mishra, P., Rajnish, R., & Kumar, P. (2016). Sentiment analysis of Twitter data: Case study on digital India. In *International Conference on Information Technology (InCITe)* (pp. 3–5).
14. Cenni, D., Nesi, P., Pantaleo, G., & Zaza, I. (2017). *Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis* (pp. 2–6). Department of Information Engineering (DINFO), University of Florence, Florence, Italy.
15. Kasture, N., & Bhilare, P. (2015). An approach for sentiment analysis on social networking sites. In *Computing Communication Control and Automation (ICCUBEA)* (pp. 390–395).
16. Munjal, P., Kumar, S., Kumar, L., & Banati, A. (2017). Opinion dynamics through natural phenomenon of grain growth and population migration. In *Hybrid Intelligence for Social Networks* (pp. 161–175). Springer, Cham.
17. Munjal, P., Narula, M., Kumar, S., & Banati, H. (2018). Twitter sentiments based suggestive framework to predict trends. *Journal of Statistics and Management Systems, 21*(4), 685–693.
18. Munjal, P., Kumar, L., Kumar, S., & Banati, H. (2019). Evidence of Ostwald Ripening in opinion driven dynamics of mutually competitive social networks. *Physica A: Statistical Mechanics and its Applications*.