Analysis of Airline Dataset using Machine Learning Techniques

Rishabh Jain
Student, Computer Engineering Department
Poornima College of Engineering, Jaipur.
2020pcecsrishabh155@poornima.org

Sharma Chetan Ramkishore
Student, Computer Engineering Department
Poornima College of Engineering, Jaipur.
2020pcecsshetan208@poornima.org

Tushar Sharma
Student, Computer Engineering Department
Poornima College of Engineering, Jaipur.
2020pcecstushar189@poornima.org

Sarthak Bhardwaj
Student, Computer Engineering Department
Poornima College of Engineering, Jaipur.
2020pcecssarthak167@poornima.org

Abstract

In the contemporary world, Data analysis is a challenge in the era of varied inters- disciplines though there is a specialization in the respective disciplines. Then, it is increasingly difficult to handle huge amount of data that gets generated no matter what's the business is like, range of fields from social media to finance, flight data, environment and health. the burgeoning research in which data on air passenger flows are used to analyse a network of world citics. Rather than taking the relevance of such airline statistics on trust, we consider their advantages and drawbacks in the con text of the different approaches devised in the empirical research at large. To assess the potential of data on air passenger flows in this context, we construct a taxonomy of approaches that distinguishes between information on global corporate organization and large-scale infrastructure networks. Social media is a media that many users need to be connected with other users. One of the most widely used social media is Twitter. This Twitter contains opinions or short messages called tweets. The invited company also needs feedback from its customers to find out their view of the requested service. Logarithmic multivariate Gaussian models are trained to evaluate the performance of aircrafts at different flight phase separately. By including a forward synchronization, feature selection, and mini-batch training process, this model overcomes challenges introduced by the large size and high dimensionality of flight datasets. the aviation industry generates massive data every day. By analyzing the aviation big data, aviation manufacturers and airlines can optimize the flight of civil aircraft including risk reducing, operation optimization, and personalized services. Building a platform for storing and analyzing the aviation big data becomes an important task for civil aviation.

Keyword: Machine Learning, Skypra, Random Forest Algorithms, Big Data

I INTRODUCTION

Big Data is not only a Broad term but also a latest approach to analyze a complex and huge amount of data; there is no single accepted definition for Big Data. But many researchers working on Big Data have defined Big Data in different ways. Nowadays, the aviation industry generates massive data rapidly, e.g., the operational data in every day. The big data technology can deal with the data and bring opportunities for the aviation industry chain, including design, development, production, and operation. For aircraft manufacturing, the big data technology helps analyze the historical manufacturing data and suggest the light and reliable aircraft materials. For the aircraft design, the big data technology can suggest the optimal design ideas considering safety and reliability, comfort, and fuel consumption by analyzing the massive data generated in the flight test. For the operations, airlines can provide personalize services to users according to the recommendations of the big data technology. For the aircraft health management, the big data technology can provide fault diagnosis for aircrafts. It can forecast the possible faults by exploring the flight data. The big data technology can also propose the maintenance suggestions, namely predictive maintenance. In this paper, we propose a big data platform for civil aircrafts. The platform collects massive civil aviation data from multiple data sources and analyzes the big data using various approaches. The proposed platform provides frameworks for big data storage and processing. The platform manages data models and analysis models for providing different types of analysis services. Moreover, the platform provides application systems for civil aviation with the help of the analysis services. Social media is become a reference for several companies that aim to find a description about customer behavior today [1]. The results of the analysis can also be used as a guideline for taking policies related to future business direction decisions for the company concerned. This analysis results in the classification of opinions or sentiments obtained from tweets provided in the form of datasets by Kaggle. Therefore, the collection of texts from this tweeter is very valuable because it contains hidden information and must be revealed for the purpose of the company. Disclosure of such information involves mining data with various types of classifiers as well. This data mining process has main task in converting unstructured text data into structured data. Text data that has been structured is a requirement for the data mining process. In addition, sentiment analysis also involves a Natural Language Preprocessing approach to get meaning from the collection of words in the tweet. One of social media that is often used is Twitter. Twitter users currently reach 330 million people in the world and this social media can also generates 8000 data in every second. On Twitter, there is the term tweet which

means a collection of texts or sentences that can contain news content, opinions, arguments, and several other types of sentences. Twitter users are not only from the hands of the youth but various circles even from government and business circles as well. The usage of Twitter can be through various platforms such as mobile devices, websites, or desktop applications by connecting to the internet. This ease of access has resulted in a growing number of users over time. An airline is an organization that provides flight services for passengers or goods. They rent or own aircraft to provide these services and can form partnerships or alliances with other airlines for mutual benefit. The company needs feedback from its customers to find out their views on the services of the airline company. The desired feedback will be difficult to obtain if done alone using a questionnaire, sampling, or interviewing samples from customers.

II LITERATURE REVIEW

Customer satisfaction is a complex customer experience in the service industry, and can be defined as an evaluation on which the customers have experienced. Understanding what consumers expect from a service industry is important in order to provide a standard of comparison against which consumers judge an organization's performance regarding the expectation. Service quality can be defined as a consumer's overall impression of the relative efficiency of the organization. In addition, customer satisfaction can be defined as experience made on the basis of a specific service encounter, and it is contributed to customer loyalty, repeat purchase, favorable word-of-mouth, and ultimately higher profitability. The customer sets expectations for the product or service and these expectations are becoming the standard before purchasing. Once the product or service is used, the outcomes or perceptions are compared to pre-purchase expectations. Consumer generated content contains a variety of media forms and types. Online reviews that reflect how customers explain and share their experiences in various forms are a valuable way of figuring out what customers think, and online platforms allow customers to share experience with information, opinions, and knowledge about products, services and brands. Customers seek out a variety of information to be confident of their choices, thereby reducing the perceived risk. Therefore, in this study, data was collected through the online review written by those who have already experienced it. Due to the advance of technology, it is easy for customers to post their experience with products and services on the website. It is especially relevant for service industries because of intangible characteristics of services. Many studies have demonstrated the strong impact of online customer reviews. For example, Dellarocasetal have demonstrated that online review metrics can accurately forecast movie revenue. Minnemaetal. have demonstrated that product returns have a strong relationship with online customer reviews and the effect of it needs to be considered. The number of reviews has grown exponentially over the past few decades, and the content of the reviews has had a significant impact on the repurchase of products. Sotiriadis and van Zyl found that online reviews and recommendations affect the decision-making process of tourists towards tourism services and WOM has a significant impact on the subjective norms and attitudes towards an airline, and a customer's willingness to recommend. Therefore, the online review would be very useful for airlines to understand their diverse customer base in order to take service improvement strategies since airlines are inherently multicultural businesses. Skytrax is an airline quality assessment website that performs an online assessment after the customer directly used each airline [1]. Skytrax has worked for over 150 airlines across the globe, from the world's largest airlines through to small domestic carriers and it is a world-recognized brand that provides professional audit and service benchmarking programs for airlines on product and service quality. They employ professional auditors to assess the quality of the work done in an airline, both onboard and in the airport terminals. These evaluations are based on consistent standards. Skytrax is an airline quality assessment website that performs an online assessment after the customer directly used each airline. The best airlines in the world highly recognize these quality awards presented by the Skytrax. When an airline is awarded a 'Skytrax star-ranking' or advances to a higher ranking, they immediately announce this news by publishing press releases and posting it on their websites' most visible spots. Big Data represents a new era in data exploration and utilization. This is occurring because of new sources of data, and since the very beginning of the Internet, users have been keeping generating data on the Internet. The big data intensifies the need for sophisticated statistics and analytical skills. The big data technologies are providing unprecedented opportunities for statistical inference on massive analysis, but they also bring new challenges to be addressed, especially when compared to the analysis of carefully collected smaller data sets. A semantic network analysis is becoming its own research paradigm as well as a method for analysis of the big data. The semantic network analysis is a method of identifying and analyzing relationships between words to describe a part of a connected network. The semantic network analysis, as a method of qualitative textual analysis, provides a strong theoretical and methodological foundation with which to describe the semantic nature of the online tourism domain. Centrality and proximity were employed to measure the structure of the semantic network and to compare the differences between two semantic structures in the Jo and Kim's study. The similarities matrix generated in the text analysis can be used as input into multidimensional scaling to assess both the content and structure of the semantic network. While in this study, the similarities of the top 100 frequent words were conducted by CONCOR analysis, and the methodology for visualizing data are vital for understanding the semantic network of words. The network can be visualized and verified, and the visual representation makes it easy Sustainability 2019, 11, 4066 4 of 17 to see at a glance the structure of the network or the associativity between nodes. The approach and visualization for the semantic network analysis of this study was performed by Ucinet 6.0.

It applied Logistic Regression and Decision Tree (Random Forest) algorithms on the model to predict delays. Factor analysis is used to understand the possible factors affecting the delay of a flight. Hence, the analyzed factors are implemented using the random forest algorithm. The estimate time of arrival and delays are compared from both the models. The research claims decision tree algorithm to be more effective compared to logistic regression

III Conclusion

Digital flight data are collected by airlines from all aircraft on a regular basis. These data contain a large amount of information about daily operations that could be used to inform airlines for safety improvement. Yet the analysis of such data is challenging due to the increased complexity and variability in air transportation operations. We developed a new data driven approach that can support safety experts to utilize digital flight data, better monitor flight operations and potentially improve airline safety. The new approach can automatically detect anomalous situations without extended initial tuning which are time consuming and expensive. Results of abnormal flight can inform further analysis by domain experts to identify risks, and to determine whether mitigation measures are needed to prevent accidents. The method was tested on real world datasets provided by international airlines. Results show that Cluster AD-Data Sample is able to detect anomalies that are operationally significant and may represent increased level of risks. Compared with other data-driven methods to detect anomalies in flight data, Cluster AD-Data Sample performed better in detecting known unsafe events. Further study is needed to comprehensively evaluate its performance in detecting unknown issues. several things can be concluded as follows that the Naive Bayes classification has faster training data on the airline dataset. The SVM Linear Classifier has the best accuracy in classifying tweets in this airline dataset. Features with an MI calculation value greater than other features indicate that the feature contains important information. The use of features that have less tendency to train data processing time faster. The civil aviation can generate massive data. The aviation industry is beneficial from the aviation big data, e.g., reducing flight risks, facilitating manufacturing, improving services, and reducing cost. Thus, building an integrated big data platform for civil aircraft becomes a critical issue. In this paper, we propose a civil aircraft big data platform. The platform provides management and analysis of aircraft operating around the world. The platform monitors the aircraft operational health status in real time, and reduces operating costs. In the future, we will focus on the following work: the big data-based aircraft diagnosis and prediction algorithms; optimization of the algorithms using deep learning methods; construction of big aircraft data analysis center, particularly the security of the platform. We will focus on investigating enhanced security of aviation data sharing
using the block chain technology.

IV FUTURE SCOPE

The above research methodology should be performed on the data collected for the recent years, owing to the population rise in recent years leading to increase in the number of flights. To obtain a detailed analysis, a more thorough localized search and scrutiny must be conducted to accurately determine the arrival or departure delay. Moreover, this methodology can be used for all the airports. The results of our research can be extrapolated to perform the above and determine accurately the delay and help in determining the major reasons causing it. The MI method will look for the value of each feature. Features that have high value are features that carry important information. This will affect the value of accuracy in sentiment analysis. However, the highest value of the results of the calculation of mutual information is spread on existing features. This requires certain methods to select those features with MI calculations more effectively. In this study still checking the gradual MI value, from beginning to end. Therefore, more effective methods are needed to obtain features with high MI values.

V REFERENCES

[1] Ben Derudder, frank Witlox on Mapping world city networks through Airline flows 0966-6923/$ doi:10.1016/j.jtrangeo.2007.12.005 Journal of Transport Geography 16 (2008) 305–312

[2] Hastari Utama on Sentiment Analysis in Airline Tweets using Mutual Information for Feature Selection UTC from IEEE Xplore.

[3] Lishuai Li, R. John Hansman, Rafael Palacios, Roy Welsch on Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring http://dx.doi.org/10.1016/j.trc.2016.01.007 0968-090X/ 2016 / Transportation Research Part C 64 (2016) 45–5

[4] Guoyi Li, Hyunseong Lee, Ashwin Rai, Aditi Chattopadhyay on Analysis of operational and mechanical anomalies in scheduled commercial flights using a logarithmic multivariate Gaussian model https://doi.org/10.1016/j.trc.2019.11.011 Transportation Research Part C 110 (2020)

[5] Sujie Li,Yi Yang, Lu Yang, Haixia Su1 Guigang Zhang, Jian Wang on Civil Aircraft Big Data Platform 978-1-5090-4284-5/17 $31.00 © 2017 IEEE DOI 10.1109/ICSC.2017.51

[6] Allen Wong, Sijian Tan, Keshav Ram Chandramouleeswaran, Huy T. Tran on Data-driven analysis of resilience in airline networks https://doi.org/10.1016/j.tre.2020.102068 Transportation Research Part E 143 (2020) 102068

[7] Hyun-Jeong Ban and Hak-Seon Kim on Understanding Customer Experience and Satisfaction through Airline Passengers' Online doi:10.3390/su11154066 Sustainability 2019, 11, 4066www.mdpi.com/journal/sustainability

[8] Devansh Shah, Ayushi Lodaria, Danish Jain, Lynette D'Mello on Airline Delay Prediction using Machine Learning and Deep Learning Techniques DOI:10.35940/ijrte.B4047.079