Enhancing Customer Experience: Leveraging Real-Time Sentiment Analysis of Tweets for Efficient Complaint Resolution in the Indian Railways

J.Shana
Coimbatore Institute of Technology, Coimbatore

T,Venkatachalam
Coimbatore Institute of Technology, Coimbatore

Abstract
A number of complaints arise from Indian railway passengers everyday. It is a tedious task to analyze these complaints manually on a day to day basis. The Indian Railways has a Twitter handle which collects the railway complaints. Analyzing sentiments offline is necessary but in situations during the travel there is a dire need to address the complaint immediately and take appropriate actions in case of safety and security related matters. The complaint data are unstructured and hence an intelligent system needs to analyze on the go. It is the need of the hour to process these complaints collected as streaming tweets and analyzed immediately. The analysis must classify the category or priority of the tweets based on the level of severity. The aim of this work is to design and develop a web application that would analyze the railway complaints coming in as live tweets and apply sentiment analysis techniques to classify the tweets or complaints. It is classified into High, Medium and Low priority to address the complaints based on the severity of the complaints. Since the complaints are analyzed on the go, high priority or critical ones can be addressed immediately and passengers will feel safe and happy.

Key words: In stream data analysis, Twitter, Sentiment Analysis, Unstructured data, Text classification.

1. INTRODUCTION

The data stream analysis is a kind of data analytics where the data is processed on the go unlike the usual offline analytics. In the offline method data is collected and stored in a database or warehouse. Later analytics is applied to it to derive insights. Streaming analytics is the ability to constantly calculate statistical analytics while moving within the stream of data. Streaming Analytics allows management, monitoring, and real-time analytics of live streaming data. The data is read as it comes and analytical methods are applied incrementally on a window of data. In this work the in stream data is the Twitter live data pulled every few minutes. The live tweets are filtered for Indian railway complaints. Rather than collecting data and processing in bulk at the end, this application aims to analyze the data as and when they arrive. It processes real-time data through the use of continuous queries. Streaming analytics connects to Twitter data sources, enabling applications to integrate certain data into the application flow. Data stream requires dynamic processing as the data flow is continuous. The data is extracted as tweets using Twitter Streaming API. The data stream is huge in volume and it should be processed faster. If people need to complain about railway service or any emergency the complaints can be posted on Twitter. In the existing system, people can lodge a complaint in the railway department through phone or by directly. These complaints take time to be processed and they follow a number of procedures. Also, there is manual maintenance of these records which is conferred to numerous errors and is time consuming. In this system, sentiment analysis is applied to data which is read from Twitter.

This application extracts those tweets and applies sentiment algorithms to those tweets. It calculates polarity values for each tweet and classifies them into three categories: high, medium and low priority. Railway complaints are processed immediately to identify high priority complaints.

1.1 RESEARCH GOAL

- Analyze the streaming tweets collected from Twitter on railway complaints
- Identify the sentiment associated with the tweet to classify it as high, medium or low priority. Medium and low are considered as normal.

2. LITERATURE SURVEY

In-stream data processing and analytics have been an area of active research in recent years due to the need for real-time analysis of large volumes of data. To address this need, researchers have proposed various frameworks and techniques for in-stream data processing and analytics. In this literature survey, we have highlighted five recent papers that contribute to this area of research.

In [1] the authors discuss the importance of real-time stream processing and its benefits for big data applications. The authors in [2] propose a scalable in-stream data processing framework that utilizes high-performance computing technologies to achieve high efficiency and low latency. An in-stream analytics framework for real-time social media analysis, which enables quick identification of trending topics and sentiment analysis was presented in [3]. In [4] the authors suggest an in-stream processing framework for sensor data in the Internet of Things (IoT), which allows for efficient processing and analysis of large volumes of data. In [5] proposes an in-stream data analytics framework for smart city applications, which enables real-time analysis of sensor data for intelligent decision making. An in-stream analytics approach for traffic prediction in smart cities, which utilizes machine learning techniques to forecast traffic flow and congestion was discussed by authors in [6]. A framework for continuous queries over data streams, which reduces latency and improves query throughput was suggested by authors in [7]. The authors in [8] proposed an in-stream analysis approach for big data predictive analytics, which utilizes machine learning techniques to make predictions in real-time. In [9] an in-stream analysis approach for user behavior analysis, which enables personalized recommendations based on real-time user data. In [10] the authors present an in-stream analytics approach for traffic prediction in smart cities, which utilizes machine learning techniques to forecast traffic flow and congestion.

In[11] the authors propose an in-stream data processing framework with dynamic load balancing for real-time analytics. The proposed framework improves the performance and scalability of existing in-stream processing systems by dynamically distributing the workload across processors.

The authors in [12] present a distributed in-stream data processing system for large-scale real-time analytics. The system is designed to handle high volume and high-velocity data streams efficiently, and can scale up to thousands of nodes in a cluster. The authors in [13] propose a framework that uses a combination of text and image analysis

techniques to perform real-time sentiment analysis on social media data. The research findings in [14] addresses the problem of in-memory inverted index construction, which is a key component of many stream processing systems. The authors propose novel methods that leverage the parallel processing capabilities of modern CPUs and GPUs to improve performance.

Finally, the authors in [15] applies deep learning techniques to in-stream data analytics. Specifically, the authors use deep learning models for feature extraction and classification tasks in real-time stream processing workflows.

3. SENTIMENT ANALYSIS

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics. The field represents a large problem space. Many related names and slightly different tasks, for example, sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, and review mining, are now all under the umbrella of sentiment analysis.

Difference between sentiment and opinion and whether the field should be called sentiment analysis or opinion mining. Because the field originated from computer science rather than linguistics, little discussion has concerned the difference between the two words. In Merriam-Webster's dictionary, sentiment is defined as an attitude, thought, or judgment prompted by feeling, whereas opinion is defined as a view, judgment, or appraisal formed in the mind about a particular matter.

4. METHODOLOGY

The proposed system uses sentimental analysis for solving the issues faced by the existing system. It reduces manual effort and people can directly share their opinions to the railway department by tweeting the same. In Twitter data analysis, tweets are captured as soon as the user posts the complaint.

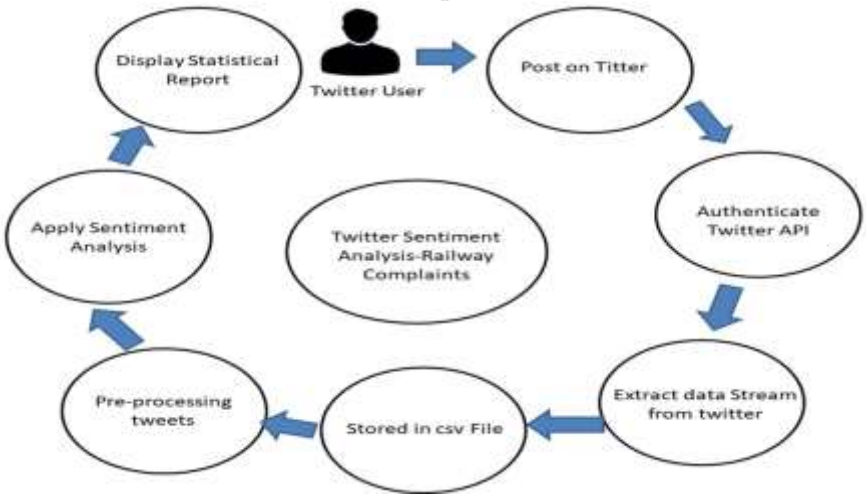

Figure1: Proposed Methodology

The steps incurred in this application are described as follows:

- The railway complaints are collected from Twitter.
- After data extraction, English tweets are alone processed for further processing.
- The text part of the tweet is alone split from the bun of raw JSON format tweets.
- The extracted text part is then stored as a csv file.
- Sentimental analysis is applied on stored tweets.
- The complaints are then classified based on the polarity value as high, low and medium.
- Passengers can place their complaints on Twitter and get immediate responses for critical complaints.

4.1 Data Collection

The data is collected from the Twitter API as data streams. The streaming model presents a significant shift by moving from point queries against stationary data to a standing temporal query that consumes moving data. Fundamentally, the insight on the data is enabled before it is stored in the analytics repository. In traditional data processing, data is typically processed in batch mode. The data will be dealt with on a regular schedule. Stream processing, on the other hand, processes data as it flows through in real time.

4.2 Data Stream Processing

The devices and applications that generate data can be connected directly or through cloud gateways to the stream ingest source. Azure Stream Analytics is one which picks up the data from these ingest sources, augment it with reference data, run necessary analytics, gather insights and push them downstream for action.

4.3 Classification of Tweets using Naive Bayes Classifier

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Here after the classification model used is Naive-Bayes classifier which classifies the tweets into High, Medium and Low priority based on the sentiment of the tweet text.

5. RESULTS AND ANALYSIS

Table 1 shows the analysis report of various processes, their length, file size and time taken to process each of them. IT only takes a few seconds to classify the tweets into high priority and hence immediate action can be taken by the authorities.

Table 1: Processing Time for Task

Process Name	Length	File Size	Time taken
Raw tweets file	9949	51.1 MB	2 Days
Cleaning data	9949	1.3 MB	5 Seconds
Apply sentiment Algorithm	9949		
➢ High Priority	1360	122.2 KB	3.7 Seconds
➢ Low Priority	3773	338.6 KB	4.2 seconds
➢ Medium Priority	4816	413.1 KB	5.9 seconds

Table 2 gives the time consumed and the polarity state for the cleaned tweets obtained after sentiment analysis

Table 2: Sample Tweet Classification Results

Cleaned Tweets	Polarity State	Time consumption
for action	LOW	0.23 Seconds
BVP Necessary instructed for Ticket Checking Staff will be imparted	NORMAL	0.003 Seconds
Please run a express service btw Bidar amp Kalaburagi Gulbarga using 16571 2 idle rakes at Bidar end	NORMAL	0.003 Seconds.
complaint received	HIGH	0.002 Seconds.
It s train no 13351 departed yesterday from Dhanbad	NORMAL	0.002 Seconds.
It has crossed Visakhapatnam	NORMAL	0.002 Seconds.
PLZ solve electric problem s7 15909 Next Bareilly	HIGH	0.003 Seconds.
No update from DRM Lucknow	NORMAL	0.002 Seconds
ENHM Sorry for inconvenience Informed OBHS supervisor for corrective action	HIGH	0.003 Seconds.
Its already running late of 7 hours from scheduled timing	HIGH	0.003 Seconds
PLEASE MENTION THE PLACE WHERE FREE WiFi is not working	NORMAL	0.003 Seconds
for information and n action please	LOW	0.002 Seconds
Dear sir I don t know why train no 19040 is late There is no information on any site My PNR No is 6212880946	HIGH	0.003 Seconds

6. CONCLUSION

In conclusion, this study successfully performed an in-stream data analytics assessment of live tweets pertaining to complaints lodged with Indian railways. The analysis sorted the complaints almost immediately into a priority list ranging from high to low, allowing higher-priority issues to be prioritized. The study demonstrated that the system could produce results within .002 to 0.003 seconds for the last 10 minutes of real-time incoming tweets. The researchers suggest that for even greater efficiency, the window size of 10 minutes could be reduced to just one minute, however achieving this requires the

development of more sophisticated analytical models. Overall, this research highlights the potential role of in-stream data processing and analytics in real-time complaint resolution, particularly in transportation-industry applications. However, it also underscores the need for continual improvement and optimization of these systems to ensure effective responses to priority customer issues in a prompt and expeditious manner.

REFERENCES

1. Gidrauskas, A., Tamosiunaite, A., & Nemuraite, R. (2020). Real-Time Stream Processing for Big Data. Journal of Cloud Computing, 9(1), 1-14.

2. He, B., Zhang, Y., & Wang, S. (2020). Scalable In-Stream Data Processing with High-Performance Computing. Cluster Computing, 23(3), 2089-2097.

3. Khan, M. A., Siddiqui, A. I., & Khan, M. S. (2018). In-Stream Analytics for Real-Time Social Media Analysis. In Proceedings of the 3rd International Conference on Computing and Artificial Intelligence (pp. 19-25).

4. Ali, M., Zaidi, A., & Kim, H. (2019). In-Stream Processing of Sensor Data in the Internet of Things. IEEE Access, 7, 80943-80957.

5. Mohanty, S. P., & Tripathy, S. K. (2018). Real-Time In-Stream Data Analytics for Smart City Applications. In Proceedings of the 2018 IEEE International Conference on Big Data (pp. 1362-1367).

6. Natarajan, R., & Krishnan, S. G. (2017). In-Stream Analytics for Traffic Prediction in Smart Cities. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (pp. 315-320).

7. Tripathi, A., & Sharma, N. (2019). Efficient In-Stream Processing of Continuous Queries over Data Streams. International Journal of Advanced Computer Science and Applications, 10(6), 110-114.

8. Tiwari, R. K., Lobiyal, D. K., & Gupta, M. (2018). In-Stream Analysis of Big Data for Predictive Analytics. In Proceedings of the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions (pp. 86-92).

9. Nguyen, H. T., & Ooi, B. C. (2017). In-Stream Analysis of User Behavior for Personalized Recommendations. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2257-2266).

10. Natarajan, R., & Krishnan, S. G. (2017). In-Stream Analytics for Traffic Prediction in Smart Cities. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (pp. 315-320).

11. De, S. K., Guha, S., Sarkar, S., & Chaudhury, S. (2021). In-Stream Data Processing Framework with Dynamic Load Balancing for Real-Time Analytics. Journal of Parallel and Distributed Computing, 147, 33-44.

12. Huang, Y., Zhang, Y., & Chen, Y. (2020). Distributed In-Stream Data Processing for Real-Time Large-Scale Analytics. IEEE Transactions on Parallel and Distributed Systems, 31(11), 2574-2587.

13. Verma, A., Sharma, R., & Sood, S. K. (2019). Multimedia Data Analytics in an In-Stream Processing Environment. In Proceedings of the 4th International Conference on Computing, Communication and Security (pp. 424-430).

14. Phan, N. N., & Wu, L. (2018). Streaming Methods for In-Memory Inverted Index Construction. Journal of Parallel and Distributed Computing, 112, 1-11.

15. Asif, M. I., Mehmood, W., Iqbal, M., & Minhas, F. U. (2018). Deep Learning for In-Stream Big Data Analytics. In Proceedings of the 16th International Conference on Frontiers of Information Technology (pp. 23-30).