Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Analyzing Performance and Accuracy in Complaint Classification using Spark MLlib

Mohammed Bachir MAHDJOUB [1], Fatima Zohra LAALLAM [2], Messaoud MEZATI [2]
[1] *Mohamed Khider University, Biskra, Algeria, bachir.mahjoub@univ-biskra.dz*
[2]*Kasdi Merbah University,Ouargla, Algeria*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Customer complaint data explosion poses the need for complaint classification methods that are scalable and efficient to facilitate Customer Relationship Management (CRM). This work surmounts the complaint classification issue of correctly classifying the unstructured customer complaints using the assistance of Apache Spark and its machine learning library, MLlib. A multi-stage PySpark pipeline used classification from the text of Amazon product reviews to "Highly Dissatisfied" and "Mildly Dissatisfied" classes. Three of the most popular classification algorithms—Naive Bayes, Logistic Regression, and RandomForestClassifier—were evaluated on the entire set of metrics like accuracy and macro F1-score and weighted recall and precision. Our experiments show that while the best accuracy was produced by the model of RandomForestClassifier all things being equal, the most balanced performance was provided by the model of Naive Bayes with the best macro F1-score of 0.6884 and highest weightage of precision of 0.7022. This optimal trade-off makes the model best suited for practical deployment. Our discovery is that for this specific classification task the most efficient solution for consistently and correctly classifying the customer complaints on large scale is the algorithm of Naive Bayes.

Keywords: Spark MLlib, Complaints Classification, Customer Relationship Management (CRM), Big Data, Machine Learning, Performance Analysis.

1. INTRODUCTION

A determinant of today's highly competitive world success is the comprehension of customer remarks. At the heart of a mature Customer Relationship Management (CRM) initiative is comment handling and processing efficiently. With the complaints from the customers having grown manifold in all sectors, be it banking or e-commerce sectors, the complaints now actually form a trove of business intelligence. The complaints offer instantaneous inputs on the defects in the product, service gaps, and evolving requirements of the customer. The classifications of the complaints therefore become critical to keep the customer contented, defend the brand name, and invoke continuous improvements. The classifying used to be a manual process in the past, but that became unsustainable once the complaints grew manifold. The issues of classifying the unstructured text of various language and ambiguity only made the application of the fast speed and high-quality classification a herculean task.

Customer data explosion has highlighted the critical role of Big Data technologies in the modern-day CRM. Big Data technologies provide the infrastructure support to collect, store, process, and analyze large batches of customer data from various sources like social networking websites, transcripts of the call center, and e-mails. With Big Data, the CRM systems go beyond just responding to problems to the detection of trends and system problems proactively, thereby driving well-informed decisions and individualized customer experiences. Apache Spark is one of the leaders in this group and is a distributed computing system designed for large-scale data processing. Its equivalent, Spark MLlib, is a horizontally scaling machine learning library with a rich collection of algorithms that is optimized for the distributed environment. Together, Spark and MLlib offer a very good solution for large-scale text classification tasks like the challenging task of resolving customer complaints.
This paper presents the problems and results of using Spark MLlib for complaint classification. Our objective is to provide useful, empirical insights on its usage for this crucial business activity. Our research provides the answer to some crucial questions: How do the different algorithms of Spark MLlib vary from one another and the state-of-art

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

in terms of accuracy and computational performance on practical complaint data from the field? What are the main performance bottlenecks of Spark MLlib for this application? Besides the above research questions, we also demonstrate techniques and best practices on solving the above problems and enhancing the overall performance of the Spark MLlib on large complaints datasets. The structure of the paper is the following. In Sect. 2, we summarize CRM in the Big Data world. In Sect. 3, we describe Spark and MLlib. Our technical discussion on complaint classification comes next in Sect. 4. Our methodology is given in Sect. 5. Our results and discussion come in Sect. 6. Our conclusions and recommendation on future work come in Sect. 7.

2. CUSTOMER RELATIONSHIP MANAGEMENT (CRM) IN THE ERA OF BIG DATA

Customer Relationship Management (CRM) has come a long way from its humble start as an electronic filing system. In the beginning, the CRM systems featured not much more than electronic rolodexes and sales automation tools with the sole concern being administrative productivity and capturing transactional interactions. The emphasis during the formative years was on records and sales automation and limited analytics. The first contact management application was ACT, started in 1987, and the term "CRM" was used around 1995. Big Data has actually transformed CRM and made it the center of business strategy[1][2]. The explosion of digital touchpoints—social networks, websites, mobile applications, e-mails, and live chats—created tremendous amounts of customer data. Big Data technologies allow organizations to gather, store, and process the numerous, large-volume interactions. This enables companies to move on from just keeping records to extracting customer preferences, predicting needs, and personalizing experiences on a never-before scale. Against this information abundance backdrop, complaint data is a valuable but challenging asset. Complaint classification is a principal CRM function that identifies systemic issues, offers inputs for product and service improvements, and facilitates proactive customer support to generate loyalty and reduce churn. Complaint data also provides a near-real-time snapshot of customer satisfaction and market sentiment[3][4][5]. Complaint data nevertheless raises serious analytics issues. It is typically free-form text and usually contains informalities, typoes, and industry parlance. One of the principal issues is that such datasets will most often be highly im-balanced[6]. That is, a few common complaint classes such as "bill inquiry," for example, will make up the vast majority of records, but critical but rare complaint classes such as "security breach" or "life-threatening product flaw" will be severely undersampled. If not handled properly, such imbalance can translate to machine learning models that poorly generalize on these critical minority classes.

3. LEVERAGING SPARK AND MLLIB FOR BIG DATA ANALYTICS

Data's increasingly large-scale and diverse nature, particularly unstructured text like customer complaints, demand rapid and flexible processing architectures. Apache Spark has become the star technology of the big data landscape and offers an optimal solution for large-scale machine learning and data analytics[7][8]. The Spark architecture is built around a master-slave topology and involves a Driver program running and scheduling tasks, Executors running tasks and retaining data on nodes, and a Cluster Manager (like YARN or Kubernetes), responsible for resources. The most significant plus point of Spark lies in its in-memory computation so that disk I/O comes down drastically and performance becomes faster in the case of iterative machine learning algorithms[9]. It also offers integrated fault tolerance and the scalability to scale horizontally on hundreds or even thousands of nodes and is thus well suited to process petabytes of data efficiently. Laid on top of the core of Spark is its scalable machine learning library, the MLlib. Its main function is to provide a wide variety of algorithms that can run efficiently on distributed datasets. One crucial evolution of the MLlib is its DataFrame-based API, the principal interface today. The API provides Pipelines, a higher-level API for composing and tuning machine learning work flows[10][11]. Pipelines combine many transformations (e.g., creating features) and estimators (e.g., training models) together in a single, end-to-end holistically-defined workflow. This design makes building simpler, promotes reusability of the code and makes use of the Spark Catalyst optimizer for quicker performance. The MLlib is the clear choice in big-scale complaint classification with its ability to train complex models on large numbers of customer commentary that would overwhelm the more standard single-machine libraries.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. AUTOMATED COMPLAINT CLASSIFICATION

Automated complaint classification harnesses machine learning and large language models (LLMs) to efficiently categorize consumer grievances [13]. Recent advances in zero-shot learning enable models like GPT-4 and Claude to classify financial complaints without labeled training data, making them ideal for dynamic environments [15]. Reasoning models, enhanced through reinforcement learning, bring structured decision-making and deep inference to complex text classification [15]. Studies have shown that traditional methods like SVM and logistic regression still perform competitively on structured datasets [13], [14]. Platforms that incorporate Latent Dirichlet Allocation (LDA) uncover complaint themes and predict company responses, supporting both consumers and regulators [14]. Additionally, hybrid frameworks using review-based control charts and dynamic importance–performance analysis provide real-time monitoring of complaint trends [12]. These methods address imbalanced data and shifting customer concerns over time [12]. Integrating LLMs, predictive models, and statistical techniques enhances scalability and accuracy in complaint resolution workflows [14], [15]. As data volume grows, automated systems are critical for proactive service quality management. This marks a transformative step in intelligent, real-time consumer complaint processing.

5. METHODS

The approach employs a highly effective multi-stage PySpark machine learning pipeline for text classification tasks, in this instance, to classify the Amazon reviews as "Highly Dissatisfied" or "Mildly Dissatisfied" complaints. The operation starts with Data Ingestion putting the raw dataset into Spark. The text classification architecture design is a multi-stage machine learning pipeline built with PySpark. The operation commences with Data Ingestion and a comprehensive Data Preprocessing. The preprocessing is conducted in parallel on different data types: text features are put through tokenization, HashingTF, and IDF; categorical features are converted using StringIndexer and OneHotEncoder; and numerical features get scaled using the StandardScaler. All the thus-preprocessed features are combined together as a single, combined Feature Vector that serves as the normalized input to the machine learning models.

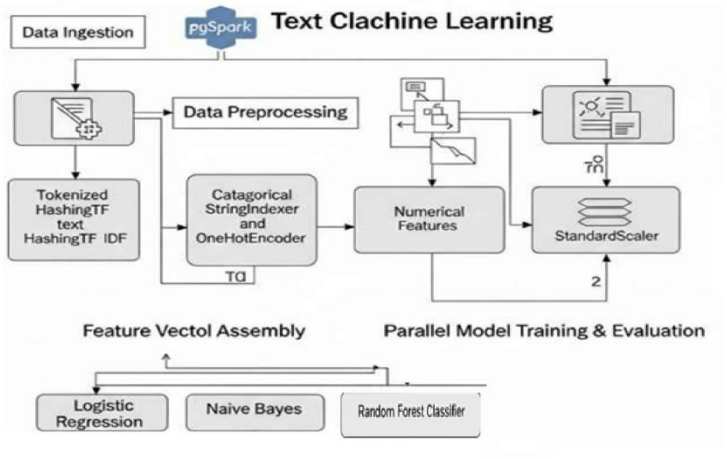

Figure 1. Machine Learning Pipeline for Comparative Model Evaluation

Parallel Model Training and Evaluation is the second step and is tasked with the comparison of the performance of four various PySpark MLlib classification models on the same set of data. The training of Logistic Regression, Naive Bayes, RandomForestClassifier, and LinearSVC is conducted on the ready-to-use feature vector. The performance of each model is tested individually and the performance of all the models is combined and compared based on crucial parameters like accuracy, precision, and recall. This design allows for a simple and efficient comparison of many algorithms together in order to determine the best-performing model for the classification task. This dataset contains the customer reviews from Amazon concerning the items of beauty products. The dataset is comprised of

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

text and meta-data fields such as rating, title, text, images, asin, parent_asin, user_id, timestamp, verified_purchase, helpful_vote, and year, all of which being nullable. For the purpose of modeling dissatisfaction, from the value of the rating the binary schema was determined as follows: the rating of 1 was marked as MajorComplaint (label 1), and the rating of 2 was marked as MinorComplaint (label 0). The resultant distribution constitutes 102,080 major complaints and 43,034 minor complaints and hence provides a total of 145,114 samples. The data was split into training and test parts with 115,928 samples and 29,186 samples being used for training and test purposes respectively while keeping the two-class category intact.

Complaint_Label	Title	Text
MinorComplaint	I'll stick to my 5-blade razor.	I think I need to stick with my 5 blade razor. This razor is not only difficult to hold & is made of a hard slick plastic with no grippers, but doesn't shave as well as my 5-blade razor. The style is nice, but the functionality is not.
MinorComplaint	Ineffective	A total waste of money. I get better results using a tea bag to alleviate under eye swelling.
MajorComplaint	halo hair extensions	This halo hair extension is simply put, garbage. Now, you get what you pay for. And this is a very cheap version. The faux hair is very shiny and looks literally like bad barbie hair. It looks WAY better in the photos than in real life. The color is horrific, in my opinion of course. The streaks are like paint strips. And all of that would be one thing - but the worst is that the hair completely fell out! I had hand fulls of hair strands just trying to put the halo on! And you might think - well, maybe a little loss is to be expected? Except this was handfuls and handfuls. I literally dropped the whole thing right into the trash. I would say this one is a pass for hair loss alone Having said all of this, I never hesitate to update my reviews should new info seem useful. All of my reviews reflect my honest, personal experience with the reviewed item - your experience may be different. I am not influenced by any outside source. I receive/accept NO free products or discounts that are not available to all shoppers- ever. For some reason our shopper ranks are no longer visible - so, to give you a little more info about me, I am a top 50 reviewer (#30 the highest rank achieved). Those numbers used to fluctuate over time - up and down but I noticed that they stopped updating regularly - perhaps to phase them out. It's a shame because it did help you see who has been around the longest and who is a trustworthy reviewer. Â I've been doing reviews for over 25 years with Amazon - over 6,000 reviews posted, those reviews have been viewed well over 50,000 times, including well over 25,000 likes. Bottom line, I pay for all my stuff, just like you do.
MajorComplaint	Donâ€™t bother	Normally I like NYX products but not this one. It leaves a funny coating on your eyebrows and pushes them down so much they look thinner than they are Not worth it.

Figure 2. Sample Cuetomer Complaints with Assigned Labels

6. RESULTS AND DISCUSSION

This section presents the comparison and the performance evaluation of the trained machine learning models for the classification task. The overall objective of the assessment was to test the performance of three algorithms—Naive Bayes, Logistic Regression, and RandomForestClassifier—to classify the data. Their performance was rigorously tested using an exhaustive set of metrics including overall accuracy, macro F1-score, and weighted measures for recall and precision. These parameters play a significant role in getting a complete understanding of the predictive ability of each model, its confidence level in making positive class assignment, and its ability to show balanced performance for all the classes. The results tabulated in the subsequent table and explained in detail will guide the selection of the best model for this specific application.

model	accuracy	f1-score_macro	precision_weighted	recall_weighted
NaiveBayes	0.6805	0.6884	0.7022	0.6805
LogisticRegression	0.6745	0.6632	0.6568	0.6745
RandomForestClassifier	0.6978	0.5736	0.4869	0.6978

Table1 : Performance Metrics of Machine Learning Models

Based on the performance measures listed above, this table displays the summary outcome of a classification test with the use of three machine learning models: Naive Bayes, Logistic Regression, and Random Forest. The performance measures applied are overall accuracy, macro F1-score, and weighted recall and precision. The higher the value for each measure applied, the higher the model performance. The above statistical test is aimed at finding the most effective of the above models in predicting the data based on the above measures.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

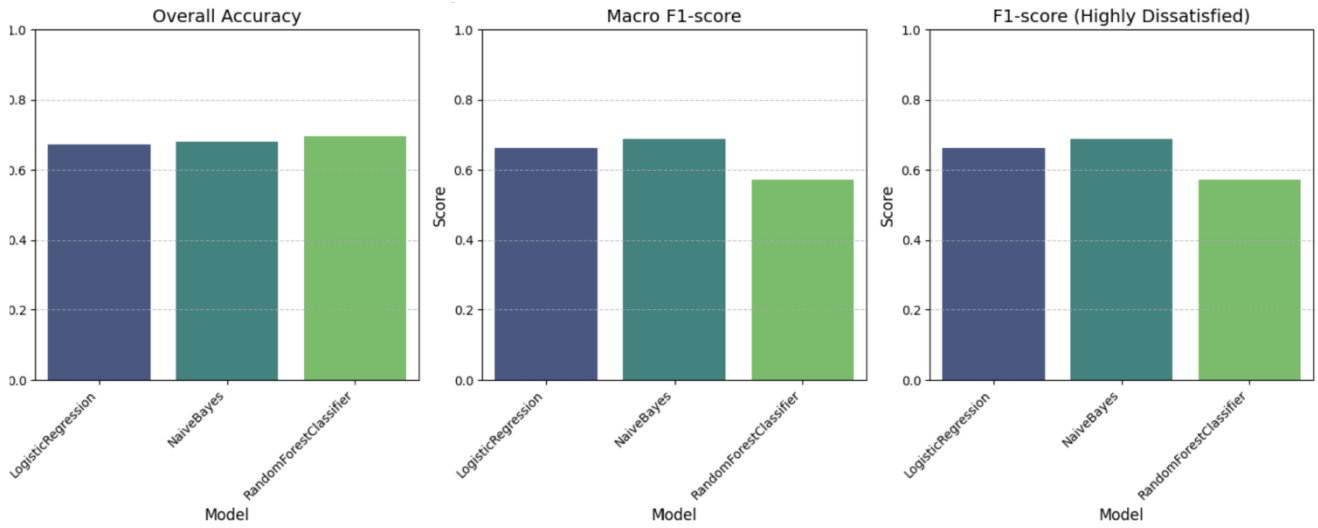

Figure 3. Performance Comparison of Spark MLlib Classifiers for Compalaint Classification

Naive Bayes has the most balanced performance all-round, obtaining the best f1-score_macro (0.6884) and the best precision_weighted (0.7022). This stems from the fact that it strikes the best balance between recall and precision for all the classes and therefore its results are reliable and predictable. In contrast, while the Random Forest classifier achieves the highest average accuracy (0.6978) and recall_weighted (0.6978), it obtains the worst precision_weighted (0.4869) and f1-score_macro (0.5736). This is a suggestion that the Random Forest is highly accurate making positive case predictions but often so much so to the disadvantage of falsely predicting negative cases and thus ends up having many false positives. Logistic Regression achieves the worst performance on all the most important performance metrics. The best model from the three models to use for a well-balanced and reliable classifier that has the most balanced performance on all the classes and metrics tested is therefore Naive Bayes. Apart from raw counts, choosing the best model depends on the trade-off between recall and precision in a complaint management system. While Random Forest shows high accuracy and recall, its low precision leads to many false positives. This burdens human agents with filtering non-complaints. In contrast, Naive Bayes offers better precision and a balanced F1-score, reducing false positives. Thus, model selection becomes both a statistical and strategic decision.

7. CONCLUSION

This work managed to evaluate the performance of the Naive Bayes, LogisticRegression, and RandomForestClassifier machine learning algorithms on a large-scale complaint classification task using the Spark MLlib. The experiment revealed that while the overall accuracy corresponding to the RandomForestClassifier was the highest achieved, its performance was unbalanced and sufferered from low precision that suggests the prevalence of a vast number of false positives. In opposition, the performance of the Naive Bayes model was the most balanced and reliable with the optimal macro F1-score and weighted precision. This balanced performance proves most critical to practical CRM systems as it ensures a very high level of reliability in the identification and classification of complaints. Thus, from the above results, we conclude that the most suitable and efficient model from the considered algorithms for the specific classification task is the algorithm of Naive Bayes.

Some potential avenues for future work also have the potential to further boost the performance of complaint classification systems. An important avenue in this category is to consider the impact of higher-level text preprocessing techniques, e.g., the use of stemming and lemmatization, to potentially standardize words and boost the generality of the models. The performance of higher-level machine learning models such as deep architectures and higher-level gradient boosting models such as XGBoost or LightGBM should also be explored. Of further value would be the performance of a more intensive hyperparameter tuning process to further optimize the models.

Finally, the models can be tested on a larger and wider set of complaint data, potentially in the form of a streaming data pipeline, to test the scalability and generality of the models over a variety of domains and industries.

REFRENCES

[1] W. K. R. Perera, K. A. Dilini, and T. Kulawansa, 'A Review of Big Data Analytics for Customer Relationship Management', in 2018 3rd International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka: IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/ICITR.2018.8736131.

[2] P. Del Vecchio, G. Mele, E. Siachou, and G. Schito, 'A structured literature review on Big Data for customer relationship management (CRM): toward a future agenda in international marketing', IMR, vol. 39, no. 5, pp. 1069–1092, Oct. 2022, doi: 10.1108/IMR-01-2021-0036.

[3] P. Zerbino, D. Aloini, R. Dulmin, and V. Mininno, 'Big Data-enabled Customer Relationship Management: A holistic approach', Information Processing & Management, vol. 54, no. 5, pp. 818–846, Sep. 2018, doi: 10.1016/j.ipm.2017.10.005.

[4] M. Al-Bashayreh, D. Almajali, M. Al-Okaily, R. Masa'deh, and A. Samed Al-Adwan, 'Evaluating Electronic Customer Relationship Management System Success: The Mediating Role of Customer Satisfaction', Sustainability, vol. 14, no. 19, p. 12310, Sep. 2022, doi: 10.3390/su141912310.

[5] Mohammed. T. Nuseir, A. I. Aljumah, and G. A. El Refae, 'Impact of Big Data Analytics and Managerial Support on CRM: Exploring Mediating Role of Marketing Analytics', in 2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS), Milan, Italy: IEEE, Nov. 2022, pp. 1–8. doi: 10.1109/IOTSMS58070.2022.10062100.

[6] E. O. Nogueira and M. Borchardt, 'The effects of customer relationship management (CRM) on e-commerce evolution: A systematic review', TSSJ, vol. 36, pp. 433–453, Oct. 2022, doi: 10.47577/tssj.v36i1.4124.

[7] S. Tang, B. He, C. Yu, Y. Li, and K. Li, 'A Survey on Spark Ecosystem for Big Data Processing', IEEE Trans. Knowl. Data Eng., pp. 1–1, 2020, doi: 10.1109/TKDE.2020.2975652.

[8] A. Ali El-Sayed and D. Khalil Ibrahim, 'Big data resolving using Apache Spark for load forecasting and demand response in smart grid: a case study of Low Carbon London Project', Journal of Big Data (2024) https://doi.org/10.1186/s40537-024-00909-6

[9] M. Guller, Big Data Analytics with Spark. Berkeley, CA: Apress, 2015. doi: 10.1007/978-1-4842-0964-6.

[10] M. Mezati and I. Aouria, ' Flink-ML: machine learning in Apache Flink' Brazilian Journal of Technology (2024) ISSN: 2595-574 DOI:10.38152/bjtv7n4-015

[11] A.A Shehloo and G.G Varshney , ' Realizing the Potential of Big Data Analytics through Apache Spark MLlib', Nanotechnology perceptions 20 N° S14 (2024) 1813-1830 .

[12] S. Kim and M. Kwak, 'Customer Complaint Analysis via Review-Based Control Charts and Dynamic Importance–Performance Analysis', Applied Sciences, vol. 13, no. 10, p. 5991, May 2023, doi: 10.3390/app13105991.

[13] G. Alarifi, M. F. Rahman, and M. S. Hossain, 'Prediction and Analysis of Customer Complaints Using Machine Learning Techniques':, International Journal of E-Business Research, vol. 19, no. 1, pp. 1–25, Mar. 2023, doi: 10.4018/IJEBR.319716.

[14] D. Vaishnav, M. Neethinayagam, A. Khaire, and J. Woo, 'Predictive Analysis of CFPB Consumer Complaints Using Machine Learning', Jul. 08, 2024, arXiv: arXiv:2407.06399. doi: 10.48550/arXiv.2407.06399.

[15] R.Bai, Ch. Liu, L. Yuan, X. Zhu, Y.Zhou, H. Yan and X. Ouyang.'Chinese Complaints Text Classification method based on pre-trained Language Model' July 2024 3rd International Conference on Electronic Information Engineering and Data Processig (EIEMP) 2024 /https://doi.org/10.1117/12.3033025