EasyChair Preprint

№ 12557

Email Classification and Routing Using Machine Learning

Yavnik Sharma, Darshan Thakar, Dipesh Majithia and Chaithanya Kolan

March 18, 2024

AUTOMATIC EMAIL CLASSIFICATION AND ROUTING USING SUPERVISED MACHINE LEARNING

Sharma Yavnik D.
2000303124086
CSE Department
Parul Institute of Engineering and Technology

Thakar Darshan N.
2000303124092
CSE Department
Parul Institute of Engineering and Technology

Majithia Dipesh P.
2000303124055
CSE Department
Parul Institute of Engineering and Technology

Kolan Chaithanya
2000303124291
CSE Department
Parul Institute of Engineering and Technology

***Abstract*—Email has been the primary form of communication for businesses and their customers for a long time now. Be it B2B or B2C, companies rely heavily on email form of communication for internal as well as customer facing areas. A large company might receive thousands of emails per day for a vast range of customer queries such as updates and complaints. Each of these queries needs to be addressed by a specific department like Sales, HR, Marketing of the concerned company. For the ease of customer, companies tend to have a single customer facing email such as contact@comapanyname.com . This tend to make customer's work very easy as they can write up all their enquiries at a single email but managing this email is a cumbersome process for a company as an employee would have to manually look into each and every email and forward it to the relevant department. This part becomes a bottleneck in customer service as the rate of customer queries answered is limited to how many emails can be routed to concerned department. In a world full of Automation, this is still one of the big issues that has to be currently handled manually. In our project, we aim to automate this process by checking the content of the incoming email using a supervised machine learning model and classify the query for which department it belongs such as Sales, HR, Marketing etc. and then automatically routes the incoming emails to their concerned department. All without any human intervention. Such a system could easily handle the load of a large organization which might receive thousands of emails per day.**

Key words: Email, Classification, Supervised Learning, Routing, Automation.

I. INTRODUCTION

In this project, we aim to automate the routing process of customer emails in a large organization by automatically classifying them based on content and route the email to the relevant department. The machine learning model will be trained on labelled dataset of customer emails with the labels being the category to which the email belongs (Sales, HR, Marketing etc.).

A. Problem Statement

Any large organization might receive thousands of emails daily from customers with various queries. These queries need to be forwarded to relevant department in the company like Sales, HR, Marketing etc. Currently, this process is handled manually as an employee would read the content of the emails one by one and then forward them to concerned department. This method however is very time and cost ineffective and can prove to be a bottleneck in customer service. The number of queries resolved by the company comes down to how many queries reach the concerned department.

B. Scope

Using supervised machine learning techniques and Natural language processing (NLP), we are going to train a model through a labelled dataset of customer emails with the labels being the category to which the email belongs (Sales, HR, Marketing etc.). This model will then be tested against unforeseen data to verify its real-world accuracy. Once the model is developed, it will be deployed live and email router would be setup in a way that the model receives all incoming emails to the organization's customer facing email address and routes it to relevant department's email address. This router can additionally work as load balancer in selectively sending emails to customer service agents based on their work load.

C. Aim and Objective

Our aim for this project is to automate the email routing process in a large organization, a process which is currently purely manual. We aim to achieve significant increase in productivity and number of queries solved by the company. This would result in faster response time and improved customer service. A current bottleneck in customer service process would be eliminated.

II. EXPERIMENTAL SETUP AND METHODOLOGY

In this project, we'll be building a supervised machine learning model using techniques like Natural language processing (NLP). We will train a model through a labelled dataset of customer emails with the labels being the category to which the email belongs (Sales, HR, Marketing etc.). This model will then be tested against unforeseen data to verify its real-world accuracy. Once the model is developed, it will be deployed live and email router would be set up in a way that the model receives all incoming emails to the organization's customer facing email address and routes it to relevant department's email address. This router can additionally work as load balancer in selectively sending emails to customer service agents based on their workload.

A. Data set

The dataset we'll use would be a labelled dataset containing labels of the category to which the email should belong such

as Sales, HR, Marketing etc. This dataset would contain the subject of the emails, their content and their intended category. Since we'll be performing classification, we need labelled dataset for supervised machine learning. These datasets would be divided into multiple datasets in order to perform Cross-Dataset validations and address future variables.

Sr No	Subject	Content	Category
1	Issue with current month's bill	There has been an issue with my current month's bill as I've been overcharged twice for...	Billing Dept.
2	What is the current price for Product X	I'm interested in purchasing Product X and I would like to know...	Sales Dept
3	Vacancy to join as a Distributor for.	I would like to join your company as a distributor...	HR Dept

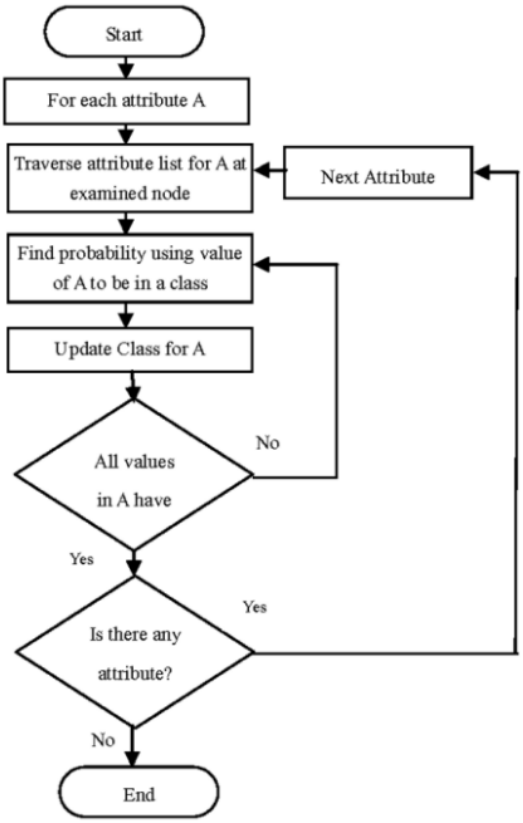
B. Training and Testing Models

Multiple models will be trained with various algorithms and later these models would be tested with unforeseen data to test accuracy and to identify the most optimal algorithm for such a use case. The criteria to select optimal algorithm would be the accuracy score, time taken to infer an email as parameters to find which model should be used in live deployment. Testing would take place with Cross-dataset testing experiments as this would provide the best insights on how the model works with unforeseen datasets and could account for future variables. If the email is forwarded to the wrong department considering a false positive, this could cause inefficiency as further forwarding would be required. To address this, if the confidence score of the model is above a certain threshold only then should the email be forwarded to relevant department or else it should be marked as un-classifiable and would then be manually classified by a person and this would be used to retrain the model. Thus, the model should also have feedback loop capabilities.

C. Classification

Incoming emails will be routed through our machine leaning model and they will be classified as to
which category and relevant department the customer's email should go to. Since there are multiple
categories and the email could belong to one of them, we would require an algorithm that supports
multi-class classification. Some algorithms that support multi-class classification are as follows:

- Na¨ıve Bayes Classification
- K-nearest Neighbours (KNN)
- Support Vector Machines (SVM)
- Decision Trees



Support vector machine: A Support Vector Machine (SVM) is a sophisticated machine learning algorithm that is used for classification and regression tasks. It works by determining the best hyperplane for separating data points into distinct classes while maximizing the margin between them. SVMs are well-known for their ability to handle high-dimensional data and to handle both linear and non-linear problems using kernel functions. They perform well in scenarios with clear class boundaries and are resistant to overfitting. SVMs have been used in a variety of fields, including image recognition, text classification, and bioinformatics, making them a versatile and widely used tool in machine learning and data analysis

Decision tree: Decision Trees are a popular machine learning algorithm that can be used for classification as well as regression. They represent a hierarchical structure, with nodes representing features and branches representing possible decisions. The tree is built by recursively partitioning data based on the most informative features, with the goal of minimizing impurity while maximizing information gain. Decision Trees are understandable and useful for feature selection. However, they are susceptible to overfitting, particularly when dealing with complex data. Techniques such as pruning and ensemble methods such as Random Forests help to mitigate this problem. Because of their simplicity and interpretability, decision trees are widely used in a variety of fields, including finance, healthcare, and natural language processing.

KNN: K-Nearest Neighbors (KNN) is a straightforward machine learning algorithm that is used for classification and regression tasks. It works by locating the K nearest data

points (neighbors) to a given query point and then predicting based on the majority class or average value of those neighbors. KNN's strength lies in its adaptability to complex data distributions. However, it is sensitive to the choice of K and requires a large amount of training data. KNN is used in recommendation systems, pattern recognition, and anomaly detection, and its simplicity and effectiveness make it a useful tool in a wide range of machine learning applications.

Naive Bayes: Bayes' Theorem It employs Bayes' theorem to determine the likelihood of a data point belonging to a specific class based on the conditional probabilities of its features. The "naive" assumption in Naive Bayes is that features are independent of one another, which simplifies calculations. Despite this simplification, Naive Bayes frequently outperforms other methods in practice and is computationally efficient. It's especially useful for large datasets and real-time applications. However, if the independence assumption is significantly violated, its performance may suffer. Sentiment analysis, document classification, and email filtering all use Naive Bayes.

Logistic regression: Logistic Regression is a statistical model that may be used for binary classification as well as multi-class classification with minor modifications. It predicts probabilities between 0 and 1, as opposed to linear regression, by mapping input characteristics to a logistic function. If the probability passes a certain threshold (typically 0.5), the data item is assigned to one of the two groups. Logistic Regression is an interpretable and effective tool in machine learning and statistics. It's utilized in a variety of applications, including medical diagnosis, spam identification, and credit risk assessment, where understanding the link between independent factors and a binary outcome is critical for making decisions.

Random forest: A Random Forest Classifier is a strong ensemble learning algorithm that is commonly utilized in classification applications in machine learning. During training, it builds numerous decision trees, each based on a subset of the data and a random selection of characteristics. After that, the final prediction is formed by aggregating the various tree forecasts by voting (for classification) or averaging (for regression). Random Forests are resistant to overfitting, can handle high-dimensional data, and can prioritize feature significance. They perform well in complicated, noisy datasets and are easy to comprehend. Random Forests are used for successful predictive modeling in a variety of industries, including banking, healthcare, and image analysis, because to their adaptability and accuracy.

Adaboost classifier: AdaBoost, an abbreviation for Adaptive Boosting, is a common ensemble learning technique for binary classification and, with expansions, multi-class problems. It trains a series of weak learners (typically decision trees) incrementally, assigning more weight to previously misclassified data points. This iterative method allows AdaBoost to concentrate on the most difficult cases, boosting overall accuracy. It makes the final forecast by combining the predictions of the weak learners using weighted majority voting. AdaBoost is well-known for its capacity to handle complicated data and frequently outperforms in practice. It has applications in face identification, text categorization, and a variety of other fields, making it an important machine learning tool.

Bagging Classifier: BaggingClassifier, which stands for Bootstrap Aggregating, is a machine learning ensemble approach. By training numerous models on random subsets of the training data, it improves the performance of basic learning algorithms, often decision trees. Each model provides its own prediction, which the BaggingClassifier aggregates by voting (for classification) or averaging (for regression), lowering the risk of overfitting and enhancing overall model accuracy. Bagging works well for minimizing variation, boosting stability, and improving generalization. It is frequently used in applications such as random forests, which aggregate many decision trees to create robust and accurate predictions in a variety of machine learning tasks.

Extra Trees Classifier: The ExtraTreesClassifier, short for Extremely Randomized Trees Classifier, is a machine learning ensemble learning algorithm. It is similar to the Random Forest method in that it creates numerous decision trees during training. What distinguishes ExtraTrees is its extraordinary unpredictability in the tree-building process. It chooses random subsets of characteristics and divides nodes using random thresholds, making it quicker and less prone to overfitting. ExtraTrees tries to increase model generalization while maintaining high predicted accuracy by incorporating more randomization. It is especially beneficial for high-dimensional data and complicated classification issues, where it may significantly reduce variance and improve performance.

Gradient Boosting Classifier: The Gradient Boosting Classifier is a sophisticated machine learning technique that may be used for classification as well as regression. It belongs to the ensemble learning family and operates by training a sequence of weak learners, often decision trees, in a sequential manner, with each successive tree correcting the errors of the prior ones. The approach optimizes for the gradient of the loss function, hence the name "gradient boosting," in order to reduce prediction errors and make correct predictions. GradientBoosting is well-known for its excellent prediction accuracy, resistance to overfitting, and ability to capture complicated data patterns. Because of its remarkable performance, it is frequently employed in a variety of sectors, including banking, healthcare, and natural language processing.

XGB Classifier: The Extreme Gradient Boosting Classifier, or XGB Classifier, is a strong and efficient machine learning technique. It is a gradient boosting framework implementation that has been particularly built to enhance performance and computing speed. XGBoost employs techniques such as regularization and parallel processing to improve model accuracy while lowering the risk of overfitting. It is extremely adaptable, handling a wide range of loss functions and enabling both classification and regression applications. Because of its remarkable prediction performance, scalability, and capacity to work with big datasets, XGBoost is a popular option in machine learning contests and real-world applications. It is regarded as a must-have tool for data scientists and machine learning practitioners.

Multinomial Naive Bayes: Multinomial NB, also known as Multinomial Naive Bayes, is a machine learning technique

that is mostly used for text classification tasks including document categorization and spam email detection. It is a Naive Bayes variation that is well-suited for dealing with data containing discrete characteristics, such as word frequencies in text texts. MultinomialNB uses the multinomial distribution to describe the probability distribution of feature occurrences in each class, making it particularly useful for text data. Despite its ”naive” assumption of feature independence, MultinomialNB frequently performs well and is computationally efficient. Because of its simplicity and efficacy, it is commonly used in natural language processing and text mining jobs.

D. Deployment Setup

For live deployment, our model could be installed on a cloud-based service or on premises. For the setup, all incoming emails to customer facing email address of the company should be routed towards our setup first. This can be done by either forwarding via inbuilt email mechanisms or via fetching all emails through domain's MX records.

The setup would be feeded with emails of all the departments of the organization and when an email is classified for a particular department, it can be internally forwarded. Along with email routing, this setup has an additional benefit of being used as a load balancer among customer service agents inside a department. The service would keep note of the number of emails that have been sent to every agent and could load balance among them as sending next email to the agent with least workload. This would further increase efficiency and reduce response time.

III. Future work and conclusion

A. Future work and Conclusion

In machine learning, we used various classifiers such as support vector machine, naive bayes, decision trees, extra tree classifier, multinomial naive bayes, XGB CLASSIFIER, gradient boosting classifier, bagging classififer, adaboost classififer, random forest, logistic regression, naive bayes, knn. We discovered the precision and accuracy after training the model with these approaches. Its support vector classifier has an accuracy of 0.92 and a precision of 0.92 out of all methods. We used SVM for email categorization and trained and evaluated the model. If we offer input depending on the parameters we specify, we get the precise outcome, which is that the email is classified into the specified department.

There are some more possibilities for altering the workflow of the project and additional features to be implemented in order to improve upon the goal of increased efficiency, response rate and overall customer experience. One of the important features to be added in future is feedback loop. If an email has been wrongly classified into a category to which it doesn't belong or it has not been able to predict with a high enough confidence, such cases would be marked manually by a person, and these would then act as feedback to the model so it can learn from example and improve its accuracy and confidence number.

The deployment is aimed to be as seamless as possible so as to not interfere much with existing email and communication systems of an organization but to mold and work along with existing systems. Thus, techniques like catching email via MX records of the domain have been implemented so existing email infrastructure of a company need not to be touched. Overall, such a system could play a massive role in any medium or large-scale organization which could eliminate bottlenecks in customer service and thus enable the full potential of their employees to be used in serving the needs of customers.

Faster response time, improved peremployee response and overall increased efficiency are some of the invaluable things any business would strive to achieve, and this project aims to provide exactly this to their users.

References

[1] Balamurugan, Suganya and Rajaram, Revathi and Govindaraj, Athiappan and Muthupandian, M., 2023, DATA MINING TECHNIQUES FOR SUSPICIOUS EMAIL DETECTION; A COMPARATIVE STUDY

[2] Ssebulime, Timothy. (2022). Email Classification Using Machine Learning Techniques

[3] Y awen, Wang and Fan, Yu and Yanxi, Wei. (2018). Research of Email Classification based on Deep Neural Network. International Journal of Advanced Network, Monitoring and Controls. pp 17-21.

[4] Tang, X., Mou, H., Liu, J., 2021, Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. Sci Rep 11, 11849.

[5] . Kumara B, Aruna and Kodabagi, Mallikarjun Kodabagi and Choudhury, Tanupriya and Um, Jung-Sup. (2021). Improved email classification through enhanced data preprocessing approach. Spatial Information Research

[6] Morales, Valentin and Gomez, Juan Carlos and Amerongen, Saskia. (2020). Cross-dataset email classification. Journal of Intelligent and Fuzzy Systems

[7] Kumar, Devendra. (2019). EMAIL CLASSIFICATION USING ARTIFICIAL NEURAL NETWORK

[8] Ramaraj, N., 2007, Automated Classification of Customer Emails via Association Rule Mining, Information Technology Journal. 6

[9] Zeng, Chao and Lu, Zhao and Gu, Juzhong. (2009). A New Approach to Email Classification Using Concept Vector Space Model.

[10] T.A Meyer, B Whateley (2004) SpamBayes: Effective open-source, Bayesian based, email classification system