# SMS Classification using NLP concepts NLP project Report

Chetana N Patil
Computer Science Engineering
PES University,EC campus,Bangalore
Email: chetananpatil2002@gmail.com

Mekala Sanjana
Computer Science Engineering
PES University,EC campus,Bangalore
Email: sanjana.mekala135@gmail.com

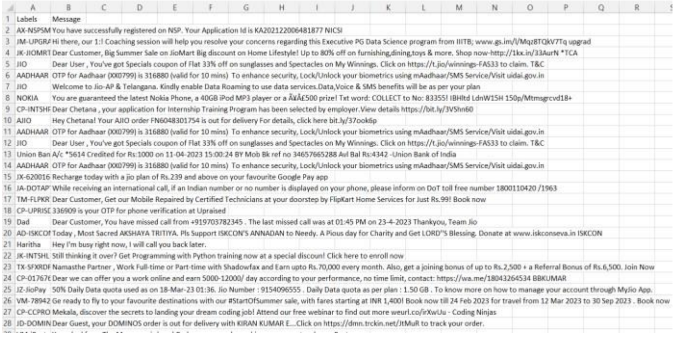
*Abstract*—Natural language processing (NLP) message classification is a sophisticated technology that uses machine learning algorithms to automatically sort text messages into predetermined classes. Message classification using NLP can help firms save time and resources while making better judgments by automating and streamlining text analysis operations. Message classification using NLP is becoming an increasingly significant tool for organizations and researchers as the volume of digital data grows.

## I. Introduction

SMS (Short Message Service) is a common means of communication used by trillions of people throughout the world. With the growing volume of SMS data created every day, there is an increasing need to classify these messages for use in applications such as spam filtering, sentiment analysis, and customer feedback analysis. Natural Language Processing (NLP) and machine learning techniques have been shown to be effective in text classification tasks and can also be used to classify SMS. The purpose of this study is to investigate the use of NLP and machine learning approaches for SMS categorization, examine their efficacy, and compare various models to choose the best performing one. The suggested method includes preprocessing SMS data to eliminate noise, feature extraction, and text representation using numerical vectors. Various machine learning algorithms, including Naive Bayes, Decision Trees, Support Vector Machines, and Neural Networks, will be trained and evaluated using a variety of performance metrics, including accuracy, precision, recall, and F1 score. The outcomes of the experiments will be examined, and the most effective model for SMS categorization will be identified. This paper will also examine the limitations and challenges of SMS classification and make recommendations for further research.

## II. Dataset Analysis

Datasets are vital and should be carefully chosen for proper machine learning model training and testing.Furthermore, to evaluate the performance of the machine learning models, the dataset must be divided into training, validation, and testing sets. The training set is used to train models, the validation set is used to tune hyperparameters and optimize the model, and the testing set is used to assess the model's performance on unknown data.The dataset has been divided into two sets, a training dataset and testing dataset which we have created it on our own but with fewer data. Here we created our own dataset with different message labels like Jio, Upgrad, AJIO, Internshala etc. Dataset basically contains 2 columns i.e, Labels, message.



| | A | B |
|---|---|---|
| 1 | Labels | Message |
| 2 | AX-NSPSM | You have successfully registered on NSP. Your Application Id is KA202122006481877 NICSI |
| 3 | JM-UPGRA | Hi there, our 1:I Coaching session will help you resolve your concerns regarding this Executive PG Data Science program from IIITB; www.gs.im/l/Mqz8TQkV7Tq upgrad |
| 4 | JK-JIOMRT | Dear Customer, Big Summer Sale on JioMart Big discount on Home Lifestyle! Up to 80% off on furnishing,dining,toys & more. Shop now-http://1kx.in/33AurN *TCA |
| 5 | JIO | Dear User , You've got Specials coupon of Flat 33% off on sunglasses and Spectacles on My Winnings. Click on https://t.jio/winnings-FAS33 to claim. T&C |
| 6 | AADHAAR | OTP for Aadhaar (XX0799) is 316880 (valid for 10 mins) To enhance security, Lock/Unlock your biometrics using mAadhaar/SMS Service/Visit uidai.gov.in |
| 7 | JIO | Welcome to Jio-AP & Telangana. Kindly enable Data Roaming to use data services.Data,Voice & SMS benefits will be as per your plan |
| 8 | NOKIA | You are guaranteed the latest Nokia Phone, a 40GB iPod MP3 player or a Ã¥Â£500 prize! Txt word: COLLECT to No: 83355! IBHltd LdnW15H 150p/Mtmsgrcvd18+ |
| 9 | CP-INTSHF | Dear Chetana , your application for Internship Training Program has been selected by employer.View details https://bit.ly/3V5hn60 |
| 10 | AJIO | Hey Chetana! Your AJIO order FN6048301754 is out for delivery For details, click here bit.ly/37ook6p |
| 11 | AADHAAR | OTP for Aadhaar (XX0799) is 316880 (valid for 10 mins) To enhance security, Lock/Unlock your biometrics using mAadhaar/SMS Service/Visit uidai.gov.in |
| 12 | JIO | Dear User , You've got Specials coupon of Flat 33% off on sunglasses and Spectacles on My Winnings. Click on https://t.jio/winnings-FAS33 to claim. T&C |
| 13 | Union Ban | A/c *5614 Credited for Rs:1000 on 11-04-2023 15:00:24 BY Mob Bk ref no 34657665288 Avl Bal Rs:4342 -Union Bank of India |
| 14 | AADHAAR | OTP for Aadhaar (XX0799) is 316880 (valid for 10 mins) To enhance security, Lock/Unlock your biometrics using mAadhaar/SMS Service/Visit uidai.gov.in |
| 15 | JX-620016 | Recharge today with a jio plan of Rs.239 and above on your favourite Google Pay app |
| 16 | JA-DOTAP | While receiving an international call, if an Indian number or no number is displayed on your phone, please inform on DoT toll free number 1800110420 /1963 |
| 17 | TM-FLPKR | Dear Customer, Get our Mobile Repaired by Certified Technicians at your doorstep by FlipKart Home Services for Just Rs.99! Book now |
| 18 | CP-UPRISE | 336909 is your OTP for phone verification at Upraised |
| 19 | Dad | Dear Customer, You have missed call from +919703782345 . The last missed call was at 01:45 PM on 23-4-2023 Thankyou, Team Jio |
| 20 | AD-ISKCOI | Today , Most Sacred AKSHAYA TRITIYA. Pls Support ISKCON'S ANNADAN to Needy. A Pious day for Charity and Get LORD''S Blessing. Donate at www.iskconseva.in ISKCON |
| 21 | Haritha | Hey I'm busy right now, I will call you back later. |
| 22 | JK-INTSHL | Still thinking it over? Get Programming with Python training now at a special discoun! Click here to enroll now |
| 23 | TX-SFXRDF | Namasthe Partner , Work Full-time or Part-time with Shadowfax and Earn upto Rs.70,000 every month. Also, get a joining bonus of up to Rs.2,500 + a Referral Bonus of Rs.6,500. Join Now |
| 24 | CP-01767E | Dear we can offer you a work online and earn 5000-12000/ day according to your performance, no time limit, contact: https://wa.me/18043264534 BBKUMAR |
| 25 | JZ-JioPay | 50% Daily Data quota used as on 18-Mar-23 01:36. Jio Number : 9154096555 . Daily Data quota as per plan : 1.50 GB . To know more on how to manage your account through MyJio App. |
| 26 | VM-78942 | Ge ready to fly to your favourite destinations with our #StartOfSummer sale, with fares starting at INR 1,400! Book now till 24 Feb 2023 for travel from 12 Mar 2023 to 30 Sep 2023 . Book now |
| 27 | CP-CCPRO | Mekala, discover the secrets to landing your dream coding job! Attend our free webinar to find out more weurl.co/irXwUu - Coding Ninjas |
| 28 | JD-DOMIN | Dear Guest, your DOMINOS order is out for delivery with KIRAN KUMAR E....Click on https://dmn.trckin.net/JtMuR to track your order. |

Fig. 1. Dataset overview

## III. Proposed Model

*a) The proposed SMS classification method employsmany phases, including data preprocessing, data cleaning, feature extraction, and model training and evaluation.:*

*b) The first stage of the proposed system is data preprocessing, which involves calculating the length and punctuations of the messages. This step helps to understand the structure of the data and identify any patterns or characteristics that may be useful for classification. By analyzing the length and punctuations of the messages, the system can determine if certain types of messages tend to be longer or use more punctuation than others.:*

*c) The second stage of the system is data cleaning, whichinvolves using stopwords and WordNetLemmatizer to remove irrelevant or redundant information from the text. Stopwords are common words that do not add any significant meaning to the text, such as ”the,” ”and,” or ”a.” WordNetLemmatizer is a technique used to reduce words to their base or root form, which helps to reduce the number of* *unique words in the text and improve the performance of the machine learning models.:*

*d) The third stage of the system is feature extraction,which involves using count vectorizer and Tf-idf vectorizer to represent the text as numerical vectors. Count vectorizer counts the number of occurrences of each word in the text, while Tf-idf vectorizer takes into account the frequency of the word in the document and the frequency of the word in the entire corpus. The result of feature extraction is a set of numerical vectors that represent the text, which can be used as input to the machine learning models.*

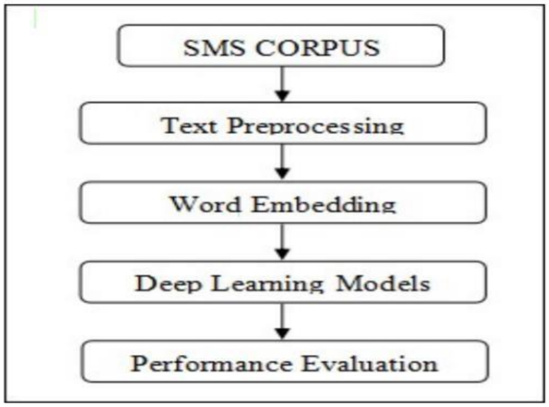


Fig. 2. High level diagram

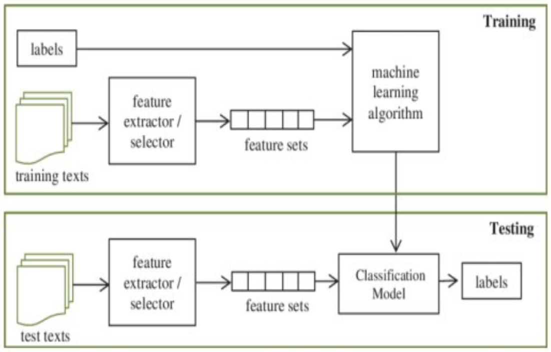


Fig. 3. Train and test model diagram

*e) The fourth stage of the system is model training and evaluation, which involves using naive Bayes and linearSVC machine learning models to classify the SMS messages. Naive Bayes is a probabilistic model that uses Bayes' theorem to calculate the probability of each class given the input features. LinearSVC is a linear classifier that tries to find a hyperplane that separates the classes in the feature space. Finally, the system uses evaluation metrics such as accuracy, precision, recall, and F1 score to measure the performance of the machine learning models. Accuracy measures the proportion of correctly classified messages, while precision measures the proportion of true positives among all positive predictions.*

*Recall measures the proportion of true positives among all actual positives, and F1 score is the harmonic mean of precision and recall.By using techniques such as count vectorizer and Tf-idf vectorizer and machine learning models such as naive Bayes and linearSVC, the system can effectively classify SMS messages and achieve high accuracy and precision. Linear SVM will give accurate prediction compared to Naive Bayes classification.*

## IV. ML ALGORITHMS USED

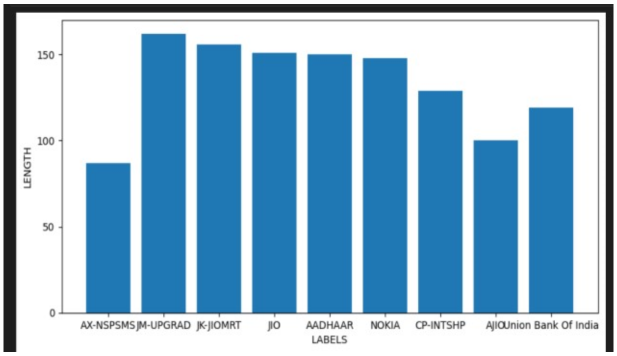
### *A. Naive Bayes Algorithm :*

Naive Bayes is a probabilistic algorithm that is based on Bayes' theorem, which states that the probability of a hypothesis (in this case, a class label) given the evidence (the input features) is proportional to the probability of the evidence given the hypothesis and the prior probability of the hypothesis. Naive Bayes assumes that the input features are conditionally independent given the class label, which simplifies the calculations and makes it computationally efficient. In text classification tasks, Naive Bayes is often used with bag-of-words representations and performs well with small to moderate-sized datasets.

### *B. LinearSVC :*

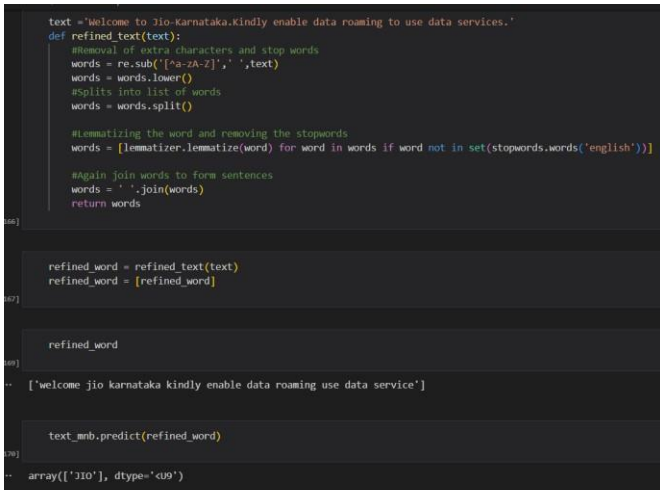
LinearSVC, on the other hand, is a linear classifier that attempts to find a hyperplane in the feature space that divides the classes. It is a popular choice for text classification tasks because it is based on the support vector machine (SVM) method. LinearSVC performs well with highdimensional datasets and is frequently employed with sparse representations like Tf-idf vectors. LinearSVC, unlike Naive Bayes, makes no assumptions about the distribution of input characteristics and can handle non-linearly separable classes using kernel functions.

Both Naive Bayes and LinearSVC have advantages and disadvantages, and the choice of method is determined by the task's specific requirements and data characteristics. LinearSVC is more difficult but can handle highdimensional datasets and non-linearly separable classes, whereas Naive Bayes is simple and efficient and performs well with small to moderate-sized datasets. Overall, both algorithms are successful for SMS categorization and have been proved in many studies to attain high accuracy and precision.



## V. Experimental Results

By using LinearSVC , we got the correct output for the new message. Compare to Naive Bayes algorithm, the LinearSVC performance is better. LinearSVC is variation of SVM algorithm.



```
    text ='Welcome to Jio-Karnataka.Kindly enable data roaming to use data services.'
    def refined_text(text):
        #Removal of extra characters and stop words
        words = re.sub('[^a-zA-Z]',' ',text)
        words = words.lower()
        #Splits into list of words
        words = words.split()

        #Lemmatizing the word and removing the stopwords
        words = [lemmatizer.lemmatize(word) for word in words if word not in set(stopwords.words('english'))]

        #Again join words to form sentences
        words = ' '.join(words)
        return words
166]

    refined_word = refined_text(text)
    refined_word = [refined_word]
167]

    refined_word
169]
··  ['welcome jio karnataka kindly enable data roaming use data service']

    text_mnb.predict(refined_word)
170]
··  array(['JIO'], dtype='<U9')
```

Fig. 4. Predicted output

## VI. Conclusion

SMS categorization is an essential task in natural language processing with numerous practical applications such as spam detection, sentiment analysis, and topic modeling. The application of machine learning techniques and NLP algorithms has enabled the development of efficient and accurate SMS classification systems.The system has been shown to achieve high accuracy and precision in various studies, and the choice of algorithm depends on the specific requirements of the task and the characteristics of the data. Naive Bayes is simple and efficient and can perform well with small to moderatesized datasets, while linearSVC is more complex but can handle highdimensional datasets and non-linearly separable classes. Overall, the development of an efficient and accurate SMS classification system using NLP and machine learning techniques involves careful consideration of the data preprocessing, feature extraction, and model selection stages. The proposed system has the potential to improve the accuracy and efficiency of SMS classification and can be applied in various practical scenarios such as spam detection, sentiment analysis, and topic modeling.

## Acknowledgment

We would like to thank Prof.Pooja Agarwal ma'am,Department of CSE, PES University,Bangalore for teaching us different techniques in NLP and for your guidance during project.

## References

[1] SMS Spam Classification–Simple Deep Learning Models With Higher Accuracy Using BUNOW And GloVe Word Embedding. Surajit Giri, Sayak Das, Sutirtha Bharati Das, and Siddhartha Banerjee Department of Computer Science, Ramakrishna Mission Residential College, Narendrapur, West Bengal, India. Pulication: 2022.

[2] Multiclass SMS Message Categorization: Beyond Spam Binary Classification. Fatia Kusuma Dewi, Mgs. M. Rizqi Fadhlurrahman, Mohamad Dwiyan Rahmanianto, Rahmad Mahendra Faculty of Computer Science, University of Indonesia. Publication: 2017

[3] SMS Text Classification Model Based On Machine Learning. XIAO FEI, LI JIANPING, GAO YUAN, ZHOU YUE,School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu. Publication: 2021

[4] A Study on Analysis of SMS Classification Using Document Frequency Thresold. N. Cristianini, and J.Shawe-Taylor,Support Vector and Kernel Methods, Intelligent Data Analysis: An Introduction Springer– Verlag, 2003

[5] A comparative study of word embedding techniques for SMS spam detection. P. Joseph and S. Y. Yerima, "A comparative study of word embedding techniques for SMS spam detection," 2022