**Cyclistic Bike-share Analysis for Targeted Casual-users marketing**

**Tamojit Roy1**

*1Vellore Institute of Technology, Chennai, India*

*1E-mail:* [*tamojit.roy2019@vitstudent.ac.in*](mailto:tamojit.roy2019@vitstudent.ac.in)

**Abstract**

Although the age calls for motor-vehicles as the major shareholders of the commute industries, cyclists still continue to have impressive shares in some of the developing heavyweights and developed nations around the world. Some of the countries like Sweden, Denmark, Germany, UK, Japan and even China have a wonderful landscape for cycling. Although some of them are casual riders, quite a many are annual members for the major cycle manufacturing companies all over the world. Data analytics plays an important role in analyzing and boosting the sales of any company. It has an upper hand when it comes to implementing market plans to target the set of customers who are the most vulnerable by suggesting them various specialized schemes and membership benefits. Hence, it is no doubt one of the most potent tools helpful in boosting the sales of a product. Hence, it can play a pivotal role in increasing the annual membership of cycles of any company for good profits. This research work intensely focusses on analyzing all the major aspects and most of the if not all of the attributes of bike-share sales of a prominent bike-share company in Chicago named Divvy. This research work mainly revolves around understanding how subscribers and customers of Divvy bike-share service use bikes differently. The comparison along with other tasks have been used to design marketing strategies aimed at converting customers of the company to the subscribers of its services.

**Keywords**

Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, Random Forest, Voting Classifier, F1 Score, Passive Aggressive Classifier, Count Vectorizer, Term Frequency-Inverse Document Frequency (TF-IDF)

**1. Introduction**

In today's Internet world, people rely on different online services / platforms for news. As the use of social media platforms such as Facebook, Twitter, and news websites increases, news spreads rapidly to millions of people in a short period of time. News websites publish news and provide authentication sources. The problem is how to authenticate messages and articles distributed on social media such as WhatsApp groups, Facebook pages, Twitter, other microblogging and social networking sites. Believing in rumors of pretending to be news is bad for society. The spread of fake news has widespread consequences, including the generation of biased opinions that affect the outcome of elections in favor of a particular candidate. In addition, spammers use compelling headlines to monetize their ads through Clickbait. Especially in developing countries like India, it takes time to put an end to rumors and focus on the right certified news articles. This white paper presents models and methods for detecting fake news. Social media platforms such as Facebook, Instagram, and Twitter provide a cheaper way to deliver news online much faster by spreading the network more easily. It's better than traditional news media, but it's a lot of fake news. News articles contain intentionally false information created online for a variety of purposes for economic and political gain. Therefore, there is an urgent need for a fake news detection system that not only distinguishes between fake news and real news, but also displays the relevant real news articles that are closest to the original. Since dealing with fake news requires precision and with current advances in machine learning algorithms to detect such fake news, it would be foolish to depend on a single algorithm/method to perform such complex classification. Therefore, it makes much more sense to divide the work into phases to completely separate the work by data mining operations such as data collection, data pre-processing, feature extraction, feature selection and implementation of machine learning models to perform predictions to classify the news as True or False and also predict the probability that the news belongs to the predicted label. Several machine learning models will be used to classify news as true or false. Each will be evaluated and compared against each other based on metrics like accuracy, f1 score, precision and recall. After the following machine learning models - SVM, Logistic Regression, Naïve Bayes and Random Forest will be trained and adjusted. A voting classifier will be implemented that will combine all the models mentioned above and form a composite classifier that uses all these classifiers to predict class labels and probabilities, and use soft voting to make the final prediction. A suitable dataset that has been divided into training set, validation set and test set will be used and will be preprocessed to apply feature extraction techniques in the next steps. Models like Random Forest Classifier, Naïve Bayes, Logistic Regression and SVM Classifier will be used to train the dataset. Finally, the different scores for each category will be calculated and compared. Below is a detailed flowchart that summarizes the entire process.

**2. Literature Survey**

This paper deals with analyzing and examining the various factors, which portray the bicycle-industry as a potent industry which can lead to the green recovery and sustainable development of economy and environment in Bangladesh. The authors performed a SWOT analysis after collecting and analyzing information with regards to the bicycle industry in the south-asian developing economy from many research publications, and by interviewing industry experts, goverment officials and university professors and finally a joint discussion by all the experts. Based on the findings in which both internal factors (strengths and weaknesses) as well as external factors (opportunities and threats) were taken into consideration, they came up with an internal factor evaluation matrix (IFEM) to find out that the bicycling industry in Bangladesh has enormous potential, given that some reforms be conducted upon the same with a view towards reducing carbon-emissions and decarbonizing the commute sector by facilitating green investment in bicycling as a non-motorized transport (NMT) option to reduce GHG emissions in the post-pandemic settings. Suggested potential strategies to facilitate the adoption of bicycles as a resilient and sustainable transport option such as by developing local manufacturing capability, dedicated infrastructure, reducing import duties, attracting FDIs, eliminating gender differences, etc., were also mentioned. [1]

The analysis of the pattern of bike-sharing in a region plays the most significant role in predicting the bike usage in any particular region. The authors in this paper, have provided a comprehensive review of bike-sharing usage prediction with many invaluable approaches to predict the bike-sharing usage pattern with deep learning. Following a set of procedures of the following modules: data aggregation to build the prediction input-features and targets, defining 3 data formats (time-series format, grid format and graph format), addressing 3 types of prediction problems (time series-input prediction, graph-input prediction and grid-input prediction), quantifying the prediction error using different evaluation metrics, and finally some prediction challenges (complex spatial dependencies and complex temporal dependencies) and prediction models like FFNN, LSTM, RNN, GNN, GRU, MLP, SVR, etc., were illustrated with a different section for each type of model was mentioned for both docked and dock-less bike-sharing systems. Finally, application scenarios within the bike-sharing systems and beyond were brought up followed by the challenges and the development directions in the research paper. More open datasets, various applications based on bike usage prediction and potential research directions were summarized to encourage future research. [2]

A vision of developing a sustainable bike-sharing system as viewed a PSS (product-service system) was expressed by the authors in this research paper. Although, it was initially developed considering a single focal-company, restricted to a particular region (southern region of developing country, Brazil) only, the authors emphasized the significance of their research-work claimimg that it was the first one to be done keeping in mind the scenario of a developing country. The design of the system-model organized in 4 stages: (i) value proposition, (ii) value configuration, (iii) value delivery and (iv) value capture was introduced. Although, the strategy doesn't hold the providers of the service as the manufacturers themselves, the business-model analyzed was a use-oriented in the context of shared mobility. By conducting face-to-face interviews with those involved in developing the business model, a research protocol was developed. The authors expressed that the PSS business model analyzed by them could represent significant contributions to improve micro-mobility. [3]

This paper entirely deals with understanding, analyzing and illustrating the multiple modes and forms of relationships between build environment (i.e., land-use, transportation system and urban design) and bike-share usage. Quite a many variances between the build environment and bike-usage were stated and described with some outliers in notable cases. Variance in relationship in the build environment across different mobility patterns, docked and dockless bike-share patterns, w.r.t. trip purpose, between arrival and departure patterns, based upon the day of week, etc. and the bike-share usage were elaborated. The paper concluded with a brief summary of the major findings of the authors and them encouraging the recommendations for the future research works. [4]

The study attempted to examine the associations of BnR (bike and ride) activities with metro area w.r.t. DBS (dockless bike sharing) systems, in the city of Shanghai, China. The study signalled that BnR behaviors were affected by features like station features, land use, socio-demographics, roadway designs, transportation facilities, etc. Mainly four metrics were employed in the entire study to understand BnR behaviors from the perspective of different participators viz. local govt., DBS users, etc. The metrics were BnR trip count, shared-bike utilization rate, metro catchment area and BnR rate, for the assessment of BnR performance. The generalized additive model (GAM) was utilized to build statistical inference. Several statistical issues such as over-dispersion, skewness and spatial autocorrelation were addressed while modelling DBS usage. The spatial distribution of the 4 metrics suggested that shared bikes were oversupplied in the city center while undersupplied in the suburb. Based on other things, various other conclusions were drawn for comprehensive analysis. [5]

In this research paper, the authors sought to investigate the correlation of the various factors of the perceived value upon the users willingness to pay for bike-sharing services in the first-tier and second-tier cities of China. A structural analysis was also conducted to validate the findings and visualize the significance of the different factors as variables. The paper analyzed the direct and indirect factors that affect bike-sharing users' willingness to pay. Based upon the findings, the authors concluded that, perceived usefullness and perceived ease-of-use have positive impact on perceived value; and perceived trust, perceived value, individual paying consciousness and environmental protection have positive impact on perceived value; the users' word-of-mouth and perceived entertainment have no significance; and finally perceived cost and perceived risk have negative impact on perceived value. [6]

This paper investigates the various factors which make the bike-sharing services to be retained by the users, and not just be opted by them in the first place. For data collection purposes, questionnaires were collected through both online and offline survey. A total of 650 questionnaires were collected, including 500 field surveys and 150 online questionnaires, resulting in 622 valid questionnaires. The authors introduced the related concepts of participation in purchasing decisions, customer engagement, and customer-perceived value, and uses a structural equation model to identify the interaction and influence mechanisms between the three variables and usage intent. It not only extends the scope of consumer behavior theory, but also provides management and marketing strategies for bike-sharing companies. [7]

Through the paper, the authors tried to analyse and discuss the current bicycle market scenario in India and where the developing country presently makes its stand in the world when it comes to manufacturing, exporting and ranking in terms of bikes' usage and procurement of raw materials. The objectives which they proposed in the research papers include, gaining knowledge about India's cycle industry, learning abouth the industry's development, comparative analysis between sales and production, study of the future growth and analysis of the industry, the bicycle industry's contribution in international economic development, research on how latest gadgets can be added in bikes, understanding CORONA's impact on the Indian cycle industry, and finally, a SWOT analysis to aggregate the facts and figures for recommendation purposes favoring the success of the Indian cycle manufacturing industry in the future. After collecting the data from various sources, viz., journal, published papers, archived newspaper articles, official bicycle industry websites, and other ventures, the authors discussed various aspects: the growth of bicycle industry in India; future analysis of bicycle industry in India; major competitors of India in the industry; the contribution of bicycle industry in international economy development; the research and development centre for bicycle in India; the analysis highlighting the strengths, weaknesses, opportunities and threats of the Indian bicycle industry; CSR activities; COVID-19 impact on the Indian bicycle industry. Based upon the aforementioned, they provided their recommendations and concluded with areas for future development of the industry. [8]

The paper provides a systematic literature review (SLR) of various research papers specializing in exploring and analyzing the multifarious machine learning approaches applies to bike-sharing systems (BSS). Based on a preferred reporting items for systematic reviews and meta-analyses (PRISMA) methodology, that consists of a checklist and a flow diagram, the systematic literature survey was performed. A four-phased flow diagram consisting of the following phases: identification, screening, eligibility, was aimed to describe and understand the items of the different sections. The authors framed a process workflow to understand all the stages of the study, viz., keyword identification and search, repositories, bibliometric analysis, etc. The open source tool VOSviewer, was used as the bibliometric research tool for network analysis. The tool helped to to identify the main keywords, authors, co-authors and their respective relations, within the SLR data set, for quantitative analysis. Many graphs were created for each of the sections of interest, by the authors in order to bring out reasonable insights from the study. Based on the findings, the authors asserted that the two main problems addressed by the machine learning techniques in the context of bike-share systems are clustering (classification) and prediction. Three clustering algorithms, viz., hierarchical-clustering, community-detection-clustering and k-means-clustering are more commonly used. The authors additionally also discussed the various research and study limitations in their whole study. The study tried to raise some recommendations for the future work within the overarching theme of machine learning techniques applied to BSSs. [9]

The research paper examines the various factors influencing the buying of two wheeler vehicles by the females in Palghar, distant suburb of Mumbai city. For collecting the data, the authors used structured questionnaires (primary data) and websites, journals, research articles and news reports (seconday data). For their research purposes, the authors collected data samples from a total of 150 respondents. Random sampling was used for the collection of primary data. The authors specified a set of hypothesis and performed a data analysis report, which put light upon the various aspects of the female buying behavior and more, viz., popularity of brands among female riders, use of internet in buying decision, awareness of celebrity endorsement, mode of purchasing the vehicle, etc. They concluded that necessity was the most influential factor affecting the buying decision of female two wheeler users. Based on the calculations, as the Chi square value was significantly high, it made them to accept the alternate hypothesis which they set. Other conclusion drawn were, celebrity endorsement had no significant imapact on the buying behavior and most of the women use the two-wheelers jointly with other family members. [10]

Information from survey of local respondents and other sources were used for studying the factors which influences the purchase of two wheeler modes of transport in the city of Chennai. Both primary and seconday data were involved in the study by the authors. The primary data was collected from the owners of two-wheelers using questionnaires. The secondary data was sourced from published reports, records, books, Journals, bulletins, magazines, internet and newspapers. About 100 respondents were selected for collecting the primary data with suitable sampling techniques. Some of the objectives put forward by the authors included: to analyze the age of the respondents, to identify the type of media created awareness among the consumers, to analyze the factors influencing the purchase of two-wheeler, and to know the expectation of consumers in the purchase of two-wheeler. This was followed by the analysis and implementation. The authors concluded the research paper with suggestions and conclusion. [11]

The authors scrutinized revenue management strategies for unlimited usage bike-share scenarios in their research paper. Citi Bike public data has been used by the authors for the overall analysis. Summarization of the basic data for understanding the behavioral patterns of the casual users and the number of trips that casual users take was estimated for relating between the two, because such data was not publicly available. The parameters of these distributions were derived from sample means and standard deviations using linear regression of daily short-term ticket sales and occasional passenger numbers. A path choice model was built using variables resampled from the fitted distribution by the bootstrap method. The multinomial logit model was used because it could represent not only individual vote probabilities but also aggregated market shares. As a result, the price combination of the two plans was optimized to maximize revenue based on estimated model output, and the impact on consumer surplus was quantified. [12]

This paper deals with providing a set of illustrative examples and performing a series of sensitivity analysis tests to propose an original model to find out the number and layout of the bike stations as well as the number of bicycles and free-racks for modeling bike-sharing systems based upon spatial equity concept aimimg towards reducing implementation and operational cost of the existing such systems. In the research paper proposed by the authors, a linear programming model was developed and tested to determine the number and layout of stations and the number of bikes and racks available at each station to set up and operate the BSS for each level of defined service, terms and conditions cost was minimized. The proposed linear mathematics problem minimized the setup and execution costs of a new BSS using a set of constraints that reflect the idea of ​​a balanced and fair level of service provided to all users of the system. [13]

Cities all around the world are increasingly getting concerned about the detrimental effects of increasing number of vehicles on their roads, viz., greater pollution, emission rates, congestion, etc., which has led many of then to take active steps in order to prevent such things. One of the most effective and efficient alternative and life-saver which has come out from the horizon in this regards is the usage of BSSs or bike-sharing systems. Many cities have also been seen to adopt BSSs to tackle such growing problems and more. This increased use of bicycles has prompted many cities to either expand existing systems or introduce new ones. Due to the unbalanced spatio-temporal demands of cycle tours, many cycle stations are either empty or full during the day. This could have a significant impact on the reliability and usefulness of BSS and encourage drivers to return to using their vehicles or choose alternative modes of transport, resulting in increased congestion and emissions and pollution may increase. This reduces the number of BSS users and reduces system revenue. Operators recognized the imbalance and began building more bike stations closer to each other with the aim of keeping them within a five-minute walk. However, this solution is both financially and practically difficult to implement. This extensive and enormous report and aggregation of research paper methodologies proposes a new generation of BSSs, in which some can be portable, i.e, it takes into account both the types of BSSs (docked and dockless bike-sharing systems). Specifically, the framework consists of two levels: We use fast, online, and incremental learning approaches to predict the number of bikes at stations and balance the system with wearable stations. The goal is a framework that solves the dynamic bike-sharing relocation problem to minimize unmet demand and increase or decrease user satisfaction. The dissertation by the authors contributes to the field in 5 ways. First, a multi-objective supervised clustering algorithm was developed to identify similarities in bicycle use with respect to time events. Second, a dynamic, interpretable and fast approach was developed for predicting the number of bikes at stations within a BSS. Third, a univariate inventory model was created using a Markov chain process that provides an optimal selection of bike levels at stations. The fourth was to explore the benefits of portable bike stations using an agent-based simulation approach as a proof of concept. Fifth, mathematical and heuristic approaches are proposed. [14]

The authors published this research paper which is actually based on the Be4Schools R&D project implemented in the Portugal based city of Águeda. The intent of this study conducted was to analyze the preferences of the students aged between 15-21 in the context of using e-bikes while going daily to school. It also aimed at assessing their longings and preferences towards ICT related attributes.The information for examination of the results was collected in three parts. It comprised of a mobility survey and a stated-choice (SC) experiment. The first part was aimed at collecting the responses from the students of their travel preferences about what mode they use, thoughts regarding inclusion of ICT, perceptions barrier for not cycling and previous cycling experiences. This part was named as the "Simplifying Cycling Mobility". The second part dealt with the household budget and business perpectives and in this both the students and their parents were questioned in order to get insights on what equipments to be installed, ICT preferences, household budget constraits, etc. This part was named as the "Assessing students and their parents' preferences". In the third part a SC experiment was performed to understand the trade-offs information between car travel and e-bike relevant attributes by gathering 2232 observations in that regard. A comprehensive econometric analysis was conducted to assess the nature and degree of heterogeneity in student preferences, also considering gender issues. The authors of this research paper had also taken the final shot at studying the main determinants of both the traditional as well as electric bikes impact on school-to-home/home-to-school trips, using a detailed study of the design using exploratory data analysis and multivariable logistic regression model. Statistical analysis was performed using the IBM SPSS Statistics 21 software and R. [15]

The authors of this research-paper worked on a very special and distinct idea and approach of reviewing the existing schemes and research-suggestions and findings of previous researches on bike-sharing: how well the schemes of implementation and operation suggested by the past research-approaches have actually come out victorious or to what extent such were actually guiding the working of the present bike-sharing schemes present in different parts of the world actually implementing such schemes. They provided a comprehensive review of such evidences. By following a two-fold measure: understanding and examining evidence-gaps and limitations that need further investigation; and by drawing on the evidence review and justifying whether the positive-sides of the approaches mentioned actually aid in transferability, beneficial impacts and operation-processes or not, they sought to put light on both the impacts as well as the processes (rather than just processes) as a target to enhance the current body of knowledge on bike-sharing and also contributing towards the ongoing academic and policy discourse on the increasing popular cycling measure. A sectional segmented approach was taken by the authors, as they sought to provide a critical overview of the increasing number of information sources and the growing body of knowledge about bike-sharing; discussing the evidence on users, usage and impacts of bike sharing's significance and limitations after summarizing them; providing an evidence on managing the business of bike-sharing from a process evaluation perspective; they concluded the paper by discussing how the evidence presented here can help strengthen and transfer positive results to other contexts in terms of impact and implementation processes, and identify key areas for further investigation. [16]

The long paper which comprised of multiple case study explored what various strategies the microenterprise owners in the artisan economy need to market using social media. Through the means of semistructured interviews and open-ended questions, data were collected from 5 bicycle framebuilding companies from a south-western US state. The diffusion of innovations theory was used. A thematic analysis identified seven themes from the data, viz., technical proficiency, building a social media presence, effective use of social media platforms, effective communication skills, building a brand identity, time management, and obtaining external support. The overall study findings aimed to help artisan microenterprises learn to use social media effectively, which would lead to boost in sales and profits, further leading to good positive environment for growth and development. The findings also expect their way out towards helping the local economy as it helps to prevent the money from leaving local economies, thus building strong communities. [17]

The study methodology proposed by the authors in this research paper aims to study the demands of the bicycle sharing systems, at the level of defining zones, with higher potential demand in urban areas. The literature survey conducted by the authors of this research paper presented some broad ideas, which included types of bicycle-sharing systems right from the antique to the modern systems, viz., free bike system, coin-deposit systems, information technology-based systems and multimodal systems; demand studies for cycling and bike-sharing: latent demand score method, 'revealed' or 'stated' preference surveys; etc. The methodology suggested by the authors to study the demand distributions for bike-sharing involves the two parts: quantifying the demand based upon other case studies; and defining the effect on demand caused by trip characteristics. The second part encompassed analyzing the factors such as purpose, distance, slopes, etc. and their effects on the the bicycle sharing demands thus making a cohesiveness study among the various factors proposed. After studying all these, the authors put the proposed methodology and their knowledge to exercise on the case study of Coimbra (a town located in the centre of Portugal). The conclusion came with pointing out the advantages being that the methodology proposed provides a quick-assessment technique, it can be adapted to other towns and cities, and that it is useful in the full design of the system. The paper ended with the authors pointing out the areas in which there is a scope for further studies in this area. [18]

In this research paper, the authors investigated the travel behavior characteristics of bicycle system users. A comparative analysis was done to understand the differences between annual members and short-term user profiles on Capital Bikeshare (CaBi). The data used for the overall research was gathered from Washington, DC area regional household travel survey of 2007-2008, an intercept 10 survey of short-term CaBi users, and an online survey of annual CaBi members. This paper deals with a case study of short-term and annual bikeshare users of Capital 13 Bikeshare (CaBi) in Washington, DC and Arlington County, VA. The monthly and annual users were subjected to an online membership survey, while the users who has a 24-hour or 5-day membership were examined on the basis of 23 survey questions. The objectives for the long-term members were to collect info on demographic and use characteristics of CaBi members, 4 satisfaction with the system, and changes in travel patterns based on bikeshare availability. Short-term users who did not have time to complete the verbal 19 survey were provided with instructions for completing the survey online. The analysis part comprised of the discussion on various aspects: Demographics of Area Cyclists, Short-Term Users, and Annual Members; Income, Car Ownership, and Access to a Bicycle; Trip Purpose, Mode Shift, and Helmet Use. Finally before concluding the paper, the authors also touched upon some of the scopes for future research which included whether there could be significant differences or not between Area and CaBi users when controlling for other factors, admitting that the analysis cannot fully consider the spatial restrictions of the two CaBi surveys, presence of potential biases over the sample as CaBi annual member survey respondents self-selected from email and online solicitations, and 6 responded over the internet, Results for the Washington, DC area may not apply to other US cities, etc. [19]

The approach of how the managing authorities of the bikesharing systems manage the problem of imbalance or bike rebalancing is surprisingly directly related to user-level satisfaction and thus towards profits. The authors of this research paper sought to present a thorough review of the challenges and opportunities in rebalancing of bike-sharing systems (only for 4th generation of BSS). The objective of their research points out towards collecting research papers based on the repositioning-problem in dock-based bikesharing systems, classifying them and to suggest and divert to many novel research venues. A period-wise table containing the main research-topics in BSS research over the decades has been provided in the research paper. The four key-themes addressed by the research paper are: (a) overview of the state-of-the-art related to BSS research, (b) identification of available literature regarding reviews of bike-relocation problem, (c) list and classification of studies related to algorithms used to solve bike repositioning problem, (d) evaluation of further research to be done. The methods of data acquisition included performing keyword search in Google Scholar utilizing some selective constraints, and further steps, etc. The contextualization of research in bike-sharing field was done by performing an exploratory data analysis by making a graphic representation of the main key-words and topics that appear in titles and abstracts on the bike-sharing related papers, by making use of VOSviewer. The research-topics were clustered in 5 groups, and were also further analyzed. Following the BSS Rebalancing Problem Review Analysis, Literature Review Timeline, the authors provided the Summary of Results and Discussion. Finally, an exhaustive table that will assist researchers from different disciplines to address the open challenges in the field was also provided. [20]

**3. Problem Statement**

There are many ways to detect fake news and articles using machine learning algorithms. Given today's advances in machine learning and the level of knowledge reached, it is not appropriate to blindly rely on one or two such algorithms to classify fake news. Therefore, instead of collecting more data and extensive training to improve accuracy and comparing accuracy scores for better understanding, a holistic approach for detecting fake news is the need of this crucial moment. Therefore, we decided to take a step-by-step systematic approach to classify fake news using data mining technology. The implementation of data mining operations such as data collection, data preprocessing, feature extraction, feature selection, and machine learning models can be used to make predictions to classify messages as true or false and predict the probability of the news belonging to the predicted level. A set of machine learning models will be implemented to compare the performance of machine learning models based on metrics such as accuracy, f1 value, fit rate, and recall. The primary determinant for assessing model performance can be selected as the f1 score, which takes into account the trade-off between fit and recall. After the following machine learning models (naive Bayes, SVM, logistic regression, random forest) have been trained and tuned, all of the above models are combined to form an ensemble classifier that predicts using all of these classifiers. A voting classifier will be implemented. Label and class probabilities that use the soft voting method to make the final prediction. There are two approaches for machine learning models to gain insights from headings: CountVectorizer and TfidfVectorizer. These models are trained using features extracted from both CountVectorizer and TfidfVectorizer. All models are then hyperparameter adjusted in GridSearchCV with 5 holdout cross-validation sets for all different possible parameters. This hyperparameter adjustment is intended to improve the model's f1 score. After the models have been tuned, they are tested in the test set and the model's evaluation metrics are calculated. The trained and voted models are combined into a voting classifier that uses them all as the basic estimator. When new test data is passed to the voting classifier, all underlying models are built to predict sample specifications. After receiving the labels from all models, the final label of the test sample is predicted using the soft voting mechanism.