Comparison of numerous machine learning models to solve multi-task classification problem

1st Wahidul Alam Riyad
COMP-1804: Applied Machine Learning
001188274

***Abstract*—Due to the popularity of online marketplaces in recent decades, online vendors and merchants now require their customers to provide feedback on what they have purchased. As a result, millions of evaluations are written every day, making it difficult for a potential client to decide whether or not to buy a product. For product makers, analysing many such viewpoints is difficult and time-consuming. This paper compares various machine learning algorithms to solve a multi-task classification problem from text data. The author used machine learning models to predict the ratings of the Amazon reviews and whether the product falls in the video games or musical instrument category. The author used supervised machine learning techniques such as Logistic Regression, Support Vector Machine (SVM), Naive Bayes and Random Forest to conduct this research. The accuracy of the machine learning models was compared to select the best model for the multi-task categorisation issue. Consequently, the SVM technique surpasses all other models in terms of accuracy.**

I. Introduction and related work

Due to the popularity of online marketplaces in recent decades, online vendors and merchants now require their customers to provide feedback on what they have purchased. Millions of evaluations of various products, services, and locations are generated over the internet. As a result, the internet has become the primary source of information and views about a product or service. However, as the number of product reviews accessible rises, it becomes more difficult for a potential client to decide whether or not to purchase the product. Customers are more puzzled about making the appropriate option when they hear different perspectives about the same product and read unclear evaluations. For all e-commerce enterprises, the requirement to analyse these materials appears to be critical [1].

Sentiment analysis and classification is computer research that tries to solve this challenge by extracting subjective information such as views and feelings from natural language texts. Natural language processing, text analysis, computational linguistics, and biometrics are some of the techniques that have been used to solve this challenge. Because of their simplicity and precision, machine learning algorithms have been increasingly popular in semantic and review analysis in recent years. People use Amazon every day for online shopping since it is one of the e-commerce behemoths that allows them to browse hundreds of evaluations left by other consumers about what they want. These evaluations offer vital insight into a product's features, quality, and suggestions, allowing buyers to comprehend nearly every element. This is useful not just to customers but also to vendors who manufacture their items. This research looks at the multi-task classification problem for online reviews. It uses supervised algorithms to calculate the overall scores of customer reviews by numerically categorising them and the product categories.

Sentiment analysis, often known as opinion mining, is an NLP task that involves finding and extracting personal information from text sources. The goal of sentiment categorisation is to determine a written text's general objective, adoration or criticism. The purpose of sentiment classification is to assess user evaluations and categorise them as positive or negative without requiring the machine to comprehend the semantics of each phrase or document fully. Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest are examples of machine learning techniques that can help with this. As a result, the project's problem will be as follows. On Amazon video game and music instrument product evaluations, which machine learning technique performs better in terms of accuracy?

Some difficulties exist. It is not always possible to classify words and sentences based on their previous polarity. For example, the term "wonderful" has a preceding positive polarity, but when a negation word like "not follows it," the context shifts dramatically. Different industries, such as movie reviews, travel location reviews, and product reviews, have attempted sentiment categorisation. The two most common approaches for sentiment categorisation are lexicon-based methods and machine learning methods. SVM performed effectively in all studies with lower error levels than alternative classification algorithms, according to Joachims (1998).

With SVM and Naive Bayes and maximum entropy classification, Pang (2002) attempted supervised learning for categorising movie reviews into two classes: positive and negative. They experimented with numerous features and discovered that the machine learning algorithms performed better when a bag of words was utilised as a feature in the classifiers. In terms of precision, all three approaches performed admirably. Three supervised machine learning algorithms, Naive Bayes, SVM, and N-gram model, were tested using internet evaluations about various tourism sites throughout the world in a recent study done by Ye et al. (2009). They discovered in this study that properly-trained machine learning algorithms perform very well for categorising vacation destination reviews in terms of accuracy. They also showed that the SVM and N-gram models outperformed the Naive Bayes technique. How-

ever, increasing the quantity of training data sets dramatically lowered the gap between the algorithms.

II. Ethical discussion

A formal system, which represents a natural system, will always be the foundation of a decision-making algorithm. As a result, it will always rely on a limited set of meaningful relationships, causes, and consequences. Whatever the algorithm's complexity, it will always reflect one perspective of the system being simulated. Finally, the AI algorithm's decision criteria are derived from human-made assumptions, such as where to draw the line between action and inaction and between many options. These human-made assumptions and selection rules influence how the computer responds. Only at the human or nonhuman interaction is this possible. Even the data used to train an algorithm is subject to the environment it was created.

Despite the issues expressed and the intents expressed to overcome them, the development of decision-making algorithms remains primarily unknown. There have been few shots to make the algorithms public. Attempts to make the process more inclusive, with greater stakeholder engagement, are also being made [2]. Identifying a suitable pool of social actors may necessitate a significant amount of effort in terms of stakeholder mapping to provide complete but effective governance regarding the number of participants and the ease with which working processes may be implemented. By bridging technical boundaries between developers, specialists in other disciplines, and laypeople, example-based explanations may also contribute to the successful participation of all parties.

Because they may change so fast, machine learning systems that keep learning are hazardous and difficult to comprehend. Is it possible to lock down a machine learning system with real-world implications to promote transparency? If this is the case, the algorithm may be faulty. If so, today's transparency may not assist us in grasping what the system will do tomorrow. This problem might be solved by hard coding the algorithm's rules once all stakeholders have agreed on them. As a result, the algorithm learning process would not clash with the agreed-upon norms. Making it essential to save these algorithms in a database owned and controlled by this entrusted super parts authority might help the whole process evolve more quickly.

III. Dataset preparation

Python was chosen as the programming language for this project due to its vast library and ease of usage. Python is one of the most extensively used programming languages in machine learning and data research and offers an extensive library of machine learning libraries. A simple python method was built to eliminate the unnecessary characteristics before producing the relevant data. Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest are some of the categorisation algorithms available. Other feature extraction approaches are also included. Many functionalities have been eliminated except for the review content, score, and product category. In addition, the reviewer's score is based on a scale of one to five stars. Negative reviews received one or two stars, while good reviews received four or five stars. Three-star reviews are notorious for having a lot of mixed feedback and are difficult to categorise as good or negative.

Remove any duplicate or unnecessary observations from your dataset. Duplicate words are most likely to occur during data collecting. Duplicate data might arise when you mix data sets from numerous sources, scrape data, or get from clients or several departments. One of the most critical aspects of this procedure is de-duplication. You've made irrelevant observations when you observe words unrelated to the problem you're trying to solve. For example, if you want to study data on millennial clients, but your dataset includes observations from previous generations, you might wish to eliminate such words. This may speed up analysis, focus on your primary goal, and make your dataset easier to maintain and perform [3].

There will frequently be one-off observations that don't appear to fit into the data you're examining at first sight. You may find unusual naming practices, typos, or wrong capitalisation when measuring or transferring data. These are structural mistakes. These discrepancies might result in incorrectly classified groups or classes. "N/A" and "Not Applicable" may appear in the same category, but they should be treated as such. If you have a worthy cause to delete an outlier, such as incorrect data input, doing so will improve the data you're working with. However, the presence of an outlier can occasionally be used to prove a hypothesis. Remember that the presence of an outlier does not imply that it is erroneous. Consider eliminating an outlier if it appears to be unimportant or a mistake. This step is required to determine the number's legitimacy.

Many algorithms will reject missing values, so you can't disregard them. Missing data can be dealt with in a variety of ways. Both aren't ideal, but they're worth considering. You can drop observations with missing values as a first choice, but be aware that doing so can result in the loss or loss of information, so be cautious. You may also fill in missing values based on other observations; however, you risk compromising the data's integrity because you're working with assumptions rather than facts. To successfully browse null values, you might change how the data is used as a third alternative.

IV. Methods

In a supervised learning system, the support vector machine solves classification and regression issues. Many people prefer the support vector machine because it delivers significant accuracy while using less computing power. It is mostly used to solve categorization challenges. Learning may be classified into supervised, unsupervised, and reinforcement learning. The hyperplane is divided to produce a support vector machine, a selective classifier. The method constructs the optimal hyperplane that categorizes fresh samples given labelled training data. This hyperplane is a line that divides a plane into two halves in two-dimensional space, with each class on each side. The support vector machine approach aims

to find a hyperplane in an N-dimensional space that classifies data points individually.

A classification procedure called logistic regression predicts a binary result based on independent factors. This would imply forecasting whether you will pass or fail a class in the case above. Although logistic regression may be used to address regression issues, it is most often employed to tackle classification challenges. Binomial logistic regression or binary logistic regression are other names for logistic regression. Multinomial logistic regression is used when the response variable has more than two classes. Logistic regression is one of the most used binary classification techniques in machine learning and data science, and statistics inspired it [4].

Bagged decision tree models that split on a subset of characteristics on each split are known as random forests. This is a big word, so let's break it down. We'll start with a single decision tree, then move on to bagged decision trees, and eventually to splitting on a random subset of characteristics. Because it introduces splitting on a random subset of features, random forest improves bagging by decorrelating the trees. This indicates that the model only analyses a limited subset of characteristics at each split in the tree rather than all of the model's features. This is necessary to eliminate variation. Consider what would happen if there were only a few strong predictors in the data set. These predictors will be consistently picked at the top level of the trees, resulting in similarly constructed trees. The trees would be firmly connected, in other words [5].

Naive Bayes is a supervised machine learning technique for classification. The term supervised here refers to the algorithm being trained using both input characteristics and categorization outputs. This is due to the classifier's assumption that the model's input characteristics are unrelated. As a result, altering one input characteristic has no impact on the others. As a result, it's naive in the sense that this assumption might be true or false, and it's most likely not.

V. Experiments and evaluation

The findings of the investigation are presented in this section. We have compared the precision, f-1 score, recall, and support of all the models for evaluation. At the end of each model implementation, we have calculated the accuracy of each model. The accuracy number indicates the proportion of the model's testing data set adequately categorised. We can see from the table that SVM receives 0.63 and Logistic Regression receives 0.62. On the other hand, both Naive Bayes and Random Forest gets 0.58 and 0.59, respectively. Thus, SVM is the best model for this use case scenario.

TABLE I
Comparison of ML Models Accuracy

No.	Model	Accuracy
1	Support Vector Machine	0.63
2	Logistic Regression	0.62
3	Random Forest	0.59
4	Naive Bayes	0.58

VI. Discussion and future work

Because the reviews themselves include many words, the bag of word features may be sparse. Consequently, we can observe that the algorithms' accuracies are more significant for all trials when applied to more informative summaries and have fewer words. Another obstacle in sentiment categorisation is recognising negation and its impact on sentence semantic comprehension. Future research might be helpful in further exploring this issue and providing solutions. Only four major machine learning algorithms were investigated in this research. Other efficient sentiment classifiers, such as Decision trees, should be explored in future studies.

VII. Conclusions

This study used the Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest machine learning algorithms on Amazon beauty product evaluations. When the entire data set was used as training and testing data, the SVM strategy outperformed all others in terms of accuracy.

References

[1] E. Haddi, X. Liu and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis", Procedia Computer Science, vol. 17, pp. 26-32, 2013.

[2] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward", Humanities and Social Sciences Communications, vol. 7, no. 1, 2020.

[3] Z. Zhang et al., "Realization of automatic data cleaning and feedback conditioning for J-TEXT ECEI signals based on machine learning", Fusion Engineering and Design, vol. 177, p. 113065, 2022.

[4] A. Joby, "What Is Logistic Regression? Learn When to Use It", Learn.g2.com, 2022. [Online]. Available: https://learn.g2.com/logistic-regression.

[5] J. Kho, "Why Random Forest is My Favorite Machine Learning Model", Medium, 2022. [Online]. Available: https://redirect.is/ivpyie9.