Sentiment Analysis using SVM and Naïve Bayes Classifiers on Restaurant Review Dataset

Jason Cornelius Sugitomo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jason.sugitomo@binus.ac.id

Nathaniel Kevin
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nathaniel.kevin@binus.ac.id

Nayra Jannatri
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nayra.jannatri@binus.ac.id

Derwin Suhartono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
dsuhartono@binus.edu

2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI) | 978-1-6654-4002-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCSAI53272.2021.9609776

Abstract—**Consumer reviews on the food and services of a restaurant is a significant thing to monitor for restaurant businesses. Sentiment Analysis, having another name of Opinion Mining, is a technique that was used in order to identify people's opinions and attitudes towards certain subjects, and the most widely used application of sentiment analysis is analyzing consumer reviews of their products and services. This paper will assess sentiment analysis' performance with SVM and Naïve Bayes classifiers on a dataset of restaurant reviews. A grid search with different hyperparameters of the classifiers and feature selection methods is done to compare their effects on performance. Each model will be evaluated based on accuracy, F1 score, and confusion matrix. The trained models can be further finetuned to aid restaurant businesses in tracking their business performance and reputation.**

Keywords—Sentiment Analysis, Restaurant reviews, Sentiment Classification, ML approach, Naïve Bayes, Support Vector Machines

I. Introduction

In this digital age, it has become easier for people to post reviews or read other people's reviews on restaurants. Many have developed the habit of reading a restaurant's reviews before visiting the site. According to a 2019 industry report by Toast, Inc., 35% of guests and 49% of restaurateurs choose their restaurants based on online reviews compared to other criteria of decision-making [1], shown in Fig. 1. Therefore, consumer reviews are of significant interest to restaurant businesses as their reviews can affect the restaurant's reputation and, potentially, their profits. It is important for restaurants to monitor their consumer's sentiments on their food and services.

TABLE I. Methods For Choosing Restaurants

How to choose restaurants	Restaurateurs	Guests
Online reviews	49%	35%
Restaurant social media	33%	10%
Facebook	38%	28%
Restaurant website	30%	19%
Instagram	22%	8%
Online articles	12%	21%
Consumer ordering platforms	17%	12%

Fig. 1. A part of the results of a 2019 industry report by Toast, Inc. that shows some ways in which diners choose which restaurants to dine in.

Sentiment Analysis, having another name of Opinion Mining, is a technique that was used in order to identify people's opinions and attitudes towards certain subjects. These subjects, or entities, may refer to topics or individuals. A person's opinions, or sentiments, on a subject may be positive, neutral, or negative. Sentiment analysis technology is very beneficial for organizations and businesses as it allows them to understand customer needs and monitor the reputation of their products. In businesses, the most widely used application of sentiment analysis is analyzing consumer reviews of their services and goods.

Sentiment analysis is rooted on a classification process. There are five main types of sentiment classification problems in sentiment analysis: document-level, comparative, aspect-based, sentiment lexicon acquisition, and sentence-level [2]. Document-level sentiment analysis is the most basic of sentiment analysis, which makes it the most suitable for analyzing business reviews. It expects that the document as a whole contains one opinion on the subject, which is the case for reviews. The techniques of Sentiment classification can be differentiated into three: lexicon-based, machine learning, and hybrid, which combines the previous two [3]. A study on sentiment analysis of polarizing movie reviews have shown that classification using ML techniques is the most successful, with the SVM algorithm being the most superior [4].

This paper aspires to assess sentiment analysis performance using machine learning approaches using restaurant review data. The trained model can be useful for restaurants to monitor their consumer's sentiments, or even used by other types of businesses to monitor product reputation. We believe that sentiment analysis can measure a person's true sentiments in their review more accurately than any existing rating systems (e.g., five-star rating, binary rating, etc.).

Our paper will explain the process of performing sentiment analysis with a dataset of 1000 restaurant reviews in English retrieved from Kaggle. The chosen method of sentiment analysis is to use the ML approach, more specifically using the SVM and Naïve Bayes classifiers. Furthermore, the performance of the classifiers combined with various feature

978-1-6654-4002-8/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

selection methods will be compared in order to obtain the optimal models.

Our reason for using SVM and Naïve Bayes is because of the popular usage of those algorithms in research of sentiment analysis. However, it is still quite uncertain which of them perform better. An example can be seen from one research [5] which used both Naïve Bayes and SVM to figure out people's opinion on Twitter before doing polarity analysis, but it is still undecided which one of the algorithms is the best in their conclusion. Therefore, we use both SVM and Naïve Bayes to find out people's opinions regarding the restaurant from the reviews in the Kaggle dataset.

Previous studies and results of sentiment analysis on product reviews have been successful, so we believe that sentiment analysis is a suitable tool to aid in classifying restaurant review sentiments.

II. Literature Review

A. Sentiment Analysis System Architecture

Fig. 2 shows a general sentiment analysis architecture with the following steps: data collection, data preprocessing, feature selection, sentiment classification, and evaluation.

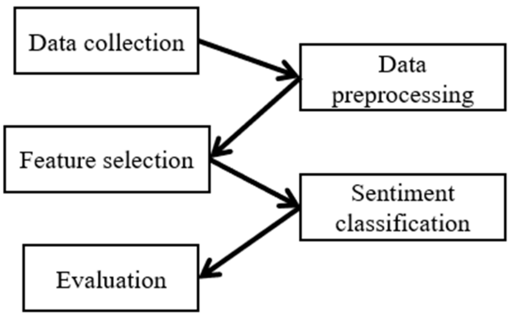

Fig. 2. The general architecture of sentiment analysis showing its workflow.

There are various methods for data collection. One of the most common methods is the use of API's (application programming interface). A study on sentiment analysis of Twitter posts uses the Twitter Streaming API to retrieve the data [6]. Another sentiment analysis study on restaurant reviews use the AYLIEN Text Analysis API [7].

The data preprocessing step significantly impacts the accuracy and performance of NLP systems [8]. The process of tokenization breaks up a stream of text to pieces known as tokens, which could contain either words, terms, or even symbols, and a typical tokenizer splits tokens based on whitespace characters or marks between words [9]. The normalization process helps to replace abbreviations, or microtext (e.g., OTW, WDYM, w8), into their actual meanings (e.g., on the way, what do you mean, wait) [10]. In POS tagging, each token is given their part-of-speech tags (e.g., subject, object, adverb, compound noun). Some other important preprocessing steps are stop words removal, stemming, or lemmatization.

In feature selection, representative features are selected from text to improve the sentiment classification step. By removing irrelevant features, sentiment classification can become more accurate and reduced running times of learning algorithms can be achieved [11]. Some commonly used feature selection methods are Document Frequency, Relief-F Algorithm, CHI Statistic, Gain Ratio, and Information Gain.

In sentiment classification, the sentiment classifier model can be developed by machine learning algorithms. Commonly used machine learning algorithms for training of sentiment classifiers are Naïve Bayes and SVMs, with the latter known to have the better performance. An example of this would be in analyzing textual reviews such as movie reviews [12]. While this is usually true, some cases prove NB to have better performance [13]. This means that both might have advantage in their accuracy at some aspect. Thus, both are usually used to find out which one has the better accuracy when doing sentiment analysis to make predictions of sentiment.

Lastly, evaluation of sentiment classification results is done. During training, results must be measured on their accuracy. This is done to ensure that the training of the model is properly done with accurate or near-accurate results.

B. Related Works

Various research on sentiment analysis over the years have used data sources such as chats, tweets, newspapers, and photos [14]. More recent sentiment analysis research focused on the online domain. Online text has evolved drastically from formal written text, so sentiment analysis methods must always adapt to the nature of online text. For example, capitalization or punctuation in online text can be a sign of seriousness or strong opinions, so methods could be adapted to include those as features in classification [15].

A study [16] found that sentiment analysis was able to be used in businesses such as restaurants. The research showed that sentiment analysis allows people who use it to understand what they need to improve according to the analysis' outcome. Sentiment analysis also proved to be very flexible, allowing people to use it in businesses of all sizes due to its flexibility. The results of another research [17] found out that people are influenced by five attributes of restaurants, namely food, service, ambience, price, and context. Further research finds that food, service, and context affect the reviews made by customers when compared to ambience and price.

A research [18] found that sentiment analysis achieved outstanding results with the SVM classifier. The research found out that sentiment analysis that was done using the SVM classifier was able to reach a high accuracy when applied to the dataset that they have used, reaching 94.56%. Another research

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

[19] supplements this as they have found that SVM produces good results in the dataset that they have used. The results of the experiment showed that the classifier reached an accuracy of 88.906% with a lambda setting of 0.0003. The data also showed that the results might help customers choose their favorite cuisine as well as giving restaurants their advantages and shortages or disadvantages. This was supplemented by another research [20] that found that SVM's performance can be tuned even more in order to make the sentiment analysis better and more accurate. The research found out that the best way to tune the classifier was to use the grid search technique as it was capable of increasing the performance of the classifier, albeit with uncertain accuracy.

Another paper shows the high results of using Naïve Bayes classifier [21]. The results of the research showed that sentiment analysis was able to be done using Naive Bayes which found out that high user evaluation relates to larger average score of constructive evaluation and low user evaluation relating to larger average score of negative evaluation. The research also found that reviews made by people can be influenced by some factors, such as the location of the person making the review. In one study on movie review datasets, both Multinomial Naïve Bayes and Bernoulli Naïve Bayes score very high in accuracy, with Multinomial Naïve Bayes having an accuracy of 88.5% [22]. Another study has shown similarly high accuracy (above 80%) for using Naïve Bayes on movie reviews, but results in lower accuracy for hotel reviews [23]. These two studies, along with another research [24] shows Naïve Bayes having the highest accuracy compared to the other classifiers in the respective studies. A research has also shown that Naïve Bayes performs with varying accuracies with different datasets, but SVM has lower variation and higher accuracy [25]. It shows that sentiment analysis accuracies vary based on context of the datasets, and the efficacy of using Naïve Bayes for sentiment analysis should continue to be investigated for other types of reviews, in this case restaurant reviews.

Different feature selection and extraction methods have been compared and evaluated. Using Chi Square in feature selection has resulted in increased speed of computation time but decrease in system performance in one study [26]. A research [27] shows that feature extraction methods are suitable for different ML classifiers. POS is best suited for SVM and Naïve Bayes while Hass tagging is best with Random Forest and linear regression. Another study proposed combining information gain and DF thresholding for feature selection which results in high testing accuracy [28].

III. Methodology

A. Data Collection

The data used is a dataset from Kaggle of 1,000 restaurant reviews in English with equal amounts of positive and negative reviews. The dataset consists of 2 columns; the first column being 'Review' which contains a string of the review, and the second column is 'Liked' which contains a Boolean value as the sentiment label (0 = negative, 1 = positive).

TABLE II. First Five Rows Of Dataset

Review	**Liked**
Wow... Loved this place.	1
Crust is not good.	0
Not tasty and the texture was just nasty.	0
Stopped by during the late May bank holiday off Rick Steve recommendation and loved it.	1
The selection on the menu was great and so were the prices.	1
Now I am getting angry and I want my damn pho.	0
Honeslty it didn't taste THAT fresh.)	0

Fig. 3. A sample of the first five reviews in the dataset.

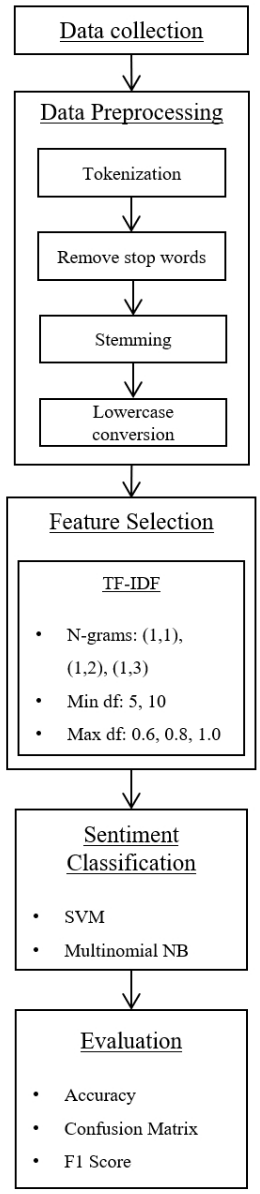

Fig. 4. The workflow showing the sentiment analysis process.

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

B. Data Preprocessing

For the data preprocessing, we will use the four steps, namely lowercase conversion, tokenization, stop words removal, and stemming.

Tokenization refers to the step in which text is turned into "tokens" before being changed into vectors. Tokens are meaningful parts of the text, such as words or phrases. The text is broken into tokens by taking only alphanumeric characters and leaving the non-alphanumeric characters.

Stop words refers to terms that occurs frequently inside the text that is not important or related to the data that we want. As such, stop words are generally expected to not be relevant for the process of text classification, so they are not included. Some stop words that will be removed are shown in Fig. 5.

Stemming is the name of the process in which the root word of a derived work is taken. This process takes the language into account since this process has algorithms that are specific to some languages.

Lowercase conversion is the process of converting uppercase letters into lowercase letters. Uppercase and lowercase letters are presumed to be identical. Because of this, all uppercase letters in the text are converted into lowercase letters before the classification process.

"a", "about", "above", "after", "again", "against", "ain", "all", "am", "an", "and", "any", "are", "aren", "aren't", "as", "at", "be", "because", "been", "before", "being", "below", "between", "both", "but", "by", "can", "couldn", "couldn't", "d", "did", "didn", "didn't", "do", "does", "doesn", "doesn't", "doing", "don", "don't", "down", "during", "each", "few", "for", "from", "further", "had", "hadn", "hadn't", "has", "hasn", "hasn't", "have", "haven", "haven't", "having", "he", "her", "here", "hers", "herself", "him", "himself", "his", "how", "i", "if", "in", "into", "is", "isn", "isn't", "it", "it's", "its", "itself", "just", "ll", "m", "ma", "me"

Fig. 5. Example of stop words that are removed in data preprocessing.

C. Feature Selection

1) TF-IDF

For feature selection, we use the vectorizer TF-IDF (Term Frequency - Inverse Document Frequency). TF-IDF is an unsupervised feature extraction algorithm that operates on the level related to the words or vocabulary of a language [29]. The algorithm is rooted on the amount a term occurs inside of a document as shown on the TF part of the name that stands for Term Frequency.

It can be computed with this formula :

$$Term\ Frequency\ (W) = \frac{Frequency\ of\ word\ (W)\ appearing\ in\ a\ document\ (D)}{Total\ number\ of\ words\ (W)\ in\ document\ (D)}$$

IDF or Inverse Document Frequency prioritizes the rarely occurring words in the text and is calculated using the following formula:

$$Inverse\ Document\ Frequency\ (W) = \log_e \frac{Total\ number\ of\ documents\ (D)}{Total\ number\ of\ documents\ containing\ word\ (W)}$$

The TF-IDF score is counted using the following formula:

$$TF - IDF(W) = TF(W) \times IDF(W)$$

2) Hyperparameter Grid Search

In order to seek the optimal combination of hyperparameters for our classifier, we will perform a grid search. The parameters that will be tuned is the gram range, minimum document frequency, maximum document frequency, and machine learning model.

TABLE III. GRID SEARCH PARAMETERS

Parameter	
Gram range	(1,1), (1,2), (1,3)
Min DF	5, 10
Max DF	0.6, 0.8, 1.0
ML model	SVM, MNB

Fig. 6. The parameters for our grid search.

The N-gram refers to the number of consecutive words or tokens that can be considered a single feature. Unigram only selects single tokens as features, while bigram and trigram can select up to pairs of adjacent tokens and three adjacent tokens respectively. In one study [30], it is observed that classification accuracy decreases as the gram range increases. However, another study [31] shows that bigram and trigram can perform better under different settings. The use of unigram, bigram, and trigram in feature selection for restaurant reviews will be tested to compare their effects towards the function of the classifiers.

D. Sentiment Classification

From a research [16], it was found out that there are advantages and disadvantages in the usage of both SVM and Naive Bayes in sentiment classification. In the research, it is

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

stated that SVM is better to be used with large datasets while Naive Bayes is more suited for small datasets. This statement contradicts another research [32] which uses 2090 twitter messages in its dataset and found out that Naive Bayes performs better than SVM, which contradicts the research before. Because of that, we will use both algorithms in our experiments in order to see which one performs better for our dataset. The Naïve Bayes algorithm we will use is Multinomial Naïve Bayes, which is most suitable for sentiment analysis as it can be used for term frequencies.

E. Evaluation methods

Because of the compact size of the available dataset, the resulting models are evaluated by cross validation. For each model, 700 reviews are selected nonspecifically for training and the leftover 300 reviews is used for testing. For each training and testing dataset, there is a balance between constructive and destructive evaluations. Each model will be evaluated based on accuracy, F1 score, and confusion matrix.

IV. RESULTS AND DISCUSSION

During the processing of the data, word clouds of the dataset are generated as an initial observation to see the most common words in restaurant reviews.

Fig. 7. Word Cloud for combined reviews.

In Fig. 7, the word cloud is the combination of the words used in the negative and positive sentiment reviews. This word cloud is used to find the most common words that are present in both positive and negative sentiment reviews in our dataset. As we can see from the word cloud, the most common words for restaurant reviews are time, great, place, good, food, service, and back. There are also words like amazing, experience, friendly, and many more. Based on this result, we can see that there seemed to be a lot of positive words even for the negative sentiment reviews. This is perhaps because the negative sentiment reviews are not written to be as direct as the positive ones.

Fig. 8. Word Cloud for positive reviews.

In Fig. 8, the word cloud is for the positive words that appear in the positive sentiment reviews. As we can see, the words that are used the most by reviewers are good, great, and food place. There are also other positive words such as amazing, friendly, best, and delicious. The positive reviews use words with strong positive sentiments.

Fig. 9. Word Cloud for negative reviews.

In Fig. 9, the word cloud is for the negative words that appear in the negative sentiment reviews. As we can see, the words that are used the most by reviewers in the word cloud are not strong negative sentiment words, but they are mostly neutral sentiment words such as service, food, and place. This is caused by the fact that the negative reviews in our dataset are not as straightforward as the positive ones, which directly praise the food and the restaurant experience. While this is the case, there are still negative sentiment words that can be identified in the word cloud such as bad, never, worst, and bland. However, some positive sentiment words are also found, which can be explained by 'vague' negative reviews. An example of a negative review which contains positive sentiment words is *"seems like a good quick place to grab a bite of some familiar pub food, but do yourself a favor and look elsewhere."*. This

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

contains positive sentiment words such as "good" and "quick", but the overall sentiment of the review is negative.

After visualization with word clouds, the data is preprocessed and fitted into a total of 36 models based on the grid search. Each model has a unique combination of hyperparameters. Fig. 10 shows a table of the results of all 36 models sorted by accuracy, then F1 score. The best model uses a Naïve Bayes classifier which achieved a precision of 77.33% and F1 score of 0.7792, with settings N-Gram = (1,3), Min DF = 5, and Max DF = 1.0. Another observation from the table of results is that Min DF = 5 is clustered at the top while Min DF = 10 is clustered at the bottom of the results.

To have a better understanding of the performance of each parameter, density distributions are plotted against accuracy for better visualization.

TABLE IV. MODEL RESULTS

N-Gram	Min DF	Max DF	Model	Accuracy	F1	TN	FN	TP	FP
(1, 3)	5	1.0	mnb	0.7733	0.7792	112	30	120	38
(1, 3)	5	0.6	mnb	0.7733	0.7733	116	34	116	34
(1, 2)	5	1.0	svm	0.7733	0.7671	120	38	112	30
(1, 3)	5	0.6	svm	0.7700	0.7596	122	41	109	28
(1, 2)	5	0.8	svm	0.7667	0.7619	118	38	112	32
(1, 3)	5	0.8	svm	0.7667	0.7482	126	46	104	24
(1, 2)	5	0.6	mnb	0.7633	0.7657	113	34	116	37
(1, 3)	5	1.0	svm	0.7633	0.7509	122	43	107	28
(1, 1)	5	0.6	mnb	0.7567	0.7653	108	31	119	42
(1, 2)	10	0.6	svm	0.7567	0.7245	131	54	96	19
(1, 3)	5	0.8	mnb	0.7500	0.7387	119	44	106	31
(1, 2)	5	0.6	svm	0.7500	0.7350	121	46	104	29
(1, 2)	5	1.0	mnb	0.7433	0.7556	104	31	119	46
(1, 1)	5	0.8	mnb	0.7433	0.7524	106	33	117	44
(1, 1)	10	1.0	svm	0.7433	0.7159	126	53	97	24
(1, 3)	10	0.8	svm	0.7433	0.7138	127	54	96	23
(1, 1)	5	1.0	mnb	0.7400	0.7383	112	40	110	38
(1, 2)	5	0.8	mnb	0.7400	0.7365	113	41	109	37
(1, 1)	5	0.6	svm	0.7400	0.7310	116	44	106	34
(1, 1)	10	0.8	svm	0.7400	0.7068	128	56	94	22
(1, 3)	10	1.0	mnb	0.7367	0.7393	109	38	112	41
(1, 3)	10	0.6	mnb	0.7333	0.7122	121	51	99	29
(1, 2)	10	1.0	svm	0.7300	0.7055	122	53	97	28
(1, 2)	10	0.8	svm	0.7233	0.7067	117	50	100	33
(1, 2)	10	1.0	mnb	0.7200	0.6934	121	55	95	29
(1, 1)	5	1.0	svm	0.7133	0.7152	106	42	108	44
(1, 1)	10	0.6	svm	0.7133	0.6861	120	56	94	30
(1, 3)	10	0.6	svm	0.7100	0.6692	125	62	88	25
(1, 3)	10	0.8	mnb	0.7067	0.6944	112	50	100	38
(1, 2)	10	0.6	mnb	0.7067	0.6923	113	51	99	37
(1, 1)	10	0.6	mnb	0.7067	0.6901	114	52	98	36
(1, 2)	10	0.8	mnb	0.7067	0.6879	115	53	97	35
(1, 1)	5	0.8	svm	0.7033	0.6983	108	47	103	42
(1, 3)	10	1.0	svm	0.6933	0.6849	108	50	100	42
(1, 1)	10	1.0	mnb	0.6900	0.6782	109	52	98	41
(1, 1)	10	0.8	mnb	0.6767	0.6255	122	69	81	28

Fig. 10. Table of results for 36 models with unique combination of hyperparameters.

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

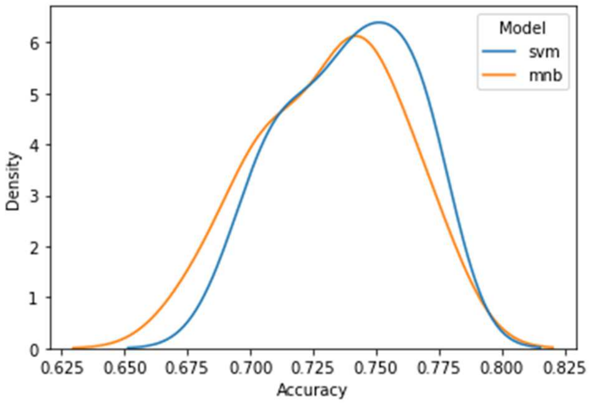

Fig. 11. Accuracy density distribution for SVM and MNB models.

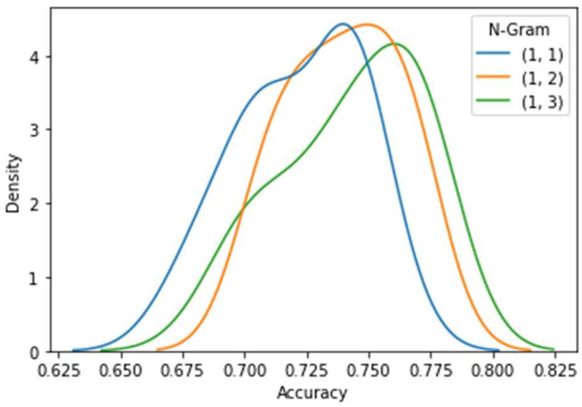

Fig. 12. Accuracy density distribution for different gram ranges.

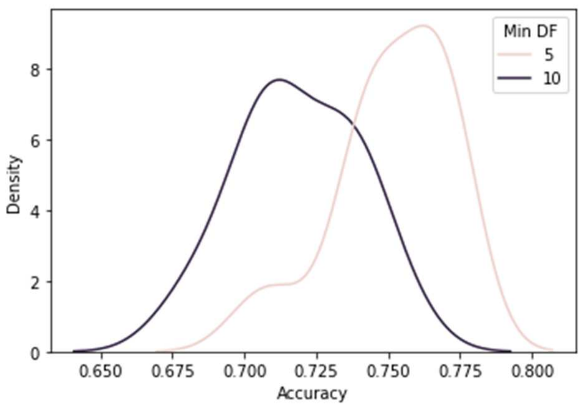

Fig. 13. Accuracy density distribution for different Min DF.

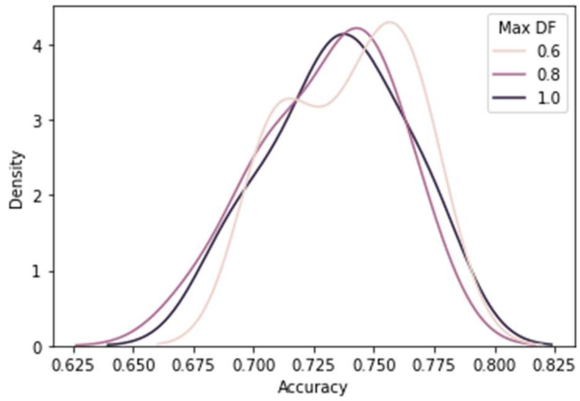

Fig. 14. Accuracy density distribution for different Max DF.

Fig. 11 compares the performance accuracy of the two classifier models used, SVM and Naïve Bayes. The line for SVM is located further right compared to the line for Naïve Bayes, meaning that the models with SVM classifier can achieve an overall greater accuracy. The accuracy of the different gram ranges is compared in Fig. 12, in which bigram and trigram seems to perform better than unigram. This may be because features selected with bigram and trigram can be much more specific and less 'vague', as it was an issue in the single-words word cloud. Fig. 13 shows that minimum DF of 5 has the better performance. This may be because using higher minimum DF may result in omitting features that are actually important for classification, which decreases accuracy. Finally, Fig. 14 compares the accuracy of using different values for maximum DF. The plotted density distribution does not really show a clear pattern, and it may be influenced by other variables. Overall, these are the settings that result in highest performances for sentiment analysis of restaurant reviews.

Fig. 15 shows all the features selected by the best model (the model on the top row of Fig. 10). Most of the features that are selected are strong sentiment terms, such as "recommend thi place" and "veri disappoint". Some of the strong sentiment terms are also specific to the topic of food and restaurants, such as "delici" (delicious) and "tasteless". This shows that the fitted models are effective specifically for restaurant reviews, because they can successfully select the sentiment terms that are strong for restaurant reviews.

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

['-', '1', '10', '2', '3', '30', '5', 'absolut', 'also', 'alway', 'amaz', 'ambianc', 'ani', 'anoth', 'anytim', 'anytim soon', 'area', 'around', 'arriv', 'ask', 'atmospher', 'attent', 'authent', 'avoid', 'away', 'awesom', 'back', 'bacon', 'bad', 'bar', 'bare', 'bathroom', 'beauti', 'becaus', 'beef', 'beer', 'befor', 'best', 'better', 'bit', 'bland', 'bread', 'breakfast', 'bring', 'buffet', 'burger', 'busi', 'came', 'check', 'chef', 'chicken', 'chip', 'clean', 'close', 'cold', 'come', 'come back', 'consid', 'cook', 'could', 'custom', 'custom servic', 'day', 'deal', 'decor', 'definit', 'delici', 'delight', 'dessert', 'dine', 'dinner', 'disappoint', 'dish', 'done', 'dri', 'drink', 'drive', 'dure', 'eat', 'eaten', 'egg', 'either', 'enjoy', 'enough', 'even', 'ever', 'everi', 'everyth', 'everyth wa', 'excel', 'expect', 'experi', 'extrem', 'famili', 'fantast', 'far', 'fast', 'feel', 'feel like', 'felt', 'first', 'first time', 'fish', 'flavor', 'food', 'food servic', 'food wa', 'found', 'fresh', 'fri', 'friend', 'friendli', 'full', 'get', 'give', 'go', 'go back', 'good', 'good food', 'got', 'great', 'great food', 'great place', 'great servic', 'ha', 'hand', 'happi', 'hard', 'heart', 'help', 'hi', 'hit', 'home', 'hope', 'horribl', 'hot', 'hour', 'huge', 'ice', 'impress', 'incred', 'insid', 'kept', 'know', 'lack', 'larg', 'last', 'leav', 'left', 'like', 'like thi', 'littl', 'live', 'locat', 'look', 'lot', 'love', 'love thi', 'lunch', 'made', 'make', 'manag', 'mani', 'may', 'meal', 'meat', 'mediocr', 'menu', 'minut', 'money', 'mouth', 'much', 'must', 'need', 'never', 'new', 'next', 'nice', 'night', 'noth', 'old', 'onc', 'one', 'onli', 'order', 'outsid', 'overal', 'overpr', 'owner', 'pay', 'peopl', 'perfect', 'pho', 'pizza', 'place', 'pleas', 'poor', 'portion', 'potato', 'pretti', 'price', 'probabl', 'qualiti', 'quick', 'quit', 'rare', 'real', 'realli', 'realli good', 'reason', 'recommend', 'recommend thi', 'recommend thi place', 'restaur', 'return', 'review', 'right', 'roll', 'rude', 'said', 'salad', 'sandwich', 'sat', 'sauc', 'say', 'seafood', 'seat', 'see', 'select', 'serious', 'serv', 'server', 'server wa', 'servic', 'servic wa', 'set', 'shrimp', 'sick', 'side', 'sinc', 'slow', 'small', 'soon', 'special', 'spici', 'spot', 'staff', 'star', 'stay', 'steak', 'still', 'suck', 'super', 'sure', 'sushi', 'sweet', 'tabl', 'taco', 'take', 'talk', 'tast', 'tasteless', 'tasti', 'tell', 'tender', 'terribl', 'thai', 'thi', 'thi one', 'thi place', 'thi wa', 'thing', 'think', 'thought', 'time', 'took', 'total', 'town', 'treat', 'tri', 'trip', 'twice', 'two', 'us', 'use', 'vega', 'veri', 'veri disappoint', 'veri good', 'visit', 'wa', 'wa amaz', 'wa delici', 'wa first', 'wa good', 'wa great', 'wa pretti', 'wa terribl', 'wa veri', 'wait', 'waiter', 'waitress', 'waitress wa', 'want', 'warm', 'wast', 'watch', 'way', 'well', 'went', 'wine', 'wonder', 'worst', 'worth', 'would', 'wrong', 'year']

Fig. 15. Features selected by the best model.

V. CONCLUSION

This study was conducted to compare the capabilities of two different machine learning classifiers for sentiment analysis of restaurant reviews. The two classifiers compared are SVM and Naïve Bayes classifiers. Using grid search to assess different hyperparameter combinations for each classifier, a total of 36 models were fitted. Evaluation of these models shows that Naïve Bayes resulted in the single best model with the highest precision of 77.33% and F1 score of 0.7792, but for the overall performance, SVM slightly outperformed Naïve Bayes, also reaching accuracies of up to 77%. Results also show that bigram and trigram result in better accuracy compared to unigram. Omitting less features can also result in better performance. Further research to compare these classifiers can be done by adding more parameters in the grid search, such as number of features selected and length of features. Overall, the results from this study show promising accuracy in sentiment analysis of restaurant reviews, and the trained models can be further finetuned to aid restaurant businesses in tracking their business performance and reputation.

VI. REFERENCES

[1] Toast Inc., "Restaurant Success in 2019 Industry Report," 2019.

[2] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013, doi: 10.1145/2436256.2436274.

[3] D. Maynard and A. Funk, "Automatic Detection of Political Opinions in Tweets," in *8th international conference on the semantic web*, 2012, pp. 88–99, doi: 10.1007/978-3-642-25953-1_8.

[4] M. Annett and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," in *Canadian AI 2008: Advances in Artificial Intelligence*, 2008, pp. 25–35.

[5] A. L. Firmino Alves, C. de S. Baptista, A. A. Firmino, M. G. de Oliveira, and A. C. de Paiva, "A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets," in *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web - WebMedia '14*, 2014, pp. 123–130, doi: 10.1145/2664551.2664561.

[6] M. S. Omar, A. Njeru, S. Paracha, M. Wannous, and S. Yi, "Mining tweets for education reforms," in *2017 International Conference on Applied System Innovation (ICASI)*, May 2017, pp. 416–419, doi: 10.1109/ICASI.2017.7988441.

[7] M. R. D. Ching and R. de Dios Bulos, "Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis," in *Proceedings of the 2019 3rd International Conference on E-commerce, E-Business and E-Government - ICEEG 2019*, 2019, pp. 62–67, doi: 10.1145/3340017.3340018.

[8] J. Camacho-Collados and M. T. Pilehvar, "On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 40–46, doi: 10.18653/v1/W18-5406.

[9] V. S and J. R, "Text Mining: open Source Tokenization Tools – An Analysis," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, Jan. 2016, doi: 10.5121/acii.2016.3104.

[10] R. Satapathy, C. Guerreiro, I. Chaturvedi, and E. Cambria, "Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 407–413, doi: 10.1109/ICDMW.2017.59.

[11] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," 2012, doi: https://doi.org/10.1145/2401603.2401605.

[12] N. Banik and M. Hasan Hafizur Rahman, "Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sep. 2018, pp. 1–6, doi: 10.1109/ICBSLP.2018.8554497.

[13] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2018, pp. 1–6, doi: 10.1109/CITSM.2018.8674352.

[14] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Comput. Sci. Rev.*, vol. 27, 2018, doi: 10.1016/j.cosrev.2017.10.002.

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

[15] M. Taboada, “Sentiment Analysis: An Overview from Linguistics,” *Annu. Rev. Linguist.*, vol. 2, no. 1, 2016, doi: 10.1146/annurev-linguistics-011415-040518.

[16] K. Kaviya, C. Roshini, V. Vaidhehi, and J. D. Sweetlin, “Sentiment analysis for restaurant rating,” in *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, Aug. 2017, pp. 140–145, doi: 10.1109/ICSTM.2017.8089140.

[17] Q. Gan, B. H. Ferns, Y. Yu, and L. Jin, “A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews,” *J. Qual. Assur. Hosp. Tour.*, vol. 18, no. 4, pp. 465–492, Oct. 2017, doi: 10.1080/1528008X.2016.1250243.

[18] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, “Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques,” 2019, pp. 687–696.

[19] B. Yu, J. Zhou, Y. Zhang, and Y. Cao, “Identifying restaurant features via sentiment analysis on yelp reviews,” 2017.

[20] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, “SVM optimization for sentiment analysis,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 4, 2018, doi: https://doi.org/10.14569/IJACSA.2018.090455.

[21] A. Micu, A.-E. Micu, M. Geru, and L. Radu, “Analyzing user sentiment in social media: Implications for online marketing strategy,” *Psychol. Mark.*, vol. 34, no. 12, 2017, doi: https://doi.org/10.1002/mar.21049.

[22] A. Rahman and M. S. Hossen, “Sentiment Analysis on Movie Review Data Using Machine Learning Approach,” Sep. 2019, doi: 10.1109/ICBSLP47725.2019.201470.

[23] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, 2016, doi: 10.5815/ijieeb.2016.04.07.

[24] K. L. S. Kumar, J. Desai, and J. Majumdar, “Opinion mining and sentiment analysis on online customer review,” 2016, doi: 10.1109/ICCIC.2016.7919584.

[25] T. K. Shivaprasad and J. Shetty, “Sentiment analysis of product reviews: A review,” Mar. 2017, doi: 10.1109/ICICCT.2017.7975207.

[26] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, “Aspect-based sentiment analysis to review products using Naïve Bayes,” 2017, p., doi: 10.1063/1.4994463.

[27] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, “Sentiment analysis: a review and comparative analysis over social media,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 1, 2020, doi: 10.1007/s12652-018-0862-8.

[28] A. I. Pratiwi and Adiwijaya, “On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis,” *Appl. Comput. Intell. Soft Comput.*, vol. 2018, 2018, doi: 10.1155/2018/1407817.

[29] A. Madasu and S. Elango, “Efficient feature selection techniques for sentiment analysis,” *Multimed. Tools Appl.*, vol. 79, pp. 6313–6335, 2020, doi: https://doi.org/10.1007/s11042-019-08409-z.

[30] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of sentiment reviews using n-gram machine learning approach,” *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016, doi: https://doi.org/10.1016/j.eswa.2016.03.028.

[31] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: opinion extraction and semantic classification of product reviews,” in *WWW '03: Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519–528, doi: https://doi.org/10.1145/775152.775226.

[32] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, “Machine Learning-Based Sentiment Analysis for Twitter Accounts,” *Math. Comput. Appl.*, vol. 23, no. 1, p. 11, Feb. 2018, doi: 10.3390/mca23010011.

Authorized licensed use limited to: Universiteit van Tilburg. Downloaded on December 30,2021 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.