Classifying Fake Reviews using Ensemble Learning

Angelo L. Guan and Concepcion L. Khan

***Abstract*— An ensemble model composed of Decision Trees, AdaBoost and Gradient Boosting was created in order to tackle the issues of using an imbalanced dataset, and using non-English datasets. The ensemble model was trained on both English and Filipino Yelp datasets and was compared to other models. The results showed that the ensemble model achieved an f-1 score of 0.6822 and an AUC score of 0.7600 using the English Yelp dataset. In addition, it also achieved an f-1 score of 0.6859 and an AUC score of 0.7441. The findings suggest that the ensemble model can reliably classify fake reviews.**

***Index Terms*— Fake reviews, Ensemble learning, Ensemble model, Majority voting**

I. Introduction

Detecting potential fake reviews is a difficult problem to solve because a reviewer can create a fake review that is convincing enough and doesn't really differentiate from other reviews. Mohawesh et al. [1] cited challenges in this research topic, such as the vast number of reviews that need to be reviewed because it can pose a challenge for a lot of consumers.

A review is a description of the experiences and opinions of a consumer regarding a product or a service. In [2], the author stated that 91% of the people surveyed read through the reviews. Also, the author in [3] found out that 92% of consumers are hesitant to purchase a product if it is unreviewed and that it can increase sales conversions by 270%. Thus, a consumer's purchasing behavior can be influenced by the product reviews that they read.

Due to the influence of reviews on consumers, some sellers have begun using fake reviews to deceive users into buying their products. A fake review is a type of review that does not reflect the buyer's experience with the product, and it can be categorized as a good, neutral, or bad review as stated in [4]. Pitman [5] states that 74% of users have encountered fake reviews while browsing online, and 54% of users said that they will not buy a product if they suspect that a product has fake reviews. Given these statistics, it implies that fake reviews can also influence a buyer's purchasing behavior.

A. Previous work done by other researchers

Researchers have been focusing on detecting fake reviews in recent years due to the threat it poses to consumers. There have been approaches to tackle the problem, such as rule-based detection, graph models, machine learning, and deep learning, as described in [6]. More specifically, Lim et al. [7] created a rule-based detection model to detect fake reviews by deriving behavior scoring models to rank reviewers based on their spamming behavior. Researchers used machine learning to detect fake reviews by using well known supervised learning algorithms such as decision trees, logistic regression, Naive Bayes, Random Forests, and Support Vector Machine. In the Deep Learning approach, it solves the shortcomings of machine learning because it can be used for feature extraction that can reduce the time cost [6].

A study [8] outlines several challenges in classifying fake reviews, such as the lack of utilization of feature selection techniques and imbalanced datasets. In particular, they stated that the lack of feature selection techniques can potentially lead to high dimensional feature sets that can lead to poor performance. Furthermore, since true reviews are more frequent than fake reviews, this leads to imbalanced datasets that can lead to poor performance using traditional machine learning algorithms because the majority class may be favored during training.

Another challenge observed by [8] found that obtaining a labeled dataset for training supervised algorithms was found to be difficult and inaccurate due to manual classification. In addition to that, training using synthetic datasets has been found to give unreliable performance when it comes to real-world data.

As described in [1], another challenge regarding fake review detection is that few studies have used fake review datasets from other languages other than English, Chinese, Malay, or Arabic. In addition to that, fake reviewers can use a translation tool to quickly convert the fake review to another language.

There have been commercial products that have dealt with detecting fake reviews on Amazon, such as Fakespot and ReviewMeta. Fakespot [9] processes and organizes information such as reviews, product information, seller information, and other data using dozens of machine learning models joined by algorithms. On the other hand, ReviewMeta uses statistical modeling to check if a review is fake or not.

B. Proposed approach in classifying fake reviews

The researcher used a type of ensemble learning technique called stacking that combines three base models and uses a majority voting mechanism to classify fake reviews. The stacking ensemble approach combines three models such as Decision Tree (DT), AdaBoost, and Gradient Boosting. The choice of these three algorithms is based on their potential to classify imbalanced datasets as found in [10]. The ensemble model was assessed against other commonly used models found in the literature review.

C. Statement of the Problem

The purpose of this study was to address the issues outlined in [8]. In particular, the ensemble model addressed the issue

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

of lowered performance due to using an imbalanced dataset and due to its robustness against the effects of an imbalanced dataset compared to traditional machine learning classifiers. Furthermore, a feature selection method was used to address the issue of high-dimensional data, and it can also be used to determine the important features of the datasets. Lastly, the datasets used in this study are the English Yelp dataset and a Filipino Yelp dataset created by translating the English version into Filipino.

D. Significance of the Study

The findings of this study provided new insights into the use of a stacking ensemble learning model in classifying fake reviews on an imbalanced dataset. Furthermore, this study can help future researchers because it can be used to show the application of other ensemble learning models such as stacking to address the issue of performance in an imbalanced dataset in classifying fake reviews. Furthermore, it can help other researchers by using other languages in evaluating fake reviews, such as Filipino.

E. Objectives of the Study

The general objective of this study was to tackle the issue of using an imbalanced dataset in classifying fake online reviews using the ensemble learning model and to address the issue of using other languages for the dataset. More specifically, the objectives that the researcher set out to accomplish are:

- To implement an ensemble model using Decision Tree, AdaBoost, and Gradient Boosting as the three base models and using a majority voting mechanism to classify fake online reviews.
- To use a Filipino dataset as one of the datasets to evaluate the ensemble model.
- To determine which features are considered most important in classifying fake reviews.
- To compare the performance of the ensemble model against other well-known models in classifying fake reviews in terms of f1-score and AUC score.

F. Scope and Limitations

The study focused on implementing an ensemble model to detect fake reviews. The ensemble model was limited to three base models such as Decision Trees, AdaBoost, and Gradient Boosting, and used a voting mechanism to classify fake reviews. Furthermore, the ensemble learning technique was limited only to the stacking technique, and only English and Filipino were used as the languages of the datasets. Also, the Filipino dataset was derived from the English dataset, and there were syntactic and semantic issues arising from translation. In addition, the study was conducted from February 2022 until June 2022 and was conducted at the University of the Philippines Los Baños.

II. Review of Related Literature

A. Fake review classification in the last decade

Rodrigues et al. [11] summarized multiple papers regarding fake review detection and the different approaches used by different researchers. The study found that the most common supervised machine learning algorithms are Naive Bayes and SVM classifiers. Other than machine learning algorithms, Mohawesh et al. [1] categorized two features commonly used by researchers: textual and behavioral features. Furthermore, they defined textual features as features that use grammar, semantics, lexicon, and metadata features to identify fake reviews. In contrast, behavioral features refer to the significance between a user's review and a user's past behavior. Another survey [8] found out that the most common approach used by researchers was using supervised machine learning to classify fake reviews. Due to the difficulty of obtaining a labeled dataset, there has been an interest in using unsupervised and semi-supervised learning to classify fake reviews. However, these approaches suffer from poor performance compared to the supervised learning approach. Furthermore, they stated that research using those two approaches is limited and rare.

B. Detecting fake reviews using linguistic features

A study [12] examined which linguistic features are the most helpful when detecting fake reviews from real reviews. They used seven classifiers such as Decision Tree (DT), Random Forest (RF), Support Vector Machine, Extreme Gradient-Boosting Trees (XGBT), Multilayer Perceptron (MLP), Logistic Regression (LR), and Naive Bayes (NB). In addition to that, the authors of [13] considered several review-centric features, such as textual features and meta-data features, to help them detect fake reviews. In terms of classifiers, they used supervised machine learning algorithms such as the Random Forests (RF) classifier. The study shows that the accuracy of the RF with review-centric features has an accuracy of 80.03%.

As shown in the studies above, using linguistic features in classifying fake reviews is an effective way of classifying fake reviews. Furthermore, there are a lot of linguistic features that were introduced by the studies, such as transcribed speech features.

C. Detecting fake reviews using supervised machine learning

The study in [13] used textual features to classify fake reviews. Furthermore, the study used an imbalanced dataset. In order to properly evaluate the models, they used the F-1 score. In terms of accuracy, SVM is the most accurate classifier, and the classifier that achieved the highest f1-score is the K-Nearest Neighbor. The authors of [14] used features such as readability, topic, n-gram features, and behavioral features. The most accurate classifier is the Logistic Regression which has a 97.2% accuracy. The researchers in [15] combined both textual and behavioral features as the input for the machine learning models. They used Gaussian EM Clustering to find hidden structures within the data. Based on the results, Random Forest is the most accurate among the three algorithms used.

As stated in [8], the supervised learning approach is the most common approach used by researchers to classify fake reviews. The studies above used this approach by training and testing models using SVM, Naive Bayes, Random Forests, and Logistic Regression, among others. Furthermore, the performance of the models is extremely accurate.

D. Detecting fake reviews using ensemble learning and ensemble models

Fayaz, et al. [16] created an ensemble model composed of Random Forests (RF), K-nearest neighbors (KNN) and Multilayer Perceptron (MLP) using the public Yelp dataset. The study also used feature selection methods such as Chi-square, Univariate, and Information Gain. The proposed model was compared against other ensemble learning models. In conclusion, the proposed ensemble model performed better than all of the other ensemble models that they compared it against. Another study [17] used decision trees, random forests, SVMs, Extreme Gradient-Boosting Trees, and Multilayer Perceptrons (MLP). It was shown that the ensemble models that they created were 77.3% accurate, and the Adaboost ensemble was the most accurate.

There are studies that use ensemble learning as their model to classify fake reviews. In particular, they used ensemble learning techniques such as bagging and boosting methods to classify fake reviews. Furthermore, the performance of the ensemble models is comparable to that of the studies above that used a supervised learning approach.

E. Acquisition of the dataset

Ma and Li [18] stated that one of the challenges in the problem of fake review detection is the acquisition of the dataset. More specifically, they said that it is easy to collect large amounts of unlabeled reviews but that collecting large amounts of labeled data is costly and time-consuming. A study [19] created a dataset that contains deceptive opinions that have been described as ”gold-standard”. This dataset is publicly available and has been used by other studies as their dataset. To acquire fake reviews, they used crowdsourcing services such as Amazon Mechanical Turks (AMT), where they asked them to create fake reviews for 20 chosen hotels.

These studies show the different approaches by previous researchers to classifying fake reviews by using different machine learning models and features. However, there are limited studies that have used a stacking ensemble model to address the issue of poor performance when using an imbalanced dataset in classifying fake reviews.

III. Materials and Methods

A. Language and libraries to be used

The Python programming language was used to program the model. The libraries used in this study are pandas, numpy, scikit-learn, nltk, textblob, spacy, spellchecker, and matplotlib. More specifically, Pandas are used to handle the dataset in an efficient manner alongside Numpy. Scikit-learn is a machine learning library designed for Python that includes numerous regression models, classifiers, and clustering models. NLTK, TextBlob, and Spellchecker will be used to derive linguistic features from the dataset. Matplotlib will be used to visualize the data. The study utilized the Cloud Translation API from Google in order to translate the English reviews into Filipino. Lastly, in order to acquire parts of speech tags in Filipino, the study used the tool provided in [20].

B. Data gathering

The dataset to be used in this study came from the publicly available dataset found in [21]. This dataset comes from Yelp. More specifically, this dataset contains a total of 39,086 restaurant reviews, with 34,818 true reviews and 4,268 fake reviews. This means that 89.1% of the dataset are considered true reviews while 10.9% are considered fake reviews. The dataset contains ratings, restaurant name, review text, date of review, user name, rating, and classification by Yelp.The English Yelp dataset can be considered as a good dataset considering that it was used in the studies of [1], [13], [14], [15] and [16]. In order to address the issue of low utilization of multilingual datasets, the study used a Filipino Yelp dataset. This dataset was derived from the English Yelp dataset, but it was translated using Cloud Translation from Google.

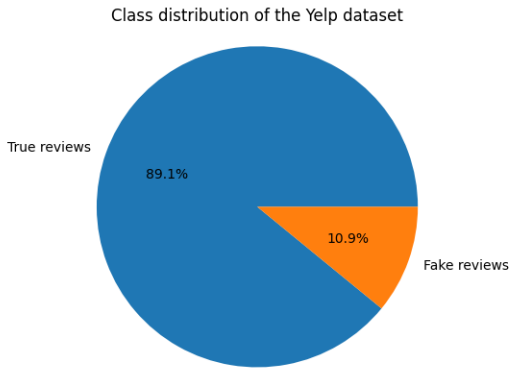

Figure 1: The distribution of the fake and true reviews in the Yelp dataset

C. Data preprocessing

Data preprocessing in text classification involves cleaning up the data in the dataset to improve the performance and speed up the time needed to train and test the models. In particular, the researcher applied tokenization, part of speech tagging, removal of stop words, and lemmatization.

1) Tokenization: It refers to the method of breaking up large units of text into smaller units of text called tokens, as described by the authors in [22]. The whole text was divided into a list of tokens by separating each word by a space. This is known as unigrams.

2) Part of speech tagging (POS tagging): It is the process of classifying each token into its corresponding part of speech based on its definition or context. For example, the word ”Chicago” is tagged as a proper noun. The purpose of this

step was to use the POS tags to compute several features, such as informality and content word diversity.

3) Removal of stop words: The author of [23] defines stop words as words that repeat in abundance in a language, such as articles, prepositions, and pronouns, among others, and it does not add any additional information to the text. Removing stop words is ideal in a text classification problem because it removes irrelevant information, helping the model run faster and perform better.

4) Lemmatization: As defined in [24], lemmatization is the removal of the inflection from words to retrieve their dictionary form. For example, if the words "swims" and "swam" undergo lemmatization, it'll return the word "swim" because it is the root word of both words.

Other text preprocessing techniques used include removing whitespace and converting the words into lowercase.

Table 1: Preprocessing the review text

Original text	Lowercase review text	Removed stop words review text	POS tagging	Lemmatized review text
Amazing! Go here!!!!!	amazing go here	amazing go	('amazing', 'VBG'), ('go', 'VB')	amaze go

Table 1 shows the different preprocessing steps that a review text undergoes. First, the text is transformed to lowercase, then the stop words are removed. After that, different parts of the speech are extracted from the review text. Finally, the review text undergoes lemmatization.

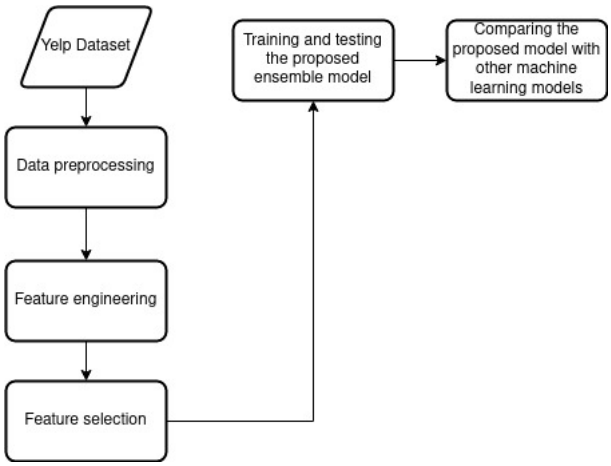

Figure 2: The flowchart of the methodology

D. Feature Engineering

Feature engineering is the process of extracting relevant features from a dataset. The features that were used in this study are linguistic and behavioral features and can be computed statistically. More specifically, these features are as follows:

Table 2: The list of basic statistical features

Feature name	Feature description
review_text_length	The length of the review text by counting all of the characters in it.
num_unique_words	Number of unique words in review text
num_total_words	Number of total words in a review text
ratio_unique_to_total_words	Ratio of unique words to total words in the review text
num_non_whitespace	Number of non-whitespace characters in a review text
num_sentence	Number of sentences in a review text
ave_word_len	Average word length of the review text
ave_sentence_len	Average sentence length of the review text
num_user_reviews	Number of reviews that a reviewer wrote in the dataset
num_unique_friends	Number of friends that a reviewer has in the dataset

The author of [12] used several linguistic features listed below.

Table 3: The list of features that used

Feature name	Feature description
num_verbs	It is the number of verbs in a review text
expressiveness	It is defined as the expressiveness of the review text and can be computed using: $\frac{no.of adjectives + no.of adverbs}{no.of nouns + no.of verbs}$
informality	It is defined as the tone of informality in the review text and can be computed using: $\frac{no.of typographical errors}{no.of words}$
lexical_diversity	It is the percentage of unique words in a review text
content_word_diversity	It is the percentage of unique content words in a review text.
redundancy	It is defined as the percentage of function words in a review text.

A study [25] used several linguistic and behavioral features to classify fake online reviews.

Table 4: The list of other linguistic and behavioral features

Feature name	Feature description
ratio_first_person_pronouns	The usage of first person pronouns such as "I", "me", "myself", etc. were measured by finding the ratio of the number of first person pronouns and the total number of words in the review text
ratio_capital_letters	The ratio of words that contains an uppercase letter and the total number of words in the review text
ratio_capital_words	The ratio of capital words and the total number of words in the review text
num_exclamatory_sentence	The number of exclamatory sentences found in the review text
rating	It is the rating given by the user alongside the review text
rating_deviation	It is the deviation of the rating given by the reviewer for a particular restaurant against the average restaurant review rating and normalized with the maximum deviation possible
max_number_reviews	It is the maximum review given by an author in a given day and normalized against the maximum number of reviews found in the data.

Lastly, a study [16] used a polarity feature to classify fake reviews.

Table 5: The list of features that uses polarity and similarity measurement

Feature name	Feature description
polarity	Polarity refers to a float value within the range of [-1.0, 1.0] where a negative float value indicates that the text is a negative statement and a positive float value means that it is a positive statement.

E. Feature selection

In an effort to improve the performance and training time of the ensemble model and to avoid the high dimensionality of the data, feature selection was used to narrow down the features to be used. The feature selection method to be used is Recursive Feature Elimination (RFE). Furthermore, this study used Recursive Feature Elimination with Cross-Validation (RFECV) to find the optimal number of features to be used in the study. In particular, the RFECV used f1-measure as its scoring measurement to address the imbalanced Yelp dataset.

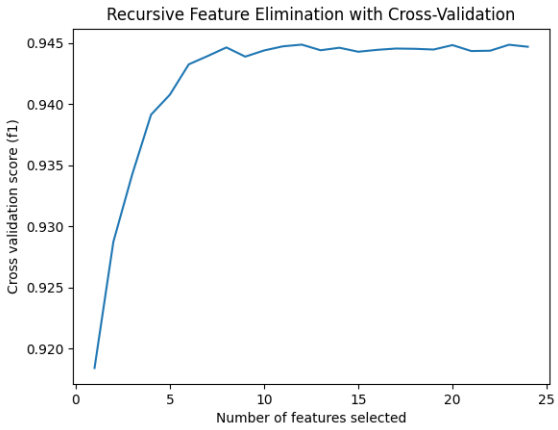

Figure 3: The result of RFECV for the English Yelp Dataset

Based on Fig. 3, the optimal number of features that the models use is twelve for the English Yelp dataset. Furthermore, Fig. 3 is not a linear graph, which means that using all of the features listed above slowed down the training time needed and affected the performance of the different models that the study used.

Table 6: The 12 features deemed as most important by RFECV for the English Yelp Dataset

Feature name
num_user_friends
num_user_reviews
review_text_length
num_non_whitespace
ave_word_len
expresiveness
content_word_diversity
redundancy
ratio_capital_letters
ave_sentence_len
polarity
informality

As shown in Table 6, the twelve most important features include both linguistic and behavioral features.

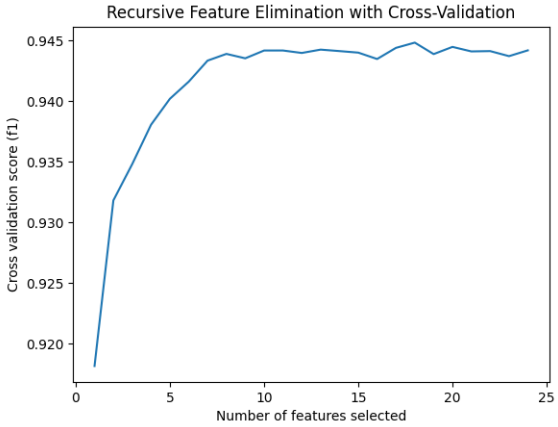

Figure 4: The result of RFECV for the Filipino Yelp Dataset

Based on Fig. 4, the optimal number of features using the

Filipino Yelp dataset is 18. In addition, the f1 score shows that the graph is not linear and that using a reduced number of features helped in training and testing the models.

Table 7: The 18 features deemed as most important by RFECV for the Filipino Yelp Dataset

Feature name
num_user_friends
num_user_reviews
review_text_length
num_unique_words
num_total_words
ratio_unique_tototalwords
num_non_whitespace
ave_word_len
ave_sentence_len
num_verbs
expresiveness
informality
lexical_diveristy
content_word_diversity
redundancy
ratio_first_personpronouns
ratio_capital_letters
polarity

As shown in Table 7, the twelve most important features include both linguistic and behavioral features.

Table 8: Example of the extracted features in the English Yelp dataset

Feature name	
Original review text	Amazing! Go here!!!!!
num_user_friends	63
num_user_reviews	16
review_text_length	8
num_non_whitespace	7
ave_word_len	3.5000
expresiveness	0.5000
content_word_diversity	1
redundancy	0
ratio_capital_letters	0.6667
ave_sentence_len	1.3333
polarity	1
informality	0.6667

Table 9: Example of the extracted features in the Filipino Yelp dataset

Feature name	
Original review text	Kahanga-hanga! Pumunta ka dito!!!!!
num_user_friends	63
num_user_reviews	16
review_text_length	12
num_unique_words	1
num_total_words	1
ratio_unique_tototalwords	1
num_non_whitespace	12
ave_word_len	12
ave_sentence_len	1.6667
num_verbs	1
expresiveness	0
informality	0.6667
lexical_diveristy	1
content_word_diversity	0.5000
redundancy	0.2500
ratio_first_personpronouns	0
ratio_capital_letters	0.5000
polarity	1

Tables 8 and 9 show the results of the feature extraction for a single sample of the review text found in the English and Filipino Yelp datasets.

F. Classification model

The ensemble model used a majority voting mechanism to determine the classification of the data and used three base models such as Decision Trees, AdaBoost, and Gradient Boosting. The choice of these base models was based on their performance when classifying imbalanced datasets as shown in [10].

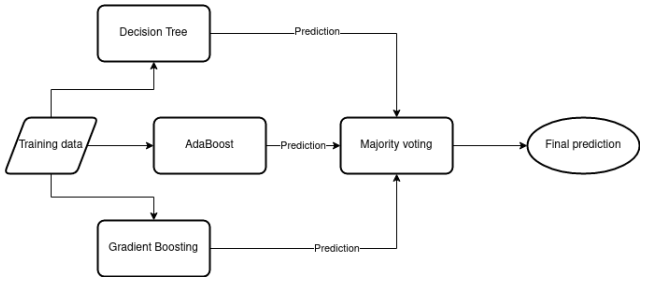

Figure 5: The flowchart of the ensemble model

G. Evaluation of the model

In order to evaluate the ensemble model, the F1-measure score and Area Under the ROC Curve (AUC) were used as the two main metrics to evaluate the performance of the model. Furthermore, the choice of using F1-measure for this study is suitable because it is a measurement that was designed to work well on imbalanced datasets. Also, using AUC as a metric is appropriate in this study because it does not depend on the distribution of the class and it is unaffected by the changes in the distribution of the class in the dataset.

The ensemble model was compared against other models that previous researchers have used as described in the literature review. The models evaluated and compared are the Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Logistic Regression (LR), and Naive Bayes (NB). This part of the study was done to give the researchers an idea of how the ensemble model compared against proven and tested models.

IV. Results and Discussion

A. Fake reviews

The ensemble model was used to classify fake reviews using the English Yelp and Filipino Yelp datasets. Listed below are the mean scores of the different features used for each dataset. Also shown below is the fake review text classified by the model as fake.

Table 10: Mean scores of the features extracted for the English dataset

Feature name	Fake reviews	Real reviews
No. of user friends	119.6129	97.4014
No. of user reviews	150.5937	139.9181
Review text length	362.9707	365.5412
Number of non-whitespace characters	307.1298	309.5855
Average word length	5.3749	5.4580
Expressiveness	0.4456	0.4592
Content word diversity	0.6734	0.6781
Redundancy	0.2824	0.2764
Ratio of capital letters	0.1206	0.1232
Average sentence length	16.5381	16.2065
Polarity	0.2630	0.2745
Informality	0.1461	0.1502

Based on Table 10, we can see that the ensemble model considers an English review text to be fake if the length of the review text is shorter, it uses fewer non-whitespace characters, it has fewer word length, is less expressive, is less diverse, is more redundant, has a longer length of sentences, is more negative, and uses more informal language. Also, another factor in classifying fake reviews is that the behavioral features, such as number of user friends and number of user reviews, are significant factors in classifying fake reviews more than the linguistic features.

Table 11: Mean scores of the features extracted for the Filipino dataset

Feature name	Fake reviews	Real reviews
No. of user friends	127.6376	95.9408
No. of user reviews	147.3098	139.5331
Review text length	441.7967	434.6687
No. of unique words	52.4880	51.7698
No. of total words	60.6734	59.4935
Ratio of unique words to total words	0.9078	0.9092
No. of non-whitespace characters	382.1232	376.1751
Average word length	6.4112	6.4836
Average sentence length	14.7613	14.2533
No. of verbs	7.7141	7.7082
Expressiveness	0.3847	0.3876
Informality	0.14622	0.1501
Lexical diversity	0.9078	0.9092
Content word diversity	0.3644	0.3646
Redundancy	0.3414	0.3429
Ratio of first person pronouns	0.0106	0.0099
Ratio of capital letters	0.1134	0.1134
Polarity	0.2637	0.2740

Based on Table 11, we can see that the ensemble model considers that a Filipino review text is fake if it is longer, uses more unique words, uses more words, uses more non-whitespace characters, less average word length, higher average sentence length, uses more verbs, is less expressive, less informal, less diverse, less redundant, uses more first person pronouns, and more negative in writing. Furthermore, another factor in classifying fake reviews is that the behavioral features, such as number of user friends and number of user reviews, are significant factors in classifying fake reviews more than the linguistic features. However, a caveat with this is that it may not represent how Filipinos actually write reviews online, given several problems with translations.

Table 12 shows some examples of the reviews classified by the ensemble model as fake reviews. We can see here that the texts use excessive punctuation marks. Some of it can be described as negative, and some of it can be described as overwhelmingly positive. Also, we can see several typographical errors in the review text. Furthermore, the number of user friends and the number of user reviews show that some of the users aren't active considering they have few friends and few reviews posted.

Table 13 shows some examples of the reviews classified by the ensemble model as fake reviews. We can see here that the texts use excessive punctuation marks. Some of it can be described as negative, and some of it can be described as overwhelmingly positive. In addition, we can see several unnecessary words such as "bit.ly" which seems to indicate a web URL, and there are several typographical errors in the review text. Furthermore, the number of user friends and the number of user reviews show that some of the users aren't active considering they have few friends and few reviews posted.

Table 12: Examples of English fake reviews

Fake review text	No. of user friends	No. of user reviews
probably the only american brunch place in this area. food was decent, but the service was odd. We had a party of 3, and there was a table for 4 open. We asked if we could be seated there, but the hostess kept asking us to wait so she could get permission from her boss if she could seat 3 people in a 4 person table. the restaurant didn't have a table made for 3 people, so where else would we sit?Overall, it's your typical brunch place.	127	237
Look at what was in my soup!!!!!!!!!!!!! A Roach! There is bad rumen, and then there is fucking horrible rumen. This is the latter. The absolute zenith of shitholery, MSG infested disgustingness.	0	1
Get your breakfast on!!!!!!Thank God we found this place, right around the corner from my hotel and from Madison Square Garden. Big booths which are great for kids ;). Big breakfast and a big belly when you leave. Get ready to get big!	56	197
I was a lost and hungry urban soldier, in the food desert that is west Chelsea, near Penn Station. For years, I've had to deal with reheated pizza and 1-star Chinese takeout when I'm in the neighborhood. Or that random Taco Bell. What a lowly state that area was in.But then, by the grace of god, Haymaker came down from heaven with a divine beer list, wings good enough to make your heart fly, and – what's that? – prices that don't hurt. I couldn't believe it.Just go, already!	70	2
i used to order here all the time but its not the place to eat if your trying to maintain your weight if you know what i mean.The food is fantastic The delivery service is terrible. I DON'T WANT TO ORDER LUNCH AND RECEIVE IT AT DINNER TIME.If there is a potential delay with the delivery then let people know. i am only two blocks a way but i work alone in my office so i cant leave. It is unacceptable to treat customers like this.I will not order here again. I GIVE BIG TIPS AND GET CRAP SERVICE NO WAY IM DONE.!!	9	16

Table 13: Examples of Filipino fake reviews

Fake review text	No. of user friends	No. of user reviews
Mayroon akong isang kinakailangan para sa mga weeknight dinner sa labas kasama ang mga kaibigan: dapat itong maging malusog. Ang mga katapusan ng linggo ay maaaring maging mas maluwag ngunit kulang ako upang manatili sa track tuwing linggo. Noong nakaraang linggo ay sinamahan ko ang aking kaibigan sa isang kaganapan sa Metropolitan Pavilion. Pagkatapos ng ilang oras na paghigop ng champagne at pagtikim ng cake, handa na kami para sa isang masustansyang pagkain. Sa kabutihang palad, nasa malapit lang ang isa sa mga paborito kong restaurant, ang Westville. Read more:bit.ly/JQvEpG	139	58
Tingnan mo kung ano ang nasa sabaw ko!!!!!!!!!!!!!!!! Isang Roach! Mayroong masamang rumen, at pagkatapos ay mayroong nakakatakot na rumen. Ito ang huli. Ang ganap na zenith ng shitholery, MSG infested disgustingness.	234	28
OO! A FAN! I LOVE their coffee selection @ The Library. MAHAL ang almond croissant. Ito ay kakaiba, tahimik, may magandang ilaw, puno ng mga libro! Ito ay isang magandang lugar para sa isang petsa kung tatanungin mo ako. Mapapaisip ka sa ka-date mo na napaka-"cultural" mo kung dadalhin mo sila dito... which you are of course. Ikaw AY sa yelp pag-aaral tungkol sa iba NYC buhay pagkatapos ng lahat.	26	109
hiyas ng isang bar! Magandang lugar para sa isang gabi ng date- intimate, malambot na ilaw, masasarap na hand crafted cocktail na gawa sa sariwang lahat (kahit ilang magagandang "mocktail" para sa iyo na hindi umiinom doon), napakagandang serbisyo......... Ang paborito kong bar sa ngayon sa The Big Apple! A+!	56	8
I used to order here all the time but its not the place to eat if you trying to maintain your weight if you know what i mean.The food is fantastic The delivery service is terrible. AYAW KO MAG-ORDER NG LUNCH AT MAKATANGGAP ITO SA DINNER TIME. Kung may potensyal na pagkaantala sa paghahatid pagkatapos ay ipaalam sa mga tao. I am only two blocks a way pero mag-isa akong nagtatrabaho sa opisina ko kaya hindi ako makaalis. Hindi katanggap-tanggap na tratuhin ang mga customer ng ganito. Hindi na ako mag-o-order dito. NAGBIBIGAY AKO NG MALAKING TIP AT NAKAKAKUHA NG CRAP SERVICE NO WAY IM DONE.!! !!! !!!!!!!!!!!!!!!!!!!!!!!!!!!!!	0	6

B. Evaluation of the ensemble model

The models were evaluated using a 5x2 cross-validation method. This method was chosen because of the imbalance in the datasets used. This works by splitting the datasets into two folds. Then, the models were trained on k-1 folds and tested out on the holdout fold. This is repeated four more times. This is to ensure that the model is thoroughly tested and validated.

Table 14: F1-scores of the models using the English Yelp dataset

Model name	F1-score
Support Vector Machine	0.5423
Decision Trees	0.6222
Random Forests	0.6726
Multilayer Perceptron	0.6065
Logistic Regression	0.5054
Naive Bayes	0.3993
Ensemble Model	0.6822

Table 14 shows the performance of the different models trained using the English Yelp dataset and evaluated using the F1-score. In addition, the models were all trained using the important features only. All models, except the Naive Bayes model, show that it performs well for classifying fake reviews. The model with the lowest F1-score is the Naive Bayes with a score of 0.3993. This means that the Naive Bayes method cannot reliably distinguish between fake and genuine reviews. In addition, the best performing model is the ensemble model, with a score of 0.6822. This means that the ensemble model can reliably predict the fake reviews.

Table 15: AUC scores of the models using the English Yelp dataset

Model name	AUC score
Support Vector Machine	0.7484
Decision Trees	0.6935
Random Forests	0.7703
Multilayer Perceptron	0.7708
Logistic Regression	0.7186
Naive Bayes	0.6417
Ensemble Model	0.7600

Table 15 shows the performance of the different models trained using the English Yelp dataset and evaluated using the AUC score. Furthermore, all of the models were trained using only the important features of the English Yelp dataset. All models show that it performs well for classifying fake reviews. The model with the lowest AUC score is the Naive Bayes, with a score of 0.6417. In addition, the best performing model is the Multilayer Perceptron with a score of 0.7708. This means that the Multilayer Perceptron is the most reliable in terms of identifying fake reviews. This can be attributed to the fact that Multilayer Perceptron is a deep learning algorithm. This means that they learn the hidden patterns within the given dataset. In addition, the ensemble model is the third best model in terms of AUC score. This means that the ensemble model can reliably separate the fake review from a real review, but it is not the best model in terms of separability.

Table 16: F1-scores of the models using the Filipino Yelp dataset

Model name	F1-score
Support Vector Machine	0.5346
Decision Trees	0.6172
Random Forests	0.6785
Multilayer Perceptron	0.6365
Logistic Regression	0.5890
Naive Bayes	0.4222
Ensemble Model	0.6859

Table 16 shows the performance of the different models trained using the Filipino Yelp dataset and evaluated using the F1-score. In addition, the models were all trained using the important features of the Filipino Yelp dataset. All models, except the Naive Bayes model, show that it performs well for classifying fake reviews. The model with the lowest F1-score is the Naive Bayes model, with a score of 0.4222. This means that the Naive Bayes classifier can not reliably predict fake reviews from real reviews. In addition, the best performing model is the Ensemble model, with a score of 0.6859. This means that the ensemble model can reliably predict the fake reviews.

Table 17: AUC scores of the models using the Filipino Yelp dataset

Model name	AUC score
Support Vector Machine	0.7440
Decision Trees	0.6778
Random Forests	0.7611
Multilayer Perceptron	0.7332
Logistic Regression	0.6730
Naive Bayes	0.6394
Ensemble Model	0.7441

Table 17 shows the performance of the different models trained using the Filipino Yelp dataset and evaluated using the AUC score. Furthermore, all of the models were trained using only the important features of the Filipino Yelp dataset. All models show that it performs well for classifying fake reviews. The model with the lowest AUC score is the Naive Bayes, with a score of 0.6394. In addition, the best performing model is the Random Forests with a score of 0.7611. This means that the Random Forests is the most reliable in terms of identifying fake reviews. This can be explained by the fact that Random Forests is a bagging type of model. This means that it was designed to improve the performance of its base estimator. Furthermore, the ensemble model is still the third best model in terms of AUC score. The above results show that the ensemble model performed the best among the other models evaluated using the F1-score because it uses multiple base estimators such as Decision Trees, AdaBoost, and Gradient Boosting in order to better predict fake reviews compared to other models. However, the proposed model doesn't fare well in terms of AUC ROC score despite having an acceptable score for that metric. This can be attributed to the fact that the ensemble model used Decision Trees as one of its base estimators and that this base estimator had a low score in terms of AUC score.

C. Statistical significance of the performance of the different models

In order to determine if the findings are statistically significant, they were further statistically evaluated using a 5x2 Cross Validation paired t-test. Furthermore, the t-value used is 2.571 and a p-value of 0.05. The results to be shown below are the summary of the other model's performance compared to the performance of the ensemble model.

Table 18: T-statistic and p-value of the compared models using the English dataset and using the F1-score

Model name	t-statistic	p-value
Support Vector Machine	-20.6754	0.000
Decision Trees	-6.5830	0.001
Random Forests	-1.1821	0.290
Multilayer Perceptron	-0.1105	0.916
Logistic Regression	-24.5294	0.000
Naive Bayes	-23.5311	0.000

Table 18 shows that the ensemble model's performance in terms of F1-score and using the English dataset is statistically significant compared to other models such as Support Vector Machine, Decision Trees, Logistic Regression, and Naive

Bayes. However, compared to other models such as Random Forests and Multilayer Perceptron, the findings suggest that the performance of the ensemble model is not statistically significant.

Table 19: T-statistic and p-value of the compared models using the English dataset and using the AUC score

Model name	t-statistic	p-value
Support Vector Machine	-1.3685	0.229
Decision Trees	-5.3569	0.003
Random Forests	0.1427	0.892
Multilayer Perceptron	0.5676	0.595
Logistic Regression	-2.7490	0.040
Naive Bayes	-8.0198	0.000

Table 19 shows that the ensemble model's performance in terms of AUC score and using the English dataset is statistically significant compared to other models such as Decision Trees, Logistic Regression, and Naive Bayes. However, compared to other models such as Support Vector Machine, Random Forests, and Multilayer Perceptron, the findings suggest that the performance of the ensemble model is not statistically significant.

Table 20: T-statistic and p-value of the compared models using the Filipino dataset and using the F-1 score

Model name	t-statistic	p-value
Support Vector Machine	-27.3476	0.000
Decision Trees	-12.0645	0.000
Random Forests	-2.5280	0.053
Multilayer Perceptron	0.0950	0.928
Logistic Regression	-14.1269	0.000
Naive Bayes	-48.4806	0.000

Table 20 shows that the ensemble model's performance in terms of F-1 score and using the Filipino dataset is statistically significant compared to other models such as Support Vector Machine, Decision Trees, Logistic Regression, and Naive Bayes. However, compared to other models such as Random Forests, and Multilayer Perceptron, the findings suggest that the performance of the ensemble model is not statistically significant.

Table 21: T-statistic and p-value of the compared models using the Filipino dataset and using the AUC score

Model name	t-statistic	p-value
Support Vector Machine	-0.6391	0.551
Decision Trees	-4.5875	0.006
Random Forests	1.3110	0.246
Multilayer Perceptron	-0.6620	0.537
Logistic Regression	-6.8157	0.001
Naive Bayes	-7.2581	0.000

Table 21 shows that the ensemble model's performance in terms of AUC score and using the Filipino dataset is statistically significant compared to other models such as Decision Trees, Logistic Regression, and Naive Bayes. However, compared to other models such as Support Vector Machine, Random Forests, and Multilayer Perceptron, the findings suggest that the performance of the ensemble model is not statistically significant.

D. Important features used

As shown in Tables 6 and 7, the important features used in both datasets can be classified into two categories: linguistic and behavioral features. The linguistic features include the length of the reviews; the number of unique and total words; the ratio of unique to total words; the number of non-whitespace characters, average length of words and sentences; the number of verbs; expressiveness; polarity; informality; and lexical diversity. The behavioral features include the number of friends and user reviews. We can see that most of the features used for classifying fake reviews came from the linguistic features given that the dataset omits other essential data that could have led to more behavioral features such as the date and time of the review. However, the two behavioral features are considered more important than the linguistic features since they help determine which reviews are fake or not.

V. Conclusion

It was shown that the ensemble model works reliably given that it had the highest F1-score amongst the other models tested against it in both English and Filipino datasets. However, the multilayer perceptron had the highest AUC score using the English dataset, while Random Forests scored the highest AUC score using the Filipino dataset. This can suggest that the other models are having a hard time separating a fake review from a real review. In addition, the difference between the performance of the ensemble model and the two models, Random Forests and Multilayer Perceptron, is found to be not statistically significant. On the other hand, RFECV was used to determine the important features that helped the models perform as well as they could. It was found that the important features for the English and Filipino datasets were composed of both review-centric and reviewer-centric features, with the majority of those features classified as review-centric features. But, the two reviewer-centric features help determine if a review is fake or not.

A possible venue for improvement is using a Filipino dataset that comes from well-known local e-commerce websites such as Lazada and Shopee so that it can actually capture how Filipinos write reviews. Another possible area of interest is using other estimators in the ensemble model, such as Random Forests, Multilayer Perceptron and other boosting algorithms such as XGBoost, in order to further improve the performance of the ensemble model. In addition, using unsupervised algorithms such as K-Nearest Neighbors and using Deep Learning algorithms such as Convolutional Neural Network can be used as alternative models to classify fake reviews.

References

[1] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake Reviews Detection: A Survey," *IEEE Access*, vol. 9, pp. 65 771–65 802, 2021, [Online; accessed 2021-11-13].

[2] Trustpilot, "Ratings and reviews in Above The Line advertising - Trustpilot Business Blog," https://business.trustpilot.com/guides-reports/build-trusted-brand/ratings-and-reviews-in-above-the-line-advertising, jul 5 2021, [Online; accessed 2021-11-13].

[3] WebsiteBuilder, "26 Mind-Boggling Online Review Statistics Facts for 2021," https://websitebuilder.org/blog/online-review-statistics/, jan 17 2020, [Online; accessed 2021-11-13].

[4] J. Valant, "Online consumer reviews: The case of misleading or fake reviews," https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2015)571301, [Online; accessed 2021-11-13].
[5] J. Pitman, "Fake Reviews Are a Real Problem: 8 Statistics That Show Why," https://www.brightlocal.com/learn/fake-reviews-are-a-real-problem-8-statistics-that-show-why/, nov 6 2019, [Online; accessed 2021-11-13].
[6] H. Tang and H. Cao, "A review of research on detection of fake commodity reviews," *Journal of Physics: Conference Series*, vol. 1651, p. 012055, 11 2020, [Online; accessed 2021-11-13].
[7] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. New York, New York, USA: ACM Press, 2010, [Online; accessed 2021-11-13].
[8] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, oct 5 2015, [Online; accessed 2021-11-29].
[9] Fakespot, "Fakespot," https://www.fakespot.com/our-mission, [Online; accessed 2021-11-13].
[10] R. M. Tischio and G. Weiss, "Identifying Classification Algorithms Most Suitable for Imbalanced Data."
[11] J. C. Rodrigues, J. T. Rodrigues, V. L. K. Gonsalves, A. U. Naik, P. Shetgaonkar, and S. Aswale, "Machine Deep Learning Techniques for Detection of Fake Reviews: A Survey," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2 2020, [Online; accessed 2021-11-13].
[12] F. Abri, L. F. Gutierrez, A. S. Namin, K. S. Jones, and D. R. W. Sears, "Linguistic Features for Detecting Fake Reviews," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 12 2020, pp. 352–359.
[13] A. M. Elmogy, U. Tariq, A. Mohammed, and A. Ibrahim, "Fake Reviews Detection using Supervised Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021, [Online; accessed 2021-11-18].
[14] X. Wang, X. Zhang, C. Jiang, and H. Liu, "Identification of fake reviews using semantic and behavioral features," in *2018 4th International Conference on Information Management (ICIM)*. IEEE, 5 2018, [Online; accessed 2021-11-13].
[15] H. Le and B. Kim, "Detection OF FAKE REVIEWS ON SOCIAL MEDIA USING MACHINE LEARNING ALGORITHMS," *Issues In Information Systems*, vol. 21, no. 1, pp. 183–194, 2020, [Online; accessed 2021-11-13].
[16] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, "Ensemble Machine Learning Model for Classification of Spam Product Reviews," *Complexity*, vol. 2020, pp. 1–10, dec 17 2020, [Online; accessed 2021-11-18].
[17] L. Gutierrez-Espinoza, F. Abri, A. Siami Namin, K. S. Jones, and D. R. W. Sears, "Ensemble Learning for Detecting Fake Reviews," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 7 2020, [Online; accessed 2021-11-19].
[18] Y. Ma and F. Li, "Detecting Review Spam: Challenges and Opportunities," in *Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*. IEEE, 2012, pp. 651–654, [Online; accessed 2021-11-13].
[19] M. Ott, Y. Choi, C. Cardie, and J. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 309–319.
[20] B. Cruz, *Filipino POS*, 2020 [Online]. [Online]. Available: https://github.com/jcblaisecruz02/filipino-pos
[21] K. Varadha Rajan, N. Ajagane, and S. Srivastav, "Evaluating Performance of Semi -Supervised Self Training in Identifying Fake Reviews," Tech. Rep.
[22] A. Pai, "What is Tokenization," https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/, may 25 2020, [Online; accessed 2021-11-18].
[23] C. Khanna, "Text preprocessing: Stop words removal," https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a, feb 10 2021, [Online; accessed 2021-11-18].
[24] "Stemming and lemmatization," https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html, [Online; accessed 2021-11-18].
[25] J. Fontanarava, G. Pasi, and M. Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 10 2017, [Online; accessed 2021-11-26].

Angelo L. Guan Angelo L. Guan is a senior BS Computer Science from Pagsanjan, Laguna.