WEIGHTED ASPECTS FOR SENTIMENT ANALYSIS

by

Byungkyu Yoo

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Julia Taylor Rayz, Chair

Department of Computer and Information Technology

Dr. John A. Springer

Department of Computer and Information Technology

Dr. Baijian Yang

Department of Computer and Information Technology

Approved by:

Dr. John A. Springer

Dedicated to my four family.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my academic advisor Dr. Julia Taylor Rayz. I am sincerely grateful for all the advice, support, guidance, feedback, encouragement, and all other help you provided. I would not have successfully completed my degree if it was not for all your help.

I would also like to thank my committee members Dr. John A. Springer and Dr. Baijian Yang for their valuable feedback and support.

I would also like to thank my AKRaNLU lab members for all the feedback and help that they gave me with my thesis and my degree.

Finally, I would like to thank my family and friends for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

When people write a review about a business, they write and rate it based on their personal experience of the business. Sentiment analysis is a natural language processing technique that determines the sentiment of text, including reviews. However, unlike computers, the personal experience of humans emphasizes their preferences and observations that they deem important while ignoring other components that may not be as important to them personally. Traditional sentiment analysis does not consider such preferences. To utilize these human preferences in sentiment analysis, this paper explores various methods of weighting aspects in an attempt to improve sentiment analysis accuracy. Two types of methods are considered. The first method applies human preference by assigning weights to aspects in calculating overall sentiment analysis. The second method uses the results of the first method to improve the accuracy of traditional supervised sentiment analysis. The results show that the methods have high accuracy when people have strong opinions, but the weights of the aspects do not significantly improve the accuracy.

1. INTRODUCTION

Review sites are a platform for consumers to share their experience of businesses, including restaurants and their food, service, location, and more. For business owners, these reviews are critical for improving customer experience. With the development of computers, techniques such as sentiment analysis have started to be used to predict the sentiment of such reviews. Using sentiment analysis, the overall sentiment of the restaurants can be found. However, computers don't evaluate the review as humans do. Unlike computers, humans have preferences. People write reviews and give a rating on the business based on what they experienced and what they value. Even if some features are bad, if the restaurant has good quality features that people value more, the restaurant will get a higher rating. For example, if the food is extremely good in a restaurant, but sitting arrangements are slightly uncomfortable, there is still a high chance that the restaurant will get good reviews because most people care about food more than the chairs that they sit on [1]. Another example is a restaurant might have the many good quality features like location, seating, table, music, and other things, but if the food is uneatable, there is a lower chance their restaurant will get a good review. Like the examples, many features may be good, but the overall number of good features may be outweighed by a single one, especially if it is salient to the establishment, like food for restaurants.

Traditional supervised sentiment analysis methods don't consider these personal and cultural values and preferences that humans have. There are methods to extract aspects of business [2] [3]. There are methods to do sentiment analysis using aspects of the reviews [4] [5]. There are methods to see which aspects people tend to care about [6] [7]. There are sentiment analysis methods that give higher weights to important parts [8] [9]. However, there has been very little research done that uses human preference to do sentiment analysis.

The motivation behind this research comes from the thought that, unlike computers, humans care more about certain aspects than other things. If there is a restaurant with the best food in the world, but the location is a far from home, most people will still write a positive review while mentioning mediocre location. However, when doing traditional supervised sentiment analysis, the sentiment may be neutral because computers do not have

a treat both food and location with the same weights. This research attempts to explore the issue by adding human preference into the sentiment analysis.

This research proposes a new approach on conducting domain-specific sentiment analysis using weighted aspects. The proposed methods use the idea that some aspects are considered more important to humans than others aspects. Using the idea, the proposed methods will incrementally add weights to different aspects which will be used in to calibrate the overall sentiment. This work also provides a method to use the weighted aspects as features in supervised sentiment analysis to increase the overall sentiment analysis accuracy. Finally, this paper compares the accuracy of traditional sentiment analysis and newly proposed salient-aspect-based sentiment analysis methods.

1.1 Research Questions

This thesis attempts to answer the following question:

- Can weighted aspects be used in predicting the overall sentiment of a review?

To answer the main research question, two additional questions are considered:

1. Does applying weights to unsupervised learning significantly increase the overall sentiment of a review?

2. Does using weighted aspects as a feature in traditional supervised methods significantly increase the accuracy in predicting the overall sentiment of a review?

Yelp dataset [10] is used to answer the research question.

1.2 Assumptions

The assumptions of this thesis is:

- Yelp dataset is an accurate representation of the users opinion with an insignificant amount of fake data.

- Four aspects used in previous literature [11] [12] [13] [14] [15] [16] will be an accurate representation of users opinion and provide the highest accuracy

- The aspect_based_sentiment_analysis library provides acceptable accuracy of aspect sentiment analysis

1.3 Delimitation and Limitations

The delimiation of this thesis is:

- This research will only use the restaurant review data in English from North America.

The limitation of this thesis is:

- The analysis and result of the research was based only on the restaurant review data in North America at the period the data was collected.

1.4 Organization of this Thesis

The paper is organized in the following order. The current section, chapter 1 contains the motivation and the questions this paper attempts to answer. Chapter 2 presents the existing research on this and related topics. It explains the sentiment analysis, aspect, and restaurant review related research that have been conducted. Chapter 3 presents the research methodology. There are 2 types of experiments, each consisting of 3 methods that have been described. Chapter 4 provides and explains the results of the project. Chapter 5 gives the conclusion and future work needed to be done.

2. BACKGROUND

Knowledge from previous research is required to do sentiment analysis with weights. Research on the best way to do sentiment analysis and aspect-based sentiment analysis was summarized for our research to be done. To understand how to give higher weights to aspects that are more valued by people, research on weighted sentiment analysis was summarized. To understand how to know what aspects are more important, research on aspect ranking was summarized. To know what people value more in restaurant reviews, research on restaurant aspects was summarized. To understand how other research used features for sentiment analysis to get the overall accuracy, the research on it was summarized.

This chapter provides a literature review of the papers relevant to answer the research questions. This section include sentiment analysis, aggregation of sentiments, aspect rankings, and restaurant weights.

2.1 Sentiment Analysis

Sentiment analysis is the task of "extracting and analyzing peoples opinions and sentiments from text" [17]. A sentiment is an "opinion that a person expresses towards an aspect, entity, person, event, feature, object, or a certain target" [17]. Sentiment analysis poses as a powerful tool for businesses, researchers, and government agencies to extract and analyze the public's mood and views to gain business insight and make better decisions [18]. There are generally 3 levels of sentiment analysis [19]: aspect level, sentence level, and document level. The aspect level finds the sentiment of a specific aspect. The sentence level determines the sentiment of a sentence. Document level applies weighting rules to sentences and sums the sentiment or produces an overall sentiment for the document.

There are various methods to do sentiment analysis including lexicon-based, rule-based, machine learning, and deep learning methods [19]. Many papers have been written analyzing sentiment on the Yelp dataset as well with various techniques. For instance, using the Naive Bayes algorithm, S. and Ramathmika [12] got an accuracy of 78%. Using SVM, Yu, Zhou, Zhang, *et al.* [20] got an accuracy as high as 88%. Using CNN, Cheng, Yao, Xiang, *et al.* [21] got an accuracy of 91%. Using LSTM, Barry [22] got an accuracy of 94%.

The current trend in sentiment analysis techniques, discussed below, is starting to incorporate the fusion of information from text with social and cultural contexts. Social context has also been introduced in other fields related to sentiment analysis, such as spam detection, where clues to identify spammers are usually hidden in multiple aspects of context, such as previous content, behavior, relationship, and interaction [23]. Sentiment analysis has also been used to determine opinions about different cultures and languages' views on the topic of COVID-19 and how the sentiments change over time [24]. Sentiment analysis can not only help identify cross-cultural trends but also can link actual events to users emotions that are expressed on social platforms. Despite the socioeconomic and cultural differences between the United States and Pakistan, there was a high similarity of sentiments that were shown on the topic of the COVID-19 pandemic. On the other hand, Sweden and Norway are very close to each other and share a very similar culture. However, the two countries showed very different sentiments on the topic of COVID-19, mainly assumed to be because of the difference in lockdown policy [25].

2.2 Aspect Based Sentiment Analysis

From the three levels of sentiment analysis, the aspect level is the sentiment analysis that our research focuses on. Aspect-based sentiment analysis is a type of sentiment analysis that predicts the sentiment of an aspect. Aspect is defined as an "attribute or a feature an item possesses" [26]. Aspects are not limited to just judgment but also include perspectives, thoughts, ways of thinking, points of view, and social influence [27]. It is commonly used to analyze the public's sentiments over time and across different contents [28]. Applications include aspect-level drug reviews [29], scientific reviews [30], restaurants, products, and hotel reviews [31]. The field of aspect-based sentiment analysis has been done using various methods and has had many changes [27]. There are usually three steps to aspect-based sentiment analysis. While the details of the steps may vary, the names may be different in some publications, and some papers even skip some of the steps, generally, most agree on the three steps:

1. Aspect Detection

2. Aspect Sentiment Analysis

3. Sentiment Aggregation

The aspect detection step extracts the aspect from the sentence or document. It includes explicit aspects, implicit aspects, aspect terms, and entities [27]. Some papers may include categorizing the aspect and finding the sentiment target pair in this step. The aspect sentiment analysis step classifies sentiment polarity for a predefined aspect, target, or entity. The last step, sentiment aggregation step, aggregates the sentiment values for each aspect to provide a concise overview.

There are various methods to do doing each step. The different types of methods to do the first two steps are shown in the Figure 2.1.

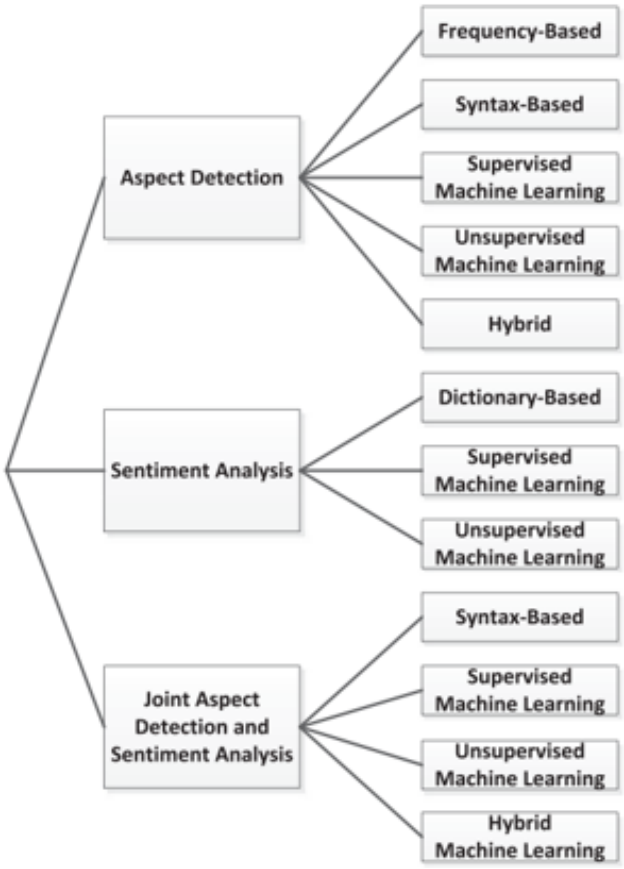

Figure 2.1. Methods of Aspect Based Sentiment Analysis by Schouten and Frasincar [32]

2.2.1 Aspect Detection

There are 5 different overall classes of aspect detection: frequency-based methods, syntax-based methods, supervised machine learning methods, unsupervised machine learning methods, and hybrid methods. The frequency-based method uses the idea that a limited set of words is used much more often than the rest of the vocabulary. These frequent words, usually nouns, are likely to be aspects [33]. The disadvantage of this method is that not all frequent nouns are actually aspects [32]. For example, if there is a sentence saying, 'I bought a good car for only $5 dollars', the aspects are car and dollars, assuming both are frequent nouns. However, the word dollar is not really an aspect of the sentence. Syntax-based methods find aspects by means of the syntactical relations of "adjectival modifier relation between a sentiment word and an aspect" [32]. An example is a word pair 'good food', a sentiment-adjective, and aspect-noun pair. Compared to the frequency-based method, the syntax-based method can find low-frequency aspect, but many syntactical relations is required [32]. The third method is the supervised model. The common method uses some combination and variant of Conditional Random Fields (CRF) [34] or Hidden Markov Modeling (HMM) [35]. The fourth method is using unsupervised models for aspect detection. The most common method is using LDA-based methods [32]. LDA assumes that documents are a mixture of topics, and topics are a mixture of tokens. Therefore, similar topics have similar tokens [36]. LDA was originally built for document-level topic modeling, so other methods are used with LDA to find local topics like syntactic dependencies [37] and emphasize aspects that have a higher influence [38]. More recent unsupervised models are models that use BERT language model [39] [40] [41]. The last methods are hybrid models. Serial hybridization uses the output of one phase, like the frequency of words, to form the input for the next phase, like a clustering algorithm. Parallel hybridization is simply using two or more methods to find complementary sets of aspects [32].

2.2.2 Aspect Sentiment Analysis

Aspect sentiment analysis itself (as apposed to aspect detection, without sentiment classification) is the second step, and it is categorized in three different ways. Dictionary-based

methods use generated sentiment dictionary to find the sentiment of relevant words, and then finds combines the scores [32]. For example, if the aspect and sentiment pair of 'good food' has been detected in the aspect extraction phase, the word good is looked up in the sentiment dictionary to get the sentiment score of the aspect. Supervised models classify sentiments by learning separate representation for each data object. LSTM, GRU, and CNN are the commonly used algorithm like many natural language processing tasks [27].

There are also hybrid approaches to detect aspects and do sentiment analysis on aspects in one step. The types are syntax-based methods, supervised machine learning, unsupervised machine learning, and hybrid machine learning. Which are combinations of the approaches that have already been explained. An example of a hybrid approach is using the aspect-based sentiment analysis and aspect sentiment analysis by using ontology [2]. It is a type of syntax-based approach that requires more manual work. Because creating an ontology requires so much labor work, it is not applicable to all datasets. However, if the dataset and target aspects are clear this method can be used.

2.2.3 Sentiment Aggregation

The last step of the aspect-based sentiment analysis, the sentiment aggregation step, is relatively simple. There are three methods to this part. All three methods are similar. The first method is a simple count of the number of positive sentiments minus the number of negative sentiments. If the total number is positive, the aspect's sentiment is positive. If the total number is zero, the aspect's sentiment is neutral. If the total number is negative, the aspect's sentiment is negative. Aggregating sentiment in an ontology-based sentiment in document-level sentiment-based analysis [42], aggregating stock sentiment per day [43], aggregating weighted sentiment for machine learning classification [44], and many other papers used this method as this was the most common method to aggregate sentiment per aspect. The second method adds to the first method. The result from the first method is divided by the total number of data that the dataset has. The detailed additional math the paper does varies, but this method is used in a paper that also calculates the degree of the sentiment polarity. Pandi, Dandibhotla, and Bulusu [45] proposed a method to evaluate a person's

reputation using this method . The last method is by using the dominant polarity class. If the positive messages significantly outnumber the negative ones, the overall sentiment is considered positive and vice versa as well [46].

2.3 Weighted Sentiment Analysis

Aspect-based sentiment analysis may use weights to aggregate the data. The type of sentiment analysis that uses weights to get the overall sentiment is generally document-level sentiment analysis. Bhatia, Ji, and Eisenstein [8] introduced a method to do document-level sentiment analysis on movie reviews . In the aggregation of sentiment phase to get the overall sentiment, the researchers used weights based on the contribution of each discourse unit. The discourse unit could be a sentence or a part of a sentence that has a role in the document such as conjunction, justify, or elaboration of the document. The weights were decided based on their position in a dependency-like representation of the discourse structure. The weights can be defined using a simple function or learned from small data. By using weights to the sentiment, the accuracy increased by 0.5% to 3%.

Singh, Piryani, Uddin, *et al.* [9] used a movie review dataset. The authors used part-of-speech (POS) tagging and SentiWordNet for sentiment analysis. At the sentiment aggregation step, that research also used weights. Weights were done based on the sentiment and on a grammar rule they made. A small portion of the rule is that if the adverb is affirmative and the sentiment of the adjective is positive, add the minimum score of the adjective and adverb. The paper reported that the adjective & verb set should have 30% weight and the adverb & adjective set should have 35% weight. Using the weights of the sentiment, the overall accuracy increased by 1.3%.

2.4 Aspect Ranking

Many papers proposed methods to retrieve aspect rankings. These aspect rankings can be used to determine the weights of the method this paper will present. There are 4 different ways aspect ranking can be done. The first method is a simple frequency-based method. The higher frequency simply has higher weight [47]. The disadvantage is that infrequent

aspects are ignored [7]. The method that we will use to detect aspect ranking is using this method. Since our paper already has predefined aspects, the disadvantage will not apply to our research (see Restaurant Aspects section). The next method is the vendor-provided product information-based method. Aspect weights are decided by the vendor. Many times vendors choose from the perspective of marketing rather than experience. Since the focus of the vendor is selling points, this method is biased [48]. The next method is the syntax-based method. The more modifier words, such as adjectives or adverbs, there are for an aspect, the more important the corresponding aspect will be [49]. The last model is the econometric-based model. This method uses a linear model with features such as demand, product, and time [50]. The downside of this method is that it fails to consider the impact of sentimental strength on consumption intention and also cannot be scaled up to high-dimensional data space [7]. Some of these methods can even be done to get the sentiment ranking, a method normally used for document-level sentiment analysis [51].

There have been several papers that detected aspect ranking from online reviews. Wang, Wang, and Song [7] used an econometric model on Amazon review data to detect aspect ranking so that the ranking can be compared to the influence on the SalesRank. This paper used the information gain theory. The summarized version is this: users will try to gain the information they care about. If there is a positive sentiment for an aspect they are looking for, they will buy it. The research measured the aspect ranking by the number of usefulness and uselessness of review. The paper tracked 386 digital cameras for 39 months consecutively on Amazon to validate the performance of the model. The result was that the aspect ranking had a strong correlation with the actual sales orders. Zha, Yu, Tang, *et al.* [52] did a form of weighted aspect-based sentiment analysis to prove that the aspect ranking matters . The research did aspect-based sentiment analysis using a syntax-based method to extract aspects and SentiWordNet for sentiment extraction. The research only used 100 reviews of 20 products with positive and negative sentiments. A comparison was done only with Boolean weighting. It is a method that basically treats all aspects equally important. As a result, the research got an average of 71% accuracy, which is 3% higher than the Boolean weighting method.

2.5 Restaurant Aspects

This paper will use a predefined aspect for restaurants. There have been already many aspect-based sentiment analyses done using the restaurant domain as well as the Yelp domain to reference from. Sydorenko, Kravchenko, Rychok, *et al.* [53] used 5 aspects of food, service, price, ambiance, and overall. Panchendrarajan, Ahamed, Sivakumar, *et al.* [14] used food, service, ambiance, price (which include discount, worthiness), and restaurant (general). The widely used dataset from SemEval2015 [54] and SemEval2015[31] labeled the aspect-based sentiment analysis task with labels of food, drinks, service, ambiance, location, restaurant (with subcategories of general, price, and others). However, looking into the distribution in Figure 3.6 below, the number of aspects related to drinks and location is relatively lower than other aspect categories.

Figure 2.2. Distribution of Aspect Category by Pontiki, Galanis, Papageorgiou, *et al.* [31]

Research that used Yelp dataset found similar aspects. Luo and Xu [55] used LDA probabilistic model to find the aspect frequency and found out there were generally 4 topics. The 4 topics were food, experience (service related), Value (price related), and location related. Prithivirajan, Lai, Shim, *et al.* [13] found 4 aspects using the dataset which were food, service, ambience, and price. The paper showed that if the aspect price exist in the dataset, the rating were relatively neutral, while if the aspect food was mentioned the rating was relatively high. Panchendrarajan, Murugaiah, Prakhash, *et al.* [1] applied several method to find the aspect rankings of these. The ranking of the aspects were similar to the distribution of dataset from SemEval. The order of the aspects were food, service,

restaurant, ambience, and price in that order. The aspect price was used as 2 separate aspects, worthiness and offer, instead using it as 1 aspect.

2.6 Aspect Based Rating Prediction for Yelp Customer Review

A somewhat similar experiment to our approach was explored by Peng [56]. The author tried to use the sentiment of the aspect to find the overall sentiment of the Yelp review. Although the research did not use weights, the papers used the idea of utilizing features to predict the overall sentiment. The focus of the research was to add features, and the sentiment of the aspects was 1 of the features the researcher added. The research used 4 predefined aspects. The 4 aspects were policy (service), location, food, and environment (ambiance). For the aspect extraction step, the research did a simple grammar-based method from 2009 to detect the aspect-noun and sentiment-word pair. For the aspect sentiment analysis, the researcher used a dictionary based method to detect the sentiment polarity. The accuracy using only the aspect's accuracy was 67.63% and using traditional machine learning combined with the aspect's sentiment was 83.39%, a 0.1% increase if the aspect's sentiment was not used. Although the accuracy increase is not high, this paper gave the idea that the concept the research tried may work.

2.7 Summary

While much work has been done on sentiment analysis, three gaps in the research, listed below, can be identified:

1. Most aspect based sentiment analysis aggregates the sentiment by aspect using simple counts of sentiment. However, extremely few papers combine aspect sentiment to get the overall sentiment.

2. There are papers that give weights to words, sentences, and sentiment. However, current research do not give weights to aspect.

3. There are methods to find the ranking of aspects. However, current research do not use the aspects ranks as weights to do analysis.

3. METHODOLOGY

This chapter describes methodology applied to answer the research questions.

3.1 Dataset

The main dataset that was used was the Yelp dataset. The Yelp dataset was an open dataset created in 2013 that could be used for academic purposes. It contains the user reviews, stars, business type, restaurant id, and more information from the actual Yelp website. The original dataset contained 8.6 million reviews of businesses. Like with most review websites, the Yelp dataset was not a perfect dataset. The dataset contained some fake reviews [57], but this papers will assume the impact was not significant. The dataset was also not balanced [57]. It contained more positive sentiments than negative sentiments. The percentage of each star in the dataset is shown in Figure 3.1 below. The left graph is the original 5 stars in the Yelp dataset. The right graph is the percentage when star 1 and 2 is combined to negative and star 4 and 5 is combined to positive.

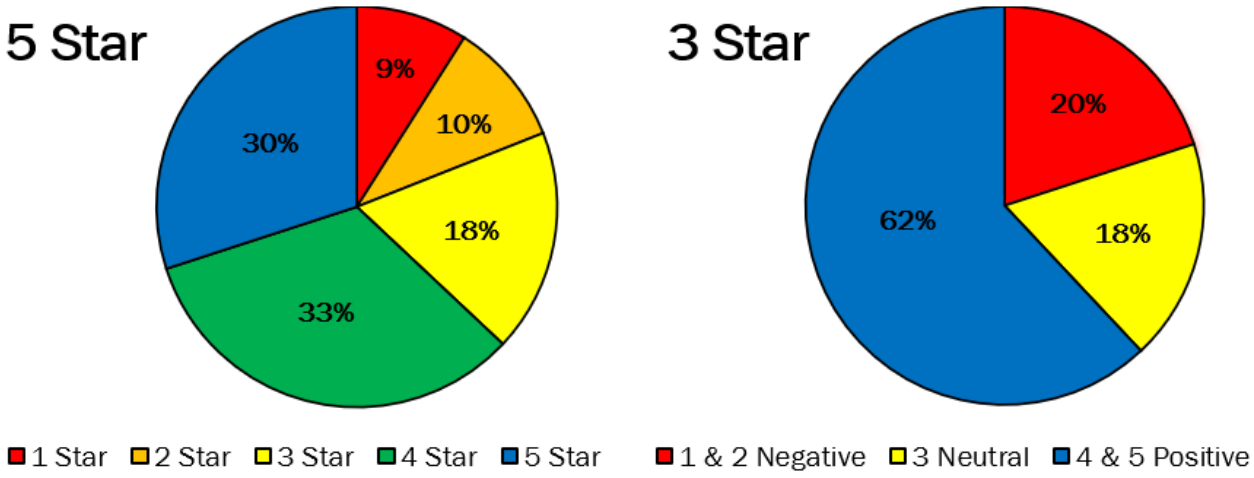

Figure 3.1. Dataset Sentiment Distribution

From the 8.6 million reviews of various businesses, the dataset was filtered to have only restaurant reviews resulting in 3.6 million reviews. Of the 3.6 million reviews, only 100,000 reviews were used due to the limited computing power available. For supervised classification, 80,000 reviews were used for training and 20,000 reviews were used for testing. For

unsupervised classification, all 100,000 reviews were used. Only the text of the review and star ratings was used and the remaining data was deleted. A dataset of 100,000 reviews is still a large enough dataset to produce accurate results for sentiment analysis [58].

3.2 Overview of Methodology

The research was divided into three steps. The first step was the data preprocessing to prepare all the data for classification. The second step was the three types of sentiment analysis. The first type was the three unsupervised classification that this paper presents. The second type was the supervised classification. Traditional supervised sentiment analysis used common machine learning algorithms. The third type, shown in the middle of Figure 3.2, was the 3 combination of supervised and unsupervised classification in attempt to increase the overall accuracy. The third and last step was simply comparing the results. The overview of the methodology used was displayed below in Figure 3.2.

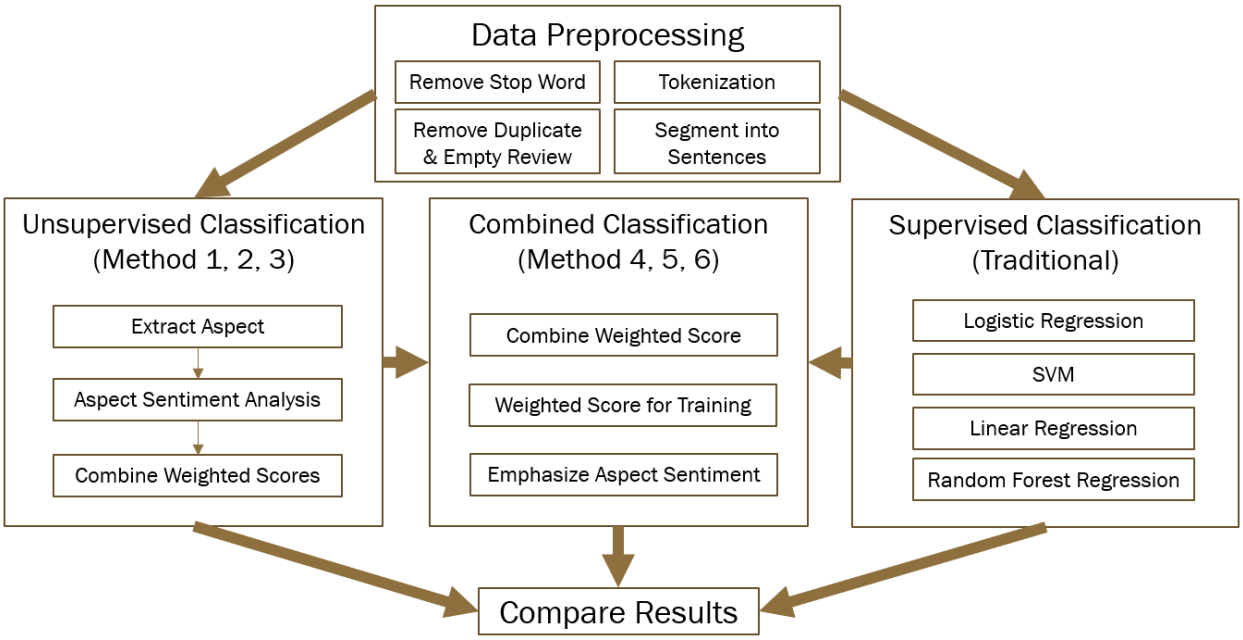

Figure 3.2. Overview of Methodology

In total, 6 methods were experimented in this paper. Methods 1, 2, and 3 used unsupervised classification to predict the star rating of the review. These methods answered whether unsupervised learning of weighted aspect based sentiment analysis can outperform

supervised classification. Methods 4 and 5 used a combination of unsupervised classification and supervised classification to predict the star rating of the review. Method 6 emphasized the sentiment of each aspect during the testing of the supervised classification. Methods 4, 5, and 6 answered whether the weighted aspect based sentiment analysis can be used to increase the accuracy of supervised classification. A short overview of the 6 methods is written below in the order of methods.

1. Combine Sentiment of Aspect in Each Sentence
2. Combine Sentiment of All 4 Aspects in a Single Review
3. Combine Sentiment of Relevant Aspects in a Single Review
4. Combine Results of Unsupervised and Supervised Classification
5. Utilize Unsupervised Results in Supervised Classification
6. Emphasize Relevant Aspect in Supervised Classification

3.2.1 Data Preprocessing

In this step, the dataset was preprocessed before the actual sentiment analysis and classifications. The Yelp dataset was downloaded from the Yelp website. The Yelp dataset contained reviews not only from restaurants but also from other businesses. Only the reviews with a business type of restaurant were kept. The remaining 5 million reviews were excluded. The data was again filtered to have only the review and the star label. Any duplicate reviews or reviews with a length of fewer than 10 characters were removed. Reviews with no alphabet were also removed. The stop words were removed from the review. The list of English stop words was retrieved from the NLTK library. The reviews were then tokenized using the BERT tokenizer.

3.2.2 Supervised Classification

Traditional sentiment analysis was used as a baseline comparison with the classification that this paper presented and was also used as a part of Methods 4, 5, and 6. Widely used

machine learning algorithms (Logistic Regression, SVM, Support Vector Regression, Linear Regression, and Random Forest Regression) were used.

3.2.3 Aspect Extraction

The first method to extract aspects was comparing the cosine similarity of the aspect's keyword embedding. For method 1, the aspect's word embedding was compared to each sentence's embedding. The aspect keyword that had the highest semantic similarity was chosen as the aspect for that sentence. For method 2, the aspect's word embedding will be compared to the whole review's embedding. The word embedding of an aspect word was compared to the whole review in method 2. An example of method 1 is shown in Figure 3.3. A sentence embedding in a review was compared to the aspect's keyword embedding 'food' and that score was 0.5631. That score was considered the similarity score. The aspect that had the highest score was considered the aspect for that sentence. On the other hand, method 2 considered all 4 sentences as 1 thing and extracted the aspect. The details of the 4 predefined aspects that were used will be explained in section 3.2.5 and the details of the methods will be explained in section 3.3 and section 3.4.

Sentence	Food	Service	Ambiance	Price	Aspect	Sentiment
I ordered the grilled chicken combo but received fried chicken.	**0.5631**	0.3956	0.3732	0.4056	Food	Negative
The inside is dirty in the afternoon like they never clean.	0.5098	0.4981	**0.5602**	0.4955	Aambience	Negative
The people behind the counter are usually polite and fast.	0.5191	**0.5282**	0.5193	0.5035	Service	Positive
I have never had an order messed-up.	0.4992	**0.5852**	0.5232	0.5787	Service	Negative

Figure 3.3. Sentence Embedding and Aspect Embedding Similarity by Alamoudi and Alghamdi [11]

The process of aspect extraction follows a part of the paper proposed by Alamoudi and Alghamdi [11]. The 4 aspects that were used were food, service, ambiance, and price. The 4 aspects were retrieved by k-mean clustering. This step was not repeated as the original paper by Alamoudi and Alghamdi [11] had already been stated in the research. Defining the 4 aspects, a pre-trained Glove model in the spaCy library was used to get the word embedding

of the 4 aspects. The pretrained model had more than one million unique vectors. Each vector had 300 dimensions and represents one word [59]. After loading the word embedding model, the sentence or review embedding was obtained. Method 1 used the embedding of each sentence that was divided in the preprocessing step. Method 1 combined the sentiment of aspect in each sentence. Method 2 used the embedding of the whole review. Method 2 combined the sentiment of all 4 aspects in a single review. The details of the methods will be explained in section 3.3 and section 3.4. The Glove embedding of each word was averaged by sentence or review. New words that were not in the corpus were replaced as zeros. The result of the sentence embedding and text embedding was a single 300-dimension embedding. It was the same shape as any word in the Glove word embedding model. This conversion of sentence to embedding used the SentenceTransformer library. The last part, calculating the semantic similarity score, was calculated to see the similarity of the aspect word vector embedding with the sentence or review embedding. The calculation was finding the cosine similarity of each dimension. The final cosine similarity score was a number from 1 to -1. If the cosine similarity score was closer to 1, it meant that the word and sentence or the word and review as similar.

Other extraction methods that the paper attempted to use included attention BERT based models, manual aspect extraction, grammar-based POS tagging, and LDA-based methods. However, none of these resulted in a meaningful aspect extraction. Therefore, aspects were extracted manually using a predefined set of words for method 3. Method 3 will be explained section 3.5. Using the predefined set of words, the keywords were simply just searched to see if they existed in each review. For example, if the word 'service' was contained in the review, the aspect service was considered as an aspect in the review. The search in the word 'service' included 'services' as the word service was still searchable in the review.

3.2.4 Aspect Sentiment Analysis

The aspect sentiment extraction step used the aspect-based-sentiment-analysis library [60]. This library combined many publicly available codes on Github so that it could be

used easily by users. The library was a combination of BERT based model [39] and fine tuning models [40] [41]. The basic logic was creating auxiliary sentences and doing sentence pair classification. The accuracy was not the highest compared to other published papers. However, the accuracy was within 5% difference with other published papers. Given a specific keyword, the aspect-based-sentiment-analysis library could extract the sentiment of the word, even if it did not exist in the review. An example is shown below as Figure 3.4. It had the keywords 'food' and 'service' with the review of 'The food is great but the staff was horrible'.

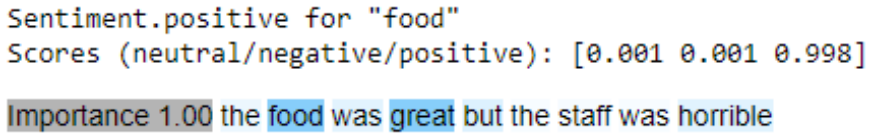

Figure 3.4. Aspect Sentiment of Food

The example above showed the sentiment of the aspect food. The sentiment of the word 'food' was successfully extracted as positive even though the overall sentiment was not as positive. The library could also find the words that had the most influence to the aspect's sentiment. The words 'food' and 'great', highlighted in blue, had the highest influence to the aspect food. The library could also detect an aspect's sentiment even if the actual word was not present in the review. An example is shown below in Figure 3.5.

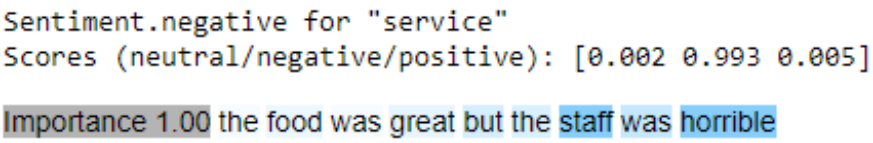

Figure 3.5. Aspect Sentiment of Service

The above example used the same review as with the aspect food. Although the sentiment of the aspect food was positive, the library could successfully detect that the sentiment of service was bad. Additionally, the library could detect the sentiment of an aspect even if the exact word was not present in the text. It could recognize that the words 'staff' and 'horrible' was highly influential to the sentiment of the aspect service.

3.2.5 Aspects and Weights

Four predefined aspects were chosen. The four aspects were the most commonly used aspects in many papers that used the restaurant and Yelp datasets. An aspect is like a category with 2 or 3 keywords. If any of the keywords were extracted in a review, the aspect it belonged to was considered extracted. The final chosen aspects and its keywords are as below and was brought from the paper written by [11] (2021) [11]. The aspect is written, followed by a colon and keywords for each aspect.

- Food: 'food', 'drink'
- Service: 'service', 'staff'
- Ambiance: 'ambiance', 'music', 'location'
- Price: 'price', 'money'

The above keywords were chosen because increasing the number of keywords per aspect did not significantly affect the result. Therefore a shorter list of keywords for each aspect was chosen. The long list of keywords for each aspect was between 5 to 20 keywords per aspect and the accuracy difference was less than 0.3%. The long list of keywords was manually obtained and categorized by looking at the most frequently used words in the dataset. The long list of keywords included is shown below:

- Food: various types of taste, drink, desert, dishes, and other food related words
- Service: staff, time, politeness, waiting system, convenience, and other service related words
- Ambiance: location, music, theme, quietness, atmosphere, overall review, restaurant review, seating, and other words not related to other the other 3 topics
- Price: money, value, worthiness, and other price related words

The weight was determined by the experiments conducted on methods 1, 2, 3, and 6. Every method had different weights to maximize the accuracy of the methods. The sum of all 4 weights for methods 1, 2, and 3 were all set to equal 1. Methods 1, 2, 3 tested all possible weight combinations using 1 decimal place, to get the highest accuracy while keeping the sum of weights as 1. Each aspect was given weights of 0.0 to 1.0. In total, more than 1000 weight combination was tested per method.

3.3 Method 1 - Combine Sentiment of Aspect in Each Sentence

The first method was divided into 4 steps as listed below.

1. Divide each review by sentence
2. Extract Aspect
3. Do aspect level sentiment analysis on each sentence
4. Combine the weighted scores

The first step was already completed in the preprocessing step. Step 2 used the average embedding of the sentence introduced in section 3.2.3 above. Step 3 used the method explained in section 3.2.4. Various details in the last step were tested to see which had the highest score. Method 1 was divided into 3 methods.

Method 1.1 was performed by using the aspect of the sentence and its simple polarity. If food had a negative sentiment, it was considered negative 1. If food had a positive sentiment, it was considered positive 1. If an aspect had 2 polarities, the polarity was averaged. Food had a weight of 0.4. The two numbers were multiplied and the value for that aspect was calculated. After that, the value of all four aspects was added up to get an overall sentiment value. If it was a negative number, the overall sentiment was negative. An example is shown in Figure 3.6 below. The original sentences and aspect similarity scores were from the paper by Alamoudi and Alghamdi [11].

Method 1.2 added the score element of the aspect. The value of each aspect was calculated by simply multiplying the polarity with weight in the previous method. The score was added

	Food	Service	Ambiance	Price	Aspect	Sentiment
I ordered the grilled chicken combo but received fried chicken.	**0.5631**	0.3956	0.3732	0.4056	Food	Negative
The inside is dirty in the afternoon like they never clean.	0.5098	0.4981	**0.5602**	0.4955	Ambiance	Negative
The people behind the counter are usually polite and fast.	0.5191	**0.5282**	0.5193	0.5035	Service	Positive
I have never had an order messed-up.	0.4992	**0.5852**	0.5232	0.5787	Service	Negative

Aspect	Polarity	Weight	Value
Food	Negative (-)	0.4	- 0.4
Service	(+ & -) = Neutral	0.3	0
Ambiance	Negative (-)	0.2	- 0.2
Price	None	0.1	0
		Overall	**- 0.6 (Negative)**

Figure 3.6. Method 1.1 Example

to this part from method 1.1. The score represented the semantic similarity of the sentence with aspect. The method to get the semantic similarity is shown in section 3.2.3. The score was multiplied with the sentiment polarity as well as the weight the final value of the aspect. The final step of adding the numbers to get the overall sentiment was the same as with Method 1.1. A diagram of the second example is shown below in Figure 3.7 below. The original sentences and aspect similarity scores were from the paper by Alamoudi and Alghamdi [11].

Method 1.3 was performed by adding a filter to the score element of the aspect. If the score of an aspect was below 0.5, that sentence was ignored as it was not close enough to any aspect. A diagram of the third example is shown in Figure 3.8 below. The original sentences and aspect similarity scores except for the number pointed with blue arrow were from the paper by Alamoudi and Alghamdi [11].

	Food	Service	Ambiance	Price	Aspect	Sentiment
I ordered the grilled chicken combo but received fried chicken.	**0.5631**	0.3956	0.3732	0.4056	Food	Negative
The inside is dirty in the afternoon like they never clean.	0.5098	0.4981	**0.5602**	0.4955	Ambiance	Negative
The people behind the counter are usually polite and fast.	0.5191	**0.5282**	0.5193	0.5035	Service	Positive
I have never had an order messed-up.	0.4992	**0.5852**	0.5232	0.5787	Service	Negative

Aspect	Polarity	Weight	Score	Value
Food	Negative (-)	0.4	0.5631	- 0.22524
Service	Negative (-)	0.3	0.0570 (0.5282 –0.5852)	- 0.0171
Ambiance	Negative (-)	0.2	0.5602	- 0.11204
Price	None	0.1	0	0
			Overall	**- 0.3544 (Negative)**

Figure 3.7. Method 1.2 Example

If below 0.5, Ignore Aspect

	Food	Service	Ambiance	Price	Aspect	Sentiment
I ordered the grilled chicken combo but received fried chicken.	**0.3631**	0.3956	0.3732	0.4056	[illegible]	[illegible]
The inside is dirty in the afternoon like they never clean.	0.5098	0.4981	**0.5602**	0.4955	Ambiance	Negative
The people behind the counter are usually polite and fast.	0.5191	**0.5282**	0.5193	0.5035	Service	Positive
I have never had an order messed-up.	0.4992	**0.5852**	0.5232	0.5787	Service	Negative

Aspect	Polarity	Weight	Score	Value
Food	Negative (-)	0.4	0	0
Service	Negative (-)	0.3	0.0570 (0.5282 –0.5852)	- 0.0171
Ambiance	Negative (-)	0.2	0.5602	- 0.11204
Price	None	0.1	0	0
			Overall	**- 0.129 (Negative)**

Figure 3.8. Method 1.3 Example

Similar to what is shown in the examples above, the details have been experimented to find the highest sentiment score. The disadvantage was that the relationship between sentences became weaker. For example, there were sentences next to each other saying "I really love chickens. However, the chicken here is terrible". Whether read by humans or aspect sentiment analysis done by computer, the two sentences result in a negative sentiment. However when aspect sentiment analysis was performed separately on the two sentences, the two sentences resulted in a positive and negative sentiment. The resulting sentiment of the two sentences were be a neutral sentiment or a very slightly positive sentiment. The example is illustrated in Figure 3.9 below.

Figure 3.9. Method 1 Disadvantage Example

3.4 Method 2 - Combine Sentiment of All 4 Aspects in a Single Review

Method 2 applies method 1, which combined the sentiment of aspects in individual sentences, to the whole review. The steps are listed below.

1. Extract all 4 aspects semantic similarity on the whole review
2. Perform aspect-based SA on the whole review using all 4 aspects
3. Combine the weighted scores (Details to be experimented)

All four aspect semantic similarity was extracted on the whole review using the method explained in section 3.3. The difference was that instead of getting the embedding vector

of the sentence, the embedding vector was from the whole review. Next, aspect sentiment analysis was performed on all four aspects. Finally, the weighted scores was combined. The assumption of this method was that even if some aspects were not explicitly in the review, the effect to the overall sentiment was minimal. The disadvantage was that if some aspects did not exist, aspect may have random sentiment. An example of the disadvantage is displayed in Figure 3.10 below.

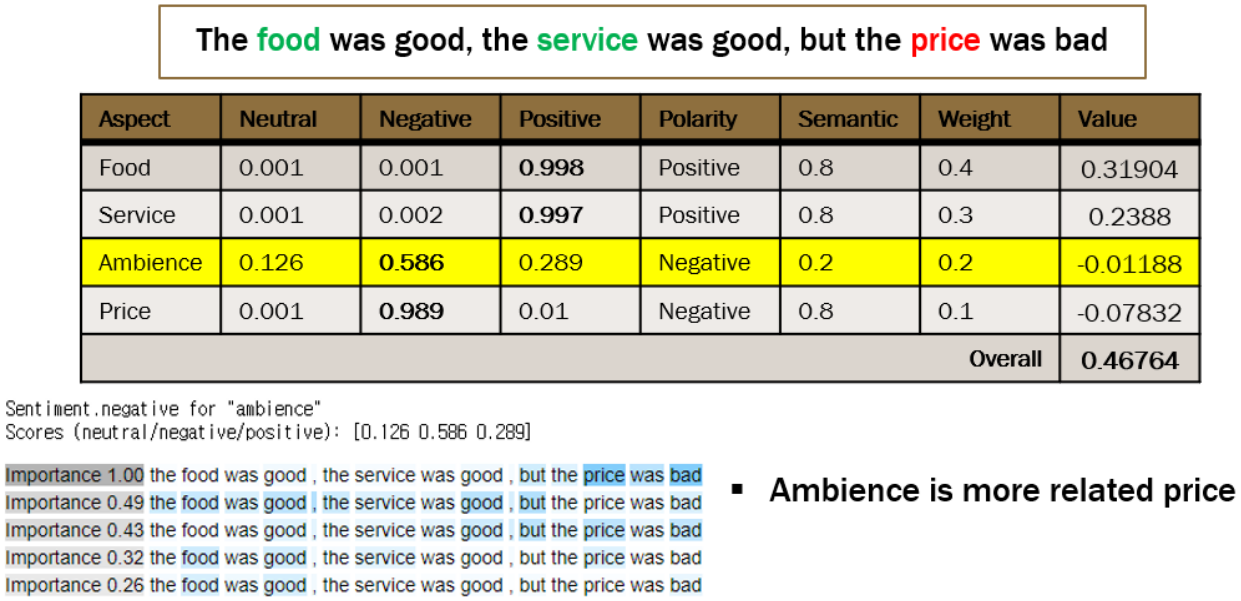

Figure 3.10. Method 2 Example

The aspect food had a positive sentiment, with a very high 0.8 semantic similarity, and a weight of 0.4. The numbers were multiplied to get a value of 0.3190. All aspects were calculated this way and added up for an overall sentiment of 0.4676 which was positive. The example above is a review that explicitly had 3 aspects. However, the aspect ambiance, highlighted in yellow, was not even included. The assumption was that the sentiment value of the aspect ambiance was low and the effect was minimal. Although the sentiment of ambiance was supposed to be neutral, it had a negative sentiment because it was more closely related to price than any other aspect. The effect of the aspect ambiance was minimal for this example because the made-up example had a low semantic value. However it might be effect the results in other reviews.

3.5 Method 3 - Combine Sentiment of Relevant Aspects in a Single Review

Method 3 followed typical aspect based sentiment analysis but gives weights to the aspects just before combining the scores. The steps are shown below.

1. Do aspect extraction on the whole review
2. Do aspect sentiment analysis
3. Give weights to the aspects
4. Combine the weighted scores

The first step extracts the aspects using manual aspect extraction explained in Section 3.2.3. After extracting the aspect, aspect sentiment analysis was performed as the second step. The third step gives weights to the 4 pre-defined aspects by multiplying the weight. Finally, the sentiment values were combined to produce an overall sentiment of the review. If a review had none of the aspects, that review was not included, but the number of reviews was very little. An example is shown below in Figure 3.11.

The food was good.
The service was good, but the price was bad

Aspect	Polarity	Weight	Value
Food	+ 0.99	0.15	+ 0.1485
Service	+ 0.92	0.35	+0.322
Price	- 0.88	0.15	- 0.132
		Overall	**+ 0.3385**
		Overall, After Weight Adjustment	**+ 0.5207 (4 Star)**

Figure 3.11. Method 3 Example

3.6 Method 4 - Combine Results of Supervised and Unsupervised Classification

Method 4 predicts the overall sentiment using both supervised and unsupervised classification. It combined the results of the unsupervised classification of methods 1, 2, and 3 with the results of the traditional supervised classification. There were 3 experiments that have been used to combine the sentiments.

Method 4.1 combined the results by simply averaging the sentiment score. For example, if the supervised classification had a result of positive 0.4 and the unsupervised classification had a result of negative 0.6, the overall sentiment was positive 0.2. An example is shown below in Figure 3.12.

Method	Sentiment	Confidence/Sentiment Score	Value
Supervised	Positive (+)	0.4	+ 0.4
Unsupervised (Method 1, 2, 3)	Negative (-)	0.6	- 0.6
		Overall	+0.2 (Positive)

Figure 3.12. Method 4.1 Example

Method 4.2 didn't combine the results. Instead, chose the sentiment that had a higher confidence score. An example is shown below in Figure 3.13.

Method	Sentiment	Confidence/Sentiment Score
Supervised	Positive (+)	0.4
Unsupervised (Method 1, 2, 3)	**Negative (-)**	**0.6** ← **Higher Score**

Figure 3.13. Method 4.2 Example

Method 4.3 considered the result of the supervised method as another aspect when doing the calculation. This was similar to method 4.1, but allowed weights to be changed so that

the weight of the supervised classification result did not have to be exactly 0.5. For example, the weight of the supervised method could be 0.2 while the weight of the food was 0.5. An example is shown below in Figure 3.14.

Aspect/Source	Sentiment	Weight	Value
Supervised	Positive (+)	0.2	+ 0.2
Food	Positive (+)	0.5	+ 0.5
Service	Positive (+)	0.15	+ 0.15
Ambience	Negative (-)	0.1	- 0.1
Price	Negative (-)	0.05	-0.05
		Overall	+ 0.7 (Positive)

Figure 3.14. Method 4.3 Example

3.7 Method 5 - Utilize Unsupervised Results in Supervised Classification

Method 5 utilized the result obtained from unsupervised classification and used it as a feature when doing supervised classification. There were 2 types for this method.

Method 5.1 included the text as well as the sentiment score of the overall unsupervised classification results of methods 1, 2, and 3 as the dataset. An example input data is shown in the below Table 3.1.

Table 3.1. Method 5.1 Example Input

Original Input Text	**Additional Input Unsupervised Result**
I loved the service and the food was good	0.7

Method 5.2 included the sentiment score of each aspect from the unsupervised classification results of methods 1, 2, and 3. An example of input data is shown in the below Table 3.2.

Table 3.2. Method 5.2 Example Input

Text	Food Sentiment	Service Sentiment	Ambiance Sentiment	Price Sentiment
I loved the food and service	+0.7	+0.8	None	None

Using the 2 different set of inputs shown above, a supervised classification was done. To see the accuracy difference, only Logistic Regression was used.

3.8 Method 6 - Emphasize Relevant Aspect in Supervised Classification

Method 6 simply emphasized words related to aspect in supervised classification. For this part, emphasizing simply meant repeating. First, the supervised learning model was trained using normal text of the Yelp dataset and the star rating. In a normal machine learning, the second part would have been using the testing dataset to see the results and accuracy of the training dataset. In method 6, the testing set emphasized the words related to the aspect. To be specific, the words that have high impact to each aspects sentiment was repeated in the text. An example is displayed in Figure 3.15 below.

Original Text - The sushi is great but the service is bad

Add weights [Aspect of Food with weight of 2] & [Aspect of Service with weight of 1]

Sentiment Enhanced Text - The sushi is great great, but the service is bad

Training with original text, testing with modified text

Figure 3.15. Method 6 Example Sentiment Enhancement

The steps of emphasizing the word is explained below.

1. Using the aspect-based-sentiment-analysis library, the words that had a high impact on the sentiment of each aspect were found. For example, in the text "Sushi is great but the service is bad", the word that had a high impact on the sentiment of the aspect food was 'great', colored in green in the above Figure 3.15. The word that had a high

impact on the sentiment of the aspect service was 'bad', colored in red, as shown in the above Figure 3.15.

2. The word that had a high influence on the aspect's sentiment was found. A minimum influence number was set and any words that had an influencing number above the set number was used to emphasize. The influence number was also retrieved from the aspect-based-sentiment-analysis library. On a scale of 0 to 1.0, the different influencing number was tested and explained in the results section. The final chosen minimum influencing number was 1.0. An example is displayed in the below Figure 3.16. The word token 'food' and 'great' had an influencing number of 1.0 for the aspect food. Since the word token 'food' was the actual word of the food aspect, it was not repeated. Only the word token 'great' was repeated based on the weight of the food aspect.

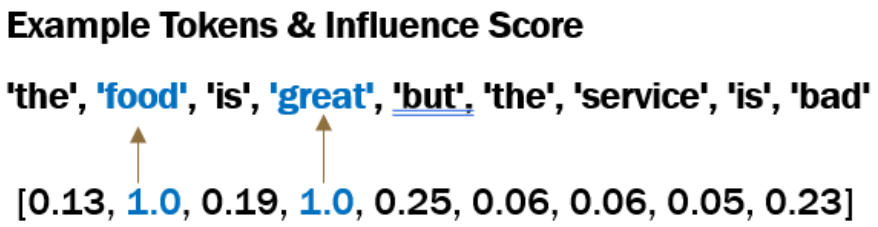

Figure 3.16. Method 6 Example Sentiment Influence by Token

3. The text was modified to repeat the word based on the set weight of the aspect. Since the food had a weight of 2 while the service had a weight of 1, the word great was repeated 2 times, while the word bad was kept at 1.

4. Only the modified text was used in the testing data and the supervised classification was completed.

Using the steps mentioned above, different weights and different minimum influence numbers were tested.

4. RESULTS AND DISCUSSION

This chapter presents the results of the six methods.

4.1 Sentiment Range

The result of unsupervised methods 1, 2, and 3 was a number from -1 to +1. To classify a star rating from this number, a set of ranges was established for each star rating. Sentiment 1 was equivalent to a star rating of 1 which was very negative. Sentiment 2 was equivalent to a star rating of 2 which was moderately negative. Sentiment 3 was equivalent to a star rating to 3 which was neutral. Sentiment 4 was equivalent to a star rating of 4 which was moderately positive. Sentiment 5 was equivalent to a star rating of 5 which was very positive. The exact range of each sentiment is listed below. The below sentiment range will be expressed as [-0.50, -0.30, 0.10, 0.55] for this paper.

- Sentiment 1: [-1.00, -0.50]
- Sentiment 2: [-0.50, -0.30]
- Sentiment 3: [-0.30, +0.10]
- Sentiment 4: [+0.10, +0.55]
- Sentiment 5: [+0.55, +1.00]

The sentiment range was decided after doing experiments on different methods. The 3 methods had slightly different sentiment ranges to maximize the accuracy. However, the accuracy difference compared with the highest accuracy was not significant from the sentiment ranges shown above. Since the sentiment ranges above fit all 3 methods well, the range shown above was used as the baseline for all 3 methods.

Some results are shown in 3 sentiments instead of 5 sentiments. The 3 sentiment was a result of combining the 4 and 5 stars as positive sentiment and 1 and 2 star as negative sentiment. The neutral star 3 was kept as is. The sentiment range of the star 1 and 2 as well as star 4 and 5 was combined as a single range each. Depending on how the weight of each

aspects were changed, the accuracy went up by 1% or down by 15%. The extreme case of going down by 15% was when only 1 aspect was used to predict the sentiment of the review.

The sentiment ranges when using star 1 and star 5 only will also be explained in each section. For the ranges of using the 2 stars, there was only 1 middle point as there are only 2 ranges of star 1 and star 5 only.

4.2 Method 1: Results

Using the baseline sentiment range explained in the previous section, methods 1.1, 1.2, and 1.3 were experimented. The results of method 1.1 and 1.2 with equal weights on the aspects are shown in the below Table 4.1 and Table 4.2. Table 4.1 displays the accuracy using 3 sentiment. Table 4.2 displays the accuracy using 5 sentiments as well as the average sentiment of the each star and overall sentiment. As explained in section 3.1, the dataset is not balanced so the average sentiment will be the average of all the reviews combined, not the average of 5 sentiments. Table 4.3 shows the result when only reviews of star 1 and star 5 were used to do the sentiment analysis. The middle sentiment range of method 1.1 1 and 5 star only was -0.00 and the middle sentiment range of method 1.2 1 and 5 star only was -0.05.

Table 4.1. Method 1 - 3 Star Results

Star	**Method 1.1 Accuracy**	**Method 1.2 Accuracy**
1 & 2 Negative	0.5689	0.7463
3 Neutral	0.4290	0.2747
4 & 5 Positive	0.8805	0.9108
Overall	0.7578	0.7913

The accuracy was about 75% and 79% in the 3 sentiment system and about 48% and 51% in the 5 sentiment system. The results were low compared to traditional sentiment analysis result of around 84% using the 3 sentiment system and 60% using the 5 sentiment system. Both method 1.1 and 1.2 showed similar pattern. The accuracy of the positive sentiment tend to be higher than those of the negative sentiment. Neutral sentiment was the lowest in the 3 sentiment system, but exactly the middle at 5 sentiment system. Method 1.2 showed

Table 4.2. Method 1 - 5 Star Results

Star	Method 1.1 Accuracy	1.1 Average Sentiment	Method 1.2 Accuracy	1. 2 Average Sentiment
1	0.3927	-0.6597	0.7280	-0.4249
2	0.2348	-0.4336	0.1581	-0.2840
3	0.4290	0.0231	0.2747	0.0129
4	0.4735	0.5330	0.3012	0.3903
5	0.5860	0.7377	0.7881	0.5809
Overall	0.4714	0.3301	0.5110	0.2655

Table 4.3. Method 1 - 1 and 5 Star Only Result

Star	Method 1.1 Accuracy	Method 1.2 Accuracy
1	0.9133	0.8938
5	0.9769	0.9781
Overall	0.9637	0.9606

more drastic change in accuracy by star compared to method 1.1. When the 2 methods used 1 and 5 stars only, it showed high accuracy. It was not as high as the traditional supervised classification accuracy of 97.41% but only a 1% difference was considered high enough.

As method 1 showed the highest accuracy, the regression score was tested as well. The result is shown below in Table 4.5. The exact label for the star was experimented to find the label with the highest score. The label that showed the highest r2 score is show in Table 4.4. The 'Equal' column was included as well to compare with method 5 in the later section.

Table 4.4. Method 1 - Regression Sentiment Label

Star	Method 1.1 Optimized Label	Method 1.2 Optimized Label	Equal Label
1	-1.0	-1.0	-1.0
2	-0.8	-0.9	-0.5
3	-0.2	-0.2	0.0
4	0.5	0.7	0.5
5	0.9	1.0	1.0

Table 4.5. Method 1 - Regression Results

Sentiment Label	Weights	Method 1.1 r2 score	Method 1.2 r2 score
Optimized	Weighted	0.5353	0.6066
Optimized	Equal Weight	0.5335	0.6001
Equal	Weighted	0.5056	0.5660
Equal	Equal Weight	0.4991	0.5568

Table 4.5 showed that method 1.2 showed higher r2 score compared to method 1.1. The highest score when optimized sentiment label and applied weights showed a 0.6066 score. It did show better result compared to simple classification, but the score overall was not high compared to method 5, which will be displayed section 4.7.

When changing the weights of each aspects, the accuracy did change drastically, but it did not increase significantly. Different combinations of weights were tested. Some of the results are shown in the below Table 4.6. The weights are displayed in the order of 'Food', 'Service', 'Ambiance', and 'Price'.

Table 4.6. Method 1.2 Results by Weights

Weight	Accuracy	R2
[0.25, 0.25, 0.25, 0.25]	0.5110	0.6001
[0.4, 0.3, 0.2, 0.1]	0.5027	0.5922
[0.4, 0.2, 0.2, 0.2]	0.5084	0.6028
[0.2, 0.2, 0.4, 0.2]	0.5090	0.5975
[0.2, 0.4, 0.2, 0.2]	0.5037	0.5902
[0.3, 0.2, 0.3, 0.2]	0.5122	0.6066

The results showed that there is a difference in accuracy and r2 score based on the weights. In an extreme case where only 1 aspect had all the weights, the accuracy went down by 15%. However, when given reasonable weights, the accuracy did not change drastically, mostly within 1% difference. Similar results were shown for r2 scores. The highest accuracy was shown where the food aspect and ambiance aspect both have a weight of 0.3 each and service aspect and price aspect both have a weight of 0.2 each. The weight was retrieved after trying all combinations of weights explained in Section 3.2.5.

More experiment results are shown in the below graph, Figure 4.1, to see the trend of accuracy when weights changes. The graph below shows the accuracy when an aspect had a certain percent of weight while all other weights had equal amount remaining of the 100% weight. For example, the first point of the blue line shows the accuracy when the aspect food was at 0.0 weight and the remaining three aspects had 0.333333 each. The second point of the blue line shows the accuracy when aspect had 0.1 weight while the remaining three aspects have 0.3 weight each. The third point of the blue line shows the accuracy when aspect food had 0.2 weight while the remaining three aspects had 0.266666 weight. The last point of the blue line shows the accuracy when the food had the whole 1.0 weight while the remaining three aspects had 0.0 weight each. The same method was performed with method 2 and 3, but the result of each method will be shown in the later sections.

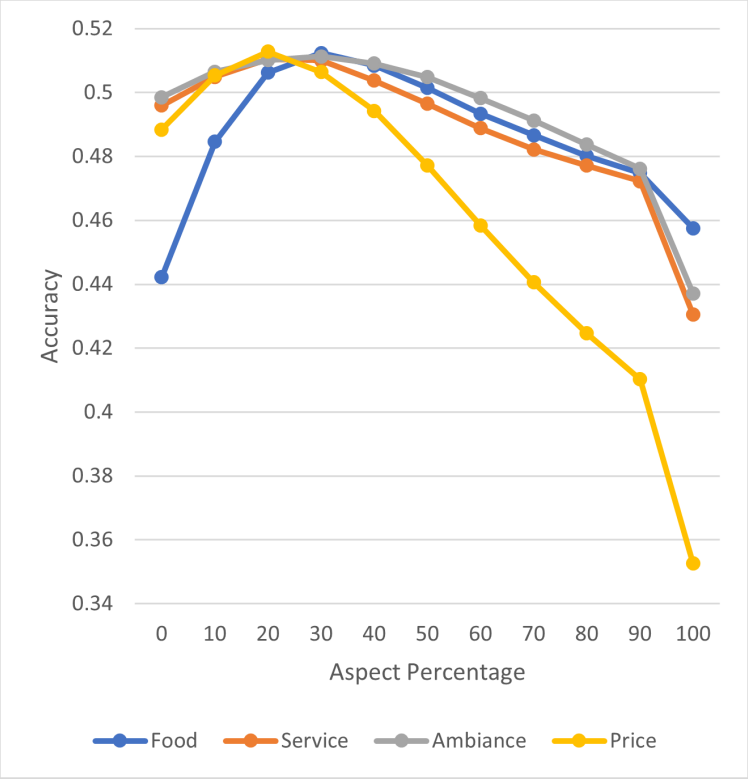

Figure 4.1. Method 1.2 Results by all weights

The sudden increase of accuracy from the first point to the next and sudden decrease in accuracy from the second last point to the last point was because of the number of reviews

included. For example, the last point in the yellow line has a price weight of 1.0. This means that if there was a review with 3 sentences that all talk about food, service, and ambiance aspect, but not the price aspect, that review was not be used. For that specific example, of the 100,000 reviews, 20,381 did not have the price aspect. Therefore, only 0.7961 percent of the reviews were used to measure the accuracy.

Further analysis was done to find the effect of number of aspects and percentage. The Table 4.7 below shows the number of aspects that was found in a review with the accuracy and the percent of review that was included from the total review. The result showed that having all 4 aspects in a single review had the highest accuracy, but the difference was small with when there were 3 aspects in the review.

Table 4.7. Method 1.2 Results by Number of Aspects

Number of Aspects	Accuracy	Percentage of Review
1	0.4904	0.0815
2	0.5018	0.3196
3	0.5202	0.3915
4	0.5217	0.2072

Another feature that was experimented was the similarity score, specifically method 1.3. When changing the minimum similarity of the word to the sentence, the accuracy did change. The result is displayed below in Table 4.8. The "Minimum Similarity" column was the minimum semantic cosine similarity score of an aspect to each sentence. If the similarity did not meet the minimum similarity, the aspect was not included. The result of method 1.3 was based using method 1.2 as method 1.2.

Looking by just the minimum similarity score, there was a definite negative correlation between the minimum similarity and the accuracy. The accuracy was highest when there were no minimum similarity, when the minimum similarity was 0. The accuracy became lower as the minimum similarity increased. The accuracy was the lowest when the minimum similarity was the highest. The reason for the decrease in accuracy was assumed to be because of the decrease in the included review and aspect. As shown in previous results, less number of aspects tend to show lower accuracy. The same logic applied method 1.3. Since

Table 4.8. Method 1.3 Results by Minimum Similarity

Minimum Similarity	Included Data	3 Sentiment Accuracy	5 Sentiment Accuracy
0.7	0.3107	0.7066	0.4158
0.6	0.6140	0.7089	0.4204
0.5	0.8658	0.7185	0.4389
0.4	0.9858	0.7612	0.4833
0.3	0.9980	0.7862	0.5051
0.2	0.9991	0.7923	0.5104
0.1	0.9995	0.7941	0.5122
0.0	1.0000	0.7942	0.5122

there are less number of aspects that was used with higher minimum similarity, the accuracy decreased.

The optimized sentiment range was tested for both methods. The optimized sentiment range for method 1.1 was [-0.3, -0.2, 0.0, 0.2] and the accuracy with weight was 0.4961. The optimized sentiment for method 1.2 was [-0.55, -0.35, 0.05, 0.60] and the accuracy with weight was 0.5135. The sentiment range was expressed in the same way as Section 4.1. As you can see, the sentiment range was more wider in method 1.2. The accuracy still remained higher for method 1.2.

4.3 Method 2: Results

The results of method 2 were lower than method 1. Method 2 also showed the highest accuracy when given different weights compared to method 1. The result is shown below in Table 4.9 and Table 4.10. The 'Equal Weights' column simply means when all 4 aspects received 0.25 weight. Table 4.11 shows the result when only reviews of star 1 and star 5 were used to do the sentiment analysis. The middle sentiment range of 1 and 5 stars only result was -0.3.

The highest accuracy was obtained when the weight was 0.3 for the aspect food, 0.4 for the aspect service, 0.3 for the aspect ambiance, and 0.0 for the aspect price. The second highest accuracy was obtained when the weight of food aspect decreased by 0.1 and price

Table 4.9. Method 2 - 5 Star Results

Star	Equal Weights	Given Weights [0.2, 0.4, 0.3, 0.1]	Given Weights [0.3, 0.4, 0.3, 0.0]	Average Sentiment
1	0.7209	0.7556	0.6836	-0.6420
2	0.1976	0.1836	0.1899	-0.5154
3	0.3367	0.3022	0.3297	-0.1838
4	0.3160	0.2893	0.2916	0.2627
5	0.5925	0.6451	0.6554	0.5471
Overall	0.4531	0.4628	0.4647	0.1485

Table 4.10. Method 2 - 3 Star Results

Star	Equal Weights	Given Weights [0.2, 0.4, 0.3, 0.1]	Given Weights [0.3, 0.4, 0.3, 0.0]
1 & 2 Negative	0.8188	0.8267	0.7715
3 Neutral	0.3367	0.3022	0.3297
4 & 5 Positive	0.7615	0.7779	0.7972
Overall	0.7143	0.7221	0.7282

Table 4.11. Method 2 - 1 and 5 Star Only Result

Star	Accuracy
1	0.8784
5	0.9583
Overall	0.9420

increased by 0.1. Although giving weights only increased the accuracy by about 1%, the weighted accuracy of 5 star system was 5% lower compared to method 1.2 5 star system. Even with experiments with different weights, the highest accuracy of method 2 was lower than of method 1. The result of more experiments with different weights are displayed in the below Figure 4.2. The same logic was used as with the method to complete Figure 4.1.

The optimized sentiment range was tested for method 2. The sentiment range was [-0.6, -0.5, -0.3, 0.25] and the accuracy with weights was 0.4909. The accuracy was still lower than method 1.

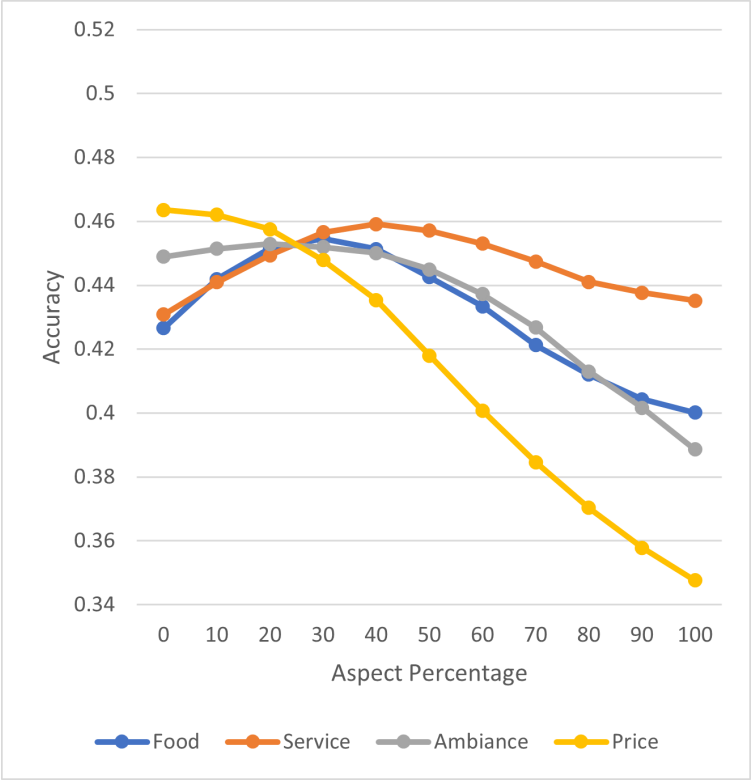

Figure 4.2. Method 2 Result by weights

4.4 Method 3: Results

The result of method 3 was lower than method 1 but slightly better than method 2. Method 3 also showed the highest accuracy when assigned different weights compared to method 1 and method 2. The result of method 3 by star rating is shown below in Table 4.12 and Table 4.13. Table 4.14 shows the result when only reviews of star 1 and star 5 were used to do the sentiment analysis. The middle sentiment range of 1 and 5 stars only result was -0.25.

The overall accuracy of method 3 was 47.70% when equal weights were used. When given weights, the accuracy increase by up to 1% and down by 8%. The results showed lower accuracy than method 1.2 but higher accuracy than method 2. The accuracy when a certain number of aspects were in the review showed that unlike method 1, method 3 performed the highest when there were 2 aspects in the review, followed by 3 aspects in the review.

Table 4.12. Method 3 - 5 Star Results

Star	Equal Weights	Given Weights [0.2, 0.2, 0.4, 0.2]	Average Sentiment
1	0.7212	0.7260	-0.6449
2	0.1607	0.17610	-0.4504
3	0.3092	0.3101	-0.0525
4	0.2573	0.2662	0.4135
5	0.7181	0.7121	0.6767
Overall	0.4765	0.4788	0.2663

Table 4.13. Method 3 - 3 Star Results

Star Rating	Equal Weights	Given Weights [0.2, 0.2, 0.4, 0.2]
1 & 2 Negative	0.7614	0.7680
3 Neutral	0.3092	0.3101
4 & 5 Positive	0.8189	0.8209
Overall	0.7382	0.7409

Table 4.14. Method 3 - 1 and 5 Star Only Result

Star	Accuracy
1	0.8938
5	0.9578
Overall	0.9447

Looking at the accuracy result when a certain number of aspects were detected, there seems to be no a specific pattern with method 1.

More experiments were tested with the different weights as done with methods 1 and 2. The below Figure 4.3 shows the result. The same logic was used in Figure 4.1.

The accuracy by number of aspect present in the review was tested. Number of aspect present was defined by how many different aspect's keyword was detected in the review. For example, if a review had the words 'food', 'service', and 'price' present but words related to ambiance were not present, it was considered as 3 aspects as present. Table 4.15 shows the accuracy when a certain number of aspects were present in review as well as the percentage of the review.

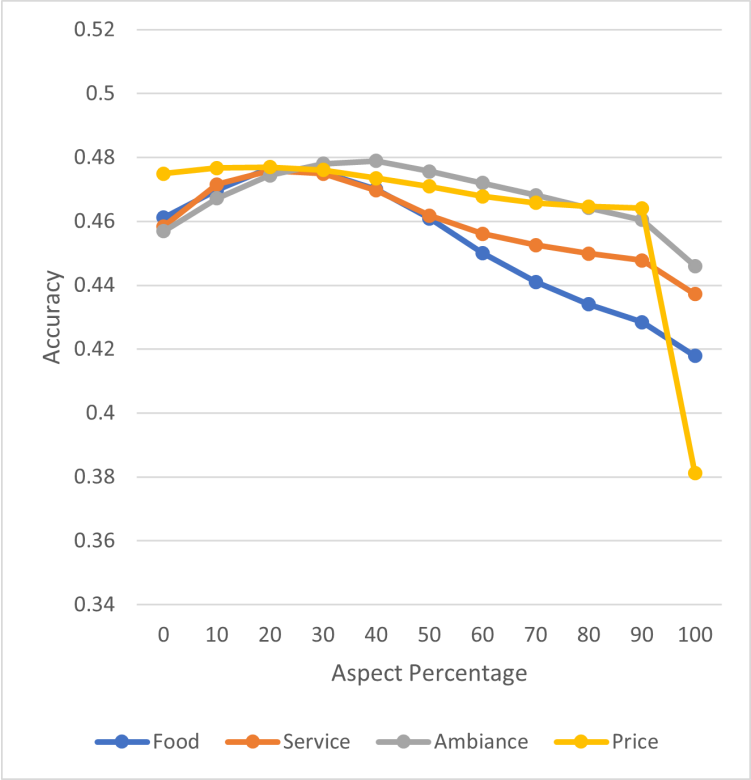

Figure 4.3. Method 3 Result by Weights

Table 4.15. Method 3 - Results by Number of Aspects

Number of Aspects	Accuracy	Percentage of Review
1	0.4736	0.1338
2	0.4871	0.3304
3	0.4781	0.4281
4	0.4408	0.1074

Method 2 assumed that all aspects are present so it was not a part of the comparison. Compared to method 1, in all number of aspects, the accuracy was low. Aspect count of 2 showed the highest accuracy and aspect count of 4 had by far the lowest accuracy. Method 1 had a pattern of increase of accuracy when the number of aspects increased, which was a different pattern with method 3. Reviews with an aspect count of 4 were 0.1074 percentage of the dataset. In the dataset of 100,000 reviews, it was still over 10,000 reviews which was still a large enough. Based on the result, it was possible to understand that the increase in

number of aspects can decrease the accuracy because the keyword detection method was not an accurate way and therefore can make the sentiment prediction wrong.

The optimized sentiment range was tested for method 3. The sentiment range was [-0.6, -0.45, -0.3, 0.5] and the accuracy with weights was 0.4890. The accuracy was still lower than method 1.2 but very similar to method 2.

4.5 Unsupervised Method Result

A comparison of the three unsupervised methods were made. These were some of the interpretations that could be made based on the results. The difference between the 3 methods are shown in Table 4.16. The results of the individual aspects were obtained by using only 1 aspect to do prediction. The last row of 'Overall' was the highest accuracy achieved, as shown in the previous results. The result of method 1 was retrieved using method 1.2.

Table 4.16. Unsupervised Method Result Comparison by Aspect

Aspects	**Method1 Accuracy**	**Method2 Accuracy**	**Method3 Accuracy**
Food	0.4027	0.4001	0.4178
Service	0.4238	0.4351	0.4372
Ambiance	0.4282	0.3886	0.4460
Price	0.3336	0.3476	0.3811
Overall	0.5110	0.4628	0.4788

Among the three methods, method 1 showed the highest accuracy and method 2 had the lowest accuracy. Method 1 ignored the relationships between the sentences since it cuts all the sentences and considers each sentence as a single aspect while methods 2 and 3 do aspect based sentiment analysis on the whole review. This accuracy difference and sentence split difference could be interpreted as the relationship between the sentences does not have an impact on the accuracy of the whole review. Another interpretation was that people who write the reviews simply do not write enough reviews where the relationship between the sentences would matter enough to affect the result. Based on the result, the relationship between sentences may not need to be considered when doing sentiment analysis of reviews.

Method 2 was the only method that assumed to contain all 4 aspects in the review. It was also the method with the lowest accuracy. This could be interpreted as that many of the reviews do not talk about all 4 aspects in the review. Therefore, doing classification with the assumption that all 4 aspects exist may have lowered the accuracy. The interpretation can be supported by the fact that although method 1 and method 2 use the same method for aspect extraction, they showed a significant decrease in accuracy. Another support is shown in Table 4.15 of method 3 and in Table 4.7 of method 1. Only 10% to 20% of the reviews had all 4 aspects and around 10% only had 1 aspect. This meant that more reviews do not mention all 4 aspects in the reviews and was reflected in the accuracy.

Looking at the result in Table 4.16, method 3 showed the highest accuracy when using only 1 aspect. This could be interpreted as that keyword based aspect extraction is better than overall cosine similarity aspect extraction when only 1 aspect is used. However, when finding the overall sentiment, method 1 gave higher accuracy.

Based on the graphs shown below in Figure 4.4, accuracy heavily depended on the existence of aspects. The left graph is the result from method 1.2 shown in Figure 4.1. The middle graph is the result from method 2 shown in Figure 4.2. The right graph is the result from method 3 shown in Figure 4.3.

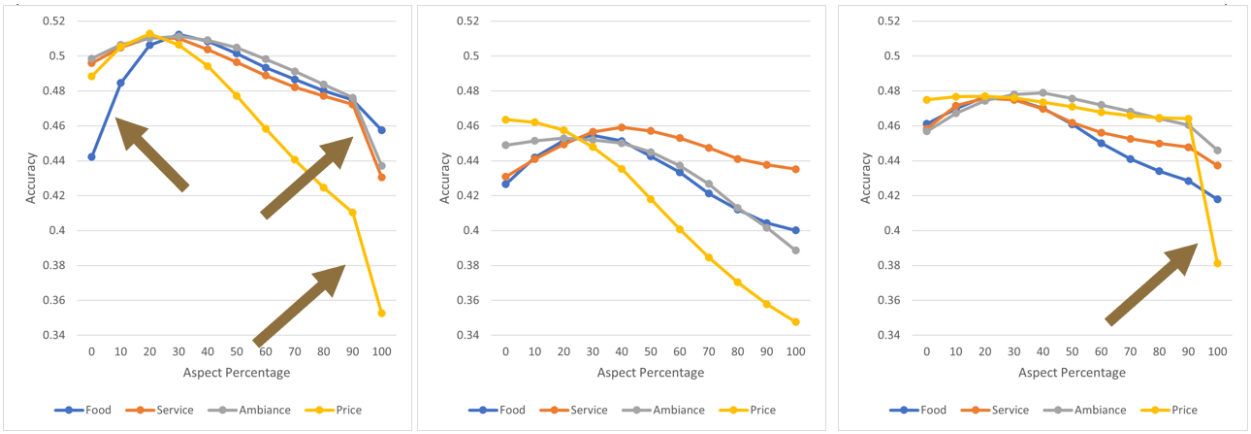

Figure 4.4. Unsupervised Method Result Comparison by Aspect

Based on the 3 graphs, there is a sudden decrease and sudden increase in accuracy based on the existence of certain aspects. The most left brown arrow shows the accuracy when only the food is not included to predict the star rating. The moment the food aspect is

included in the review, the accuracy suddenly increases. The remaining arrows are sudden drops in accuracy because only 1 aspect is used for classification. The graph in the center is accuracy of method 2. It does not show a significant change in accuracy because method 2 assumed all 4 aspects are in all reviews. This can be interpreted as that the some of the aspects effects significantly to the accuracy of the review. Also, looking at the 3 graphs and Table 4.16, price generally had low accuracy when price was the only aspect that was used for classification. This could be because price had relatively low importance compared to other aspects. The only common pattern between the 3 graphs were that the accuracy was normally the highest and most similar when the aspect percentage was similar in the 20 to 30 percent range.

4.6 Method 4: Results

Several methods were attempted using method 4 to find some pattern to increase the accuracy using both the result of supervised method and result of unsupervised methods. However, almost all methods showed lower accuracy than supervised methods. In almost all aspects, stars, confidence score, and accuracy by aspect, supervised methods showed higher accuracy. The result is shown below in 2 tables. Table 4.17 shows the accuracy by the confidence score of the supervised learning method using logistics regression algorithm. Table 4.18 shows the accuracy by star of the supervised method as well as the methods explained in the previous sections.

Table 4.17. Method 4 Results of Supervised Learning by Confidence Score

Minimum Confidence Score	Included Data	Accuracy
0.0	100%	60.83%
0.2	100%	60.83%
0.4	90%	63.18%
0.5	70%	69.25%
0.6	50%	75.11%
0.7	30%	82.11%
0.8	<20%	88.44%
0.9	<10%	93.05%

Table 4.18. Method 4 Result Comparison

Star	Supervised	Method1	Method2	Method3
1	72.01%	72.80%	59.25%	FFFE6572.60%
2	37.11%	15.81%	31.60%	17.61%
3	36.49%	27.47%	33.67%	31.01%
4	53.49%	30.12%	19.76%	26.62%
5	78.03%	78.81%	72.09%	71.21%
Overall	60.83%	51.10%	46.28%	47.88%

When compared to accuracy of methods 1, 2, and 3 by each star, supervised methods showed higher accuracy in all areas. When attempted to find any data with any confidence score range where the unsupervised methods had a higher accuracy, there were very limited data. Unsupervised method showed higher accuracy 3 times when the sentiment star was 1 and 5. Method 1 did show slightly higher accuracy by less than 1%. Even when looking into those data, there was no obvious pattern that can be utilized in a large enough dataset. For example when the star rating was 1, aspect was price, and supervised confidence score was below 0.5, the unsupervised method showed 0.04 percent higher accuracy. However the percentage of the such data in the whole review were very limited to make an meaningful increase to the accuracy. The example explained above only increased the accuracy by 0.0002. The result of method 4 was concluded that there was no practical pattern in which the unsupervised method can be used with the result of supervised methods.

4.7 Method 5: Results

Method 5 did show an increase in accuracy compared to traditional methods. However, it did not see significant change. The baseline traditional supervised learning showed an accuracy of 0.6044 and r2 score of 0.5934. The result using classification algorithm of method 5 is shown below in Table 4.19. The result of method 5 using only reviews with 1 Star and 5 Star is shown below in Table 4.20. The result using regression algorithm of method5 is shown below in Table 4.21.

Method 5 did show a slight increase in accuracy. The reason is assumed to be because there are some limited situation in which unsupervised learning is better like exampled in

Table 4.19. Method 5 Classification Results

Unsupervised Method	Method 5.1 Equal Weight	Method 5.1 Weighted	Method 5.2 No Weight
Method 1 Accuracy	0.6216	0.6217	0.6181
Method 1 r2	0.6605	0.6595	0.6545
Method 2 Accuracy	0.6182	0.6160	0.6201
Method 2 r2	0.6587	0.6550	0.6639
Method 3 Accuracy	0.6150	0.6137	0.6073
Method 3 r2	0.6447	0.6453	0.6035

Table 4.20. Method 5 - 1 and 5 Star Only Classification Results

Unsupervised Method	Method 5.1 Equal Weight	Method 5.1 Weighted	Method 5.2 No Weight
Method 1	0.9835	0.9825	0.9814
Method 2	0.9794	0.9788	0.9795
Method 3	0.9792	0.9796	0.9735

Table 4.21. Method 5 Regression Results

Unsupervised Method	Traditional Result	Method 5.1 Equal Weight	Method 5.1 Weighted	Method 5.2 No Weight
Random Forest Regression	0.5464	0.6919	0.6933	0.6717
Linear Regression	0.5455	0.6214	0.6229	0.5858

method 4. The interpretation was that method 5 was able find those better specific situations that unsupervised methods have higher accuracy. However, those situations were limited. Therefore the accuracy increase was not significant. The interesting thing was that the sentiment analysis using regression algorithm showed significant increase. To be specific, adding the result of unsupervised method increased r2 score by about 0.15, but just the significance of adding weights was low. When doing classification using the 1 and 5 stars only, the accuracy was high. Compared to a traditional sentiment analysis with an accuracy of 97.41%, method 5.1 with equal weights had an increase of 0.94%. At such a high accuracy, almost 1% increase is a significant increase. However, since this paper was focused on adding

weights to aspects, the result of method 5 can also be concluded that there was no significant increase in the r2 score when adding in weights.

Another experiment used supervised learning with only the results of method 1.2. Instead of having text and results of method 1 as input, result of method 1.2 was the only input. The result showed an accuracy of 50.86% when using the result of method 1.2. When using the sentiment of each aspect as input, the result showed 49.39% accuracy. The minor difference in accuracy of the 2 results could be interpreted as that manually giving weights to aspect could be slightly more accurate, but the difference was not big enough to be considered significant. The accuracy of 50.86% was very similar to the accuracy of method 1.2.

4.8 Method 6: Results

The result of method 6 showed similar results with other methods. The method used the Logistic Regression algorithm. With equal weights, which means the testing dataset was not changed, the accuracy was about 60.05%. Various weights were tested to see the effect on the classification accuracy as well as the r2 score. The result showed that there was a slight increase in the accuracy for some weights as well as slight increase in r2 score. However, the accuracy and r2 score increase was less than 0.5%, which is not a significant increase. The highest classification accuracy was when the aspect 'food' was emphasized 3 times. The highest r2 score depended on the algorithm that was used. The result of the logistic regression classification with various weights are shown in the below Table 4.22. The result of the regression algorithms is shown in Table 4.23. The number in the 'weights' column are the weights of each aspect in the order of food, service, ambiance, and price. The last column is the result when the training and testing used 1 star and 5 star reviews.

The weights of the above figure are the weights of ['food', 'service', 'ambiance', 'price'], in that order. For example, weights of [3, 2, 2, 2] means that the words that influenced the word 'food' was repeated 3 times in total. Words that influenced the word 'service' was repeated 2 times in total. Words that influenced the word 'ambiance' was repeated 2 times in total. Words that influenced the word 'price' was repeated 2 times in total.

Table 4.22. Method 6 Logistic Regression Results by Aspect Weights

Weights	Classification	R2 Score	Star 1 & 5 Only Classification
[1, 1, 1, 1]	0.6005	0.6057	0.9741
[3, 2, 2, 2]	0.6011	0.6008	0.9762
[2, 3, 3 ,2]	0.6014	0.6026	0.9760
[2, 1, 1, 1]	0.6031	0.6073	0.9748
[2, 2, 2, 2]	0.6003	0.6021	0.9760
[3, 1, 1, 1]	0.6042	0.6058	0.9747
[1, 3, 1, 1]	0.6002	0.6021	0.9754
[1, 1, 3, 1]	0.6005	0.6028	0.9736
[1, 1, 1, 3]	0.6003	0.6044	0.9742
[0, 0, 0, 0]	0.5818	0.5451	0.9647
[1, 0, 0, 0]	0.5951	0.5834	0.9711
[0, 1, 0, 0]	0.5866	0.5639	0.9688
[0, 0, 1, 0]	0.5838	0.5491	0.9654
[0, 0, 0, 1]	0.5835	0.5505	0.9654
[1, 0, 1, 0]	0.5959	0.5859	0.9712
[1, 1, 0, 0]	0.5990	0.5992	0.9739

Table 4.23. Method 6 Regression Results by Aspect Weights

Weights	Support Vector Regression	Random Forest Regression	Linear Regression
[1, 1, 1, 1]	0.6751	0.5631	0.5702
[3, 1, 1, 1]	0.6776	0.5599	0.5735
[1, 2, 2, 1]	0.6745	0.5627	0.5694
[1, 2, 1, 1]	0.6750	0.5633	0.5697

From the original 0.6005, the accuracy went down by only 0.0019 and up by 0.0009. The R2 score had a bigger increase of 0.0016 but is also not significant. The result of the regression algorithms showed similar result. By changing the weight, the r2 score of regression algorithm went up by at most 0.0033, which also is not a significant change.

Even when aspect weights were [0, 0, 0, 0], meaning all words that influence the aspect sentiment were deleted, the accuracy did not go down drastically. Assuming our aspect extraction method was accurate and aspect-based sentiment analysis was performed accurately,

it can be assumed that there are many other words, that is not related to the 4 aspects, that can help predict the overall sentiment of the review.

An additional experiment was performed to see if the minimum influence score would affect the accuracy. The method and explanation of the minimum influence score are in Section 3.8. The results are shown in the below Table 4.24.

Table 4.24. Method 6 Results by Minimum Influence

Minimum Influence	Accuracy	R2 Score
1.0	0.6014	0.6026
0.9	0.6012	0.6021
0.8	0.6011	0.6025
0.7	0.6010	0.6018
0.6	0.6010	0.6025
0.5	0.6015	0.6022

When the minimum influence score was changed, there was some change in the accuracy as well as the R2 score, but the change was even more small. The small influence score was the amount of influence a word token has to a specific aspect, explained in Section 3.8. The difference in accuracy was only 0.05 at most and the difference in r2 score was only 0.008. There were no distinctive pattern shown in any minimum influence score.

The small difference was not only shown in logistic regression algorithm, but also support vector machine as well. The accuracy difference for support vector machine was by only 0.001 and the r2 score was by 0.004. Both numbers are not a significant change. The result is shown in the below Table 4.25

Table 4.25. Method 6 Classification Results

Weights	Logistic Regression Accuracy	Logistic Regression R2 Score	SVM Accuracy	SVM R2 Score
[1, 1, 1, 1]	0.6057	0.6005	0.5894	0.5978
[3, 2, 2, 2]	0.6011	0.6008	0.5888	0.5969
[2, 3, 3, 2]	0.6014	0.6024	0.5883	0.5936

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

People may think that some aspects of a restaurant, such as food, is more important than other aspects, such as the weather that day. Computers, unless specifically told to do so, do not consider a certain aspect to be more important than other aspects. This paper experimented with giving more weights to specific aspects of a restaurant review to check the effect in accuracy when doing sentiment analysis. The result showed that adding weights to aspects does increase the accuracy of sentiment analysis. However, the increase in accuracy was not significant enough to conclude that adding more weights to certain aspects makes an impact to the overall classification.

Methods 1, 2, and 3 were unsupervised methods proposed in attempt to classify the overall sentiment of a restaurant review just by using the sentiment of each aspects. The result was that the unsupervised method had significantly lower accuracy than supervised methods. Applying weights to the aspects only increased the accuracy by at most 1%. While doing unsupervised learning for restaurant reviews, the relationship between sentences did not have to be considered. There was no specific aspect that had a significantly high impact on the accuracy compared to other aspects. Each method required different weights to maximize the accuracy. There were no obvious patterns on the weights of the aspect. The only common factor was that the price should have a relatively low weight when doing sentiment analysis on the review. Although all three methods had lower accuracy than traditional supervised classification, the accuracy was similar when only 1 and 5 stars were used. Method 1 showed the highest accuracy with only 1% difference compared with traditional supervised classification.

Based on the result of method 4, supervised learning is better in predicting than unsupervised methods in almost all ways. Method 5 used the result of unsupervised learning as a feature of supervised learning. Method 6 emphasized, specifically repeated, the words that had an influence on the aspect's sentiment. Method 5 and method 6 both showed only insignificant increase in accuracy, similar to methods 1, 2, and 3. However, when only 1 and

5 stars were used, method 5 showed about 1% increase from 97% of traditional supervised classification. There was a significant increase in method 6 as well, but only the regression score showed significant increase. Both method 5 and 6 showed highest increase when combining with the results of method 1. Although method 5 and 6 showed increase in scores, the effect of different weights was insignificant. When doing method 6, even deleting the words with a high influence on the aspect's sentiment did not decrease the accuracy by too much.

The conclusion was that the unsupervised methods and combined methods showed high accuracy when people have strong opinions about something, specifically when only 1 and 5 stars were used. However, the weights of the individual aspects cannot be used to significantly increase the prediction of the overall sentiment of a review. If there is enough training data, supervised methods should be used as it has significantly better accuracy. Weights of the aspects do not have to be considered when classifying reviews. Fine tuning the model or dataset would increase the accuracy more than giving weights to the aspect's sentiment.

5.2 Impact

The impact to of this research is finding the insignificance of the weights in the sentiment of relevant aspects. Through this research, it was possible to know that classification and regression of restaurant reviews does not have to consider the weights of the aspect's sentiment. Various classifications methods and regression methods were tested. The research showed that giving additional weights to certain aspects does not significantly increase the accuracy. Also the weight of the aspects does not matter significantly and there is no pattern that can be generalized in all method. Other research that does a study on specific domains do not have to consider the impact of individual aspects or have to consider the unsupervised results when the research can obtain enough training data.

5.3 Future Work

There are several future works that can be experimented with. More aspects can be tested. Only 4 aspects were tested in this paper. Additional aspects such as the 'overall' aspect could be added. Instead of location and music being under ambiance, those aspects

could be divided into different aspects. Other less common aspects such as family friendly environment and pet friendly environment could be added.

Various methods to extract aspects can also be tested. Research on whether the aspect extraction method would affect the accuracy can be useful. More experiments on how the number of keywords for each aspect would affect the result can also be done.

Other methods to give weights can be tested. This paper only focused on giving weights after the machine learning algorithms steps. Giving weights to aspects on an algorithm level could be tested to see the effects.

Finally, a validation of the results as well as the methods could be performed. The research assumed that the library and aspect extraction methods worked well based on other published research. It was not validated to see if those methods specifically work on giving weights to aspects. Experiments such as changing 'good food' to 'bad food' might be experimented to see if the results were to change.

REFERENCES

[1] R. Panchendrarajan, B. Murugaiah, S. Prakhash, M. N. Nazick Ahamed, S. Ranathunga, and A. Pemasiri, "Cheap food or friendly staff? weighting hierarchical aspects in the restaurant domain," in *2016 Moratuwa Engineering Research Conference (MERCon)*, 2016, pp. 24–29. DOI: 10.1109/MERCon.2016.7480110.

[2] S. de Kok, L. Punt, R. van den Puttelaar, K. Ranta, K. Schouten, and F. Frasincar, "Review-level aspect-based sentiment analysis using an ontology," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18, Pau, France: Association for Computing Machinery, 2018, pp. 315–322, ISBN: 9781450351911. DOI: 10.1145/3167132.3167163. [Online]. Available: https://doi.org/10.1145/3167132.3167163.

[3] S. Angelidis and M. Lapata, "Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised," *CoRR*, vol. abs/1808.08858, 2018. arXiv: 1808.08858. [Online]. Available: http://arxiv.org/abs/1808.08858.

[4] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," *CoRR*, vol. abs/1904.02232, 2019. arXiv: 1904.02232. [Online]. Available: http://arxiv.org/abs/1904.02232.

[5] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, and Z. Wu, "Content attention model for aspect based sentiment analysis," ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1023–1032, ISBN: 9781450356398. DOI: 10.1145/3178876.3186001. [Online]. Available: https://doi.org/10.1145/3178876.3186001.

[6] C. Guo, Z. Du, and X. Kou, "Products ranking through aspect-based sentiment analysis of online heterogeneous reviews," *Journal of Systems Science and Systems Engineering*, vol. 27, pp. 542–558, 2018. DOI: https://doi.org/10.1007/s11518-018-5388-2.

[7] W. Wang, H. Wang, and Y. Song, "Ranking product aspects through sentiment analysis of online reviews," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 29, no. 2, pp. 227–246, 2017. DOI: 10.1080/0952813X.2015.1132270. eprint: https://doi.org/10.1080/0952813X.2015.1132270. [Online]. Available: https://doi.org/10.1080/0952813X.2015.1132270.

[8] P. Bhatia, Y. Ji, and J. Eisenstein, "Better document-level sentiment analysis from rst discourse parsing," Sep. 2015. DOI: 10.18653/v1/D15-1263.

[9] V. K. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 2013, pp. 712–717. DOI: 10.1109/iMac4s.2013.6526500.

[10] Yelp, "Yelp open dataset," version 1.0, [Online]. Available: https://www.yelp.com/dataset.

[11] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *Journal of Decision Systems*, vol. 30, no. 2-3, pp. 259–281, 2021. DOI: 10.1080/12460125.2020.1864106. eprint: https://doi.org/10.1080/12460125.2020.1864106. [Online]. Available: https://doi.org/10.1080/12460125.2020.1864106.

[12] H. S. and R. Ramathmika, "Sentiment analysis of yelp reviews by machine learning," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 700–704. DOI: 10.1109/ICCS45141.2019.9065812.

[13] M. Prithivirajan, V. Lai, K. J. Shim, and K. P. Shung, "Analysis of star ratings in consumer reviews: A case study of yelp," in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2954–2956. DOI: 10.1109/BigData.2015.7364134.

[14] R. Panchendrarajan, N. Ahamed, P. Sivakumar, B. Murugaiah, S. Ranathunga, and A. Pemasiri, "Eatery: A multi-aspect restaurant rating system," Jul. 2017, pp. 225–234. DOI: 10.1145/3078714.3078737.

[15] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 81–88. DOI: 10.1109/ICDMW.2011.125.

[16] F. Nurifan, R. Sarno, and K. Sungkono, "Aspect based sentiment analysis for restaurant reviews using hybrid elmowikipedia and hybrid expanded opinion lexicon-senticircle," *International Journal of Intelligent Engineering and Systems*, vol. 12, pp. 47–58, Dec. 2019. DOI: 10.22266/ijies2019.1231.05.

[17] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 12, pp. 1–135, Jan. 2008, ISSN: 1554-0669. DOI: 10.1561/1500000011. [Online]. Available: https://doi.org/10.1561/1500000011.

[18] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Information Fusion*, vol. 52, pp. 344–356, 2019, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.05.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253518308704.

[19] M. Birjali, M. Kasri, and A. beni hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107 134, May 2021. DOI: 10.1016/j.knosys.2021.107134.

[20] B. Yu, J. Zhou, Y. Zhang, and Y. Cao, *Identifying restaurant features via sentiment analysis on yelp reviews*, 2017. arXiv: 1709.08698 [cs.CL].

[21] Y. Cheng, L. Yao, G. Xiang, G. Zhang, T. Tang, and L. Zhong, "Text sentiment orientation analysis based on multi-channel cnn and bidirectional gru with attention mechanism," *IEEE Access*, vol. 8, pp. 134 964–134 975, 2020. DOI: 10.1109/ACCESS.2020.3005823.

[22] J. Barry, "Sentiment analysis of online reviews using bag-of-words and lstm approaches," in *AICS*, 2017.

[23] H. Chen, J. Liu, Y. Lv, M. H. Li, M. Liu, and Q. Zheng, "Semi-supervised clue fusion for spammer detection in sina weibo," *Information Fusion*, vol. 44, pp. 22–32, 2018, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.11.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300714.

[24] A. Kruspe, M. Häberle, I. Kuhn, and X. Zhu, "Cross-language sentiment analysis of european twitter messages during the covid-19 pandemic," Aug. 2020.

[25] A. Imran, S. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," *IEEE Access*, vol. 8, pp. 181 074–181 090, Oct. 2020. DOI: 10.1109/ACCESS.2020.3027350.

[26] K. Yadav, "A comprehensive survey on aspect based sentiment analysis," Jun. 2020.

[27] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. DOI: 10.1109/TAFFC.2020.2970399.

[28] J. Wang, W. Tong, H. Yu, *et al.*, "Mining multi-aspect reflection of news events in twitter: Discovery, linking and presentation," in *2015 IEEE International Conference on Data Mining*, 2015, pp. 429–438. DOI: 10.1109/ICDM.2015.112.

[29] Y. Han, M. Liu, and W. Jing, "Aspect-level drug reviews sentiment analysis based on double bigru and knowledge transfer," *IEEE Access*, vol. 8, pp. 21 314–21 325, 2020. DOI: 10.1109/ACCESS.2020.2969473.

[30] S. Chakraborty, P. Goyal, and A. Mukherjee, "Aspect-based sentiment analysis of scientific reviews," *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 2020.

[31] M. Pontiki, D. Galanis, H. Papageorgiou, *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," Jan. 2016, pp. 19–30. DOI: 10.18653/v1/S16-1002.

[32] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, 2016. DOI: 10.1109/TKDE.2015.2485209.

[33] M. Hu and B. Liu, "Mining opinion features in customer reviews," Jul. 2004.

[34] L. Shu, B. Liu, H. Xu, and A. Kim, "Supervised opinion aspect extraction by exploiting past extraction results," Dec. 2016.

[35] W. Jin, H. H. Ho, and R. K. Srihari, "Opinionminer: A novel machine learning system for web opinion mining and extraction," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, Paris, France: Association for Computing Machinery, 2009, pp. 1195–1204, ISBN: 9781605584959. DOI: 10.1145/1557019.1557148. [Online]. Available: https://doi.org/10.1145/1557019.1557148.

[36] T. Hofmann, "Learning the similarity of documents: An information-geometric approach to document retrieval and categorization," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12, MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/1999/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[37] T.-J. Zhan and C.-h. Li, "Semantic dependent word pairs generative model for fine-grained product feature mining," May 2011, pp. 460–475. DOI: 10.1007/978-3-642-20841-6_38.

[38] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," Aug. 2011, pp. 618–626. DOI: 10.1145/2020408.2020505.

[39] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 380–385. DOI: 10.18653/v1/N19-1035. [Online]. Available: https://aclanthology.org/N19-1035.

[40] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, *Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification*, 2019. arXiv: 1908.11860 [cs.CL].

[41] X. Li, L. Bing, W. Zhang, and W. Lam, *Exploiting bert for end-to-end aspect-based sentiment analysis*, 2019. arXiv: 1910.00883 [cs.CL].

[42] S. G. Tamilselvam, S. Nagar, A. Mishra, and K. Dey, "Graph based sentiment aggregation using conceptnet ontology," in *IJCNLP*, 2017.

[43] M. Makrehchi, S. Shah, and W. Liao, "Stock prediction using event-based sentiment analysis," Nov. 2013, pp. 337–342. DOI: 10.1109/WI-IAT.2013.48.

[44] J. Märkle-HuSS, S. Feuerriegel, and H. Prendinger, "Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures," Jan. 2017. DOI: 10.24251/HICSS.2017.135.

[45] C. Pandi, T. S. Dandibhotla, and V. v. Bulusu, "Survey on sentiment analysis methods for reputation evaluation," Aug. 2018, ISBN: 978-981-13-0616-7. DOI: 10.1007/978-981-13-0617-4_6.

[46] F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou, "Textual and contextual patterns for sentiment analysis over microblogs," Apr. 2012. DOI: 10.1145/2187980.2188073.

[47] O. Tsur and A. Rappoport, "Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews," Jan. 2009.

[48] H. Liu, J. He, T. Wang, W. Song, and X. Du, "Combining user preferences and user opinions for accurate recommendation," *Electronic Commerce Research and Applications*, vol. 12, pp. 14–23, Feb. 2013. DOI: 10.1016/j.elerap.2012.05.002.

[49] M. Eirinaki, S. Pisal, and S. Japinder, "Feature-based opinion mining and ranking," *Journal of Computer and System Sciences - JCSS*, vol. 78, Jul. 2012. DOI: 10.1016/j.jcss.2011.10.007.

[50] N. Archak, A. Ghose, and P. Ipeirotis, "Show me the money!: Deriving the pricing power of product features by mining consumer reviews," Dec. 2007, pp. 56–65. DOI: 10.1145/1281192.1281202.

[51] C. Mate, "Product aspect ranking using sentiment analysis: A survey," 2016.

[52] Z.-J. Zha, J. Yu, J. Tang, M. Wang, and T.-S. Chua, "Product aspect ranking and its applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1211–1224, 2014. DOI: 10.1109/TKDE.2013.136.

[53] V. Sydorenko, S. Kravchenko, Y. Rychok, and K. Zeman, "Method of classification of tonal estimations time series in problems of intellectual analysis of text content," *Transportation Research Procedia*, vol. 44, pp. 102–109, Jan. 2020. DOI: 10.1016/j.trpro.2020.02.015.

[54] A. García-Pablos, M. Cuadros, and G. Rigau, "V3: Unsupervised aspect based sentiment analysis for semeval2015 task 12," Jan. 2015, pp. 714–718. DOI: 10.18653/v1/S15-2121.

[55] Y. Luo and X. Xu, "Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp," *Sustainability*, vol. 11, p. 5254, Sep. 2019. DOI: 10.3390/su11195254.

[56] G. Peng, "Aspect based rating prediction for yelp customer review," 2020. DOI: 10.17615/94b0-n307.

[57] A. Sihombing and A. Fong, "Fake review detection on yelp dataset using classification techniques in machine learning," in *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 2019, pp. 64–68. DOI: 10.1109/IC3I46837.2019.9055644.

[58] J. Prusa, T. M. Khoshgoftaar, and N. Seliya, "The effect of dataset size on training tweet sentiment classifiers," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2015, pp. 96–102. DOI: 10.1109/ICMLA.2015.22.

[59] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.

[60] R. Rolczyski, *Do you trust in aspect-based sentiment analysis? testing and explaining model behaviors*, Mar. 2021. [Online]. Available: https://rafalrolczynski.com/2021/03/07/aspect-based-sentiment-analysis/.