CHAPTER 2

LITERATURE REVIEW

2.1 summary

This chapter's objective is to survey recent studies and reviews of relevant literature on sentiment analysis. An introduction to the many sentiment analysis levels follows, including topics such as data collecting, data pre-processing, sentiment analysis approaches, and lastly, challenges in the field. This chapter gives a good foundation for the sentiment analysis by covering it in detail.

2.2 Degree of sentiment analysis

Several distinct degrees of investigation have been conducted on the subject of sentiment analysis. It is largely possible to identify sentiments and viewpoints at the level of the text, phrase, or aspect [Do et al., 2019]. An illustration of the degrees of sentiment analysis may be seen in Figure 2.1. The first two levels are very great to go through, but they are also really challenging. In spite of this, the third level is more challenging than the levels that came before it since it demands a more comprehensive examination. [Cambria et al., 2017]. Below you will find a brief summary of each level being discussed.

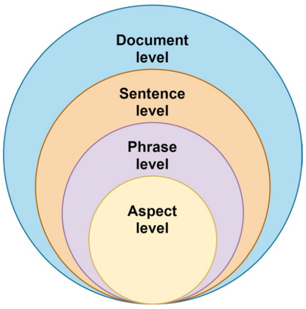

Figure 2.1 Various levels of emotional analysis [Wankhade et al., 2022].

2.2.1 Documents for sentiment analysis

This particular Part, trying to figure out if the overall attitude or viewpoint is favorable or negative. [Alqaryouti et al., 2020].Every text is classified based on the overarching opinion expressed by the author towards a certain object (such a product, for instance). When a single individual is responsible for its creation, categorization at the document level exhibits the highest level of success, and it isn't suited for publications that evaluate or compare many entities. There have been a lot of different approaches proposed for doing sentiment analysis at the document level. For the purpose of conducting document-level sentiment analysis, a framework that is independent of domains was designed [Zhao et al., 2017]. This framework makes use of weighting criteria that are taken from Rhetorical Structure Theory (RST). After creating rhetorical structure trees, the authors analyzed the papers by calculating the emotion scores of phrases using two existing lexicons. This technique was used to analyze the articles. They compiled the scores of the sentences in accordance with the weighting criteria that had been developed in order to ascertain the condition of polarity of the emotion that was expressed throughout the article. The use of analysis in sentiment is very advantageous across a wide range of application areas; yet, documents may sometimes include competing sentiments that have the potential to impact the final judgment.

When doing analysis of sentiment at the document part, it's necessary to examine that the complete content. This type of sentiment analysis is defined by assigning a single polarity to the entire text. This specific sentiment analysis is hardly used. According to the information that it gives, this tool has the capability of categorising the chapters or pages of a book as either positive, negative, or neutral. Currently, both supervised and unsupervised learning approaches may be employed for material categorisation [Bhatia et al., 2015]. The analysis of sentiment across several domains and the analysis of sentiment across numerous languages are the two most essential challenges within the arena of document-level sentiment analysis. According to [Saunders, 2021], When it comes to the domain in issue, domain-specific sentiment analysis is very accurate and highly sensitive to the language used. In the context of this discussion, the feature vector is a collection of words that in addition to being restricted in scope, must also be particular to the subject at hand.

2.2.2 Analysis of opinion at the sentence level

During this stage, the phrase is the primary focus of attention. The major purpose of this analysis is to ascertain if the language transmits a positive, negative, or neutral mood [Liu, 2022]. In order to achieve this aim, the statement must be classified as either objective, which conveys information that is true, or subjective, which reflects views and opinions.

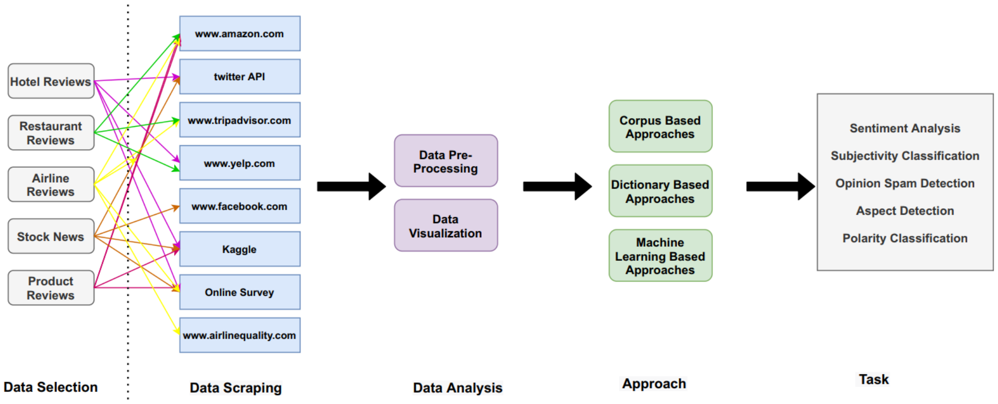

Figure 2.2 Procedures that are often used in sentiment analysis [Wankhade et al., 2022].

2.2.3 At the phrase level, an analysis of the sentiment

There is also the possibility of doing analysis of sentiment at the phrase part, which involves the extraction and that which is classified as decision-making words. It is possible for each phrase to include a single element or a handful of items. It is worth noting that a single aspect is stated in a sentence [Thet et al., 2010], which may prove to be advantageous for product evaluations that span many lines. Recently, this has been a popular topic of inquiry that has been being conducted. Given thatIt is more helpful to do sentence-level analysis rather than text-level analysis when analysing a document that includes both positive and negative statements. Analysis of the text-level focuses on categorizing the whole content as either favorably or adversely subjective. The word is the basic unit of language, and the polarity of a word is strongly tied to the subjectivity of the sentence or text in which it is used. There is a high probability that a statement that contains an adjective is subjective [Fredriksen-Goldsen and Kim, 2017].In addition, the phrase that was selected for expression is representative of the demographic features of individuals, such as their gender and age, as well as their aspirations, social status, and personality, in addition to an assortment of other psychological and social factors. [Flek, 2020]. In light of this, the phrase serves as the foundation for the interpretation of text sentiment.

2.2.4 An examination of sentiment at the aspect level

This level conducts an in-depth examination to determine the sensations that are associated with the different characteristics of entities. Take, for example, the sentence that states, ”The screen resolution of the iPhone 11 is exceptional.” ”screen,” which is a feature of the entity ”iPhone 11,” is the subject of this review, and the evaluation is positive. Therefore, because of this, the task at this level makes it easier to precisely identify the preferences and aversions of people [Indurkhya and Damerau, 2010]. It does not focus on analysing the mood of paragraphs or words but rather places an emphasis on the properties of objects (for example, the attributes of a product). Among the essential tasks of sentiment analysis is the extraction of aspects, which encompasses both implicit and explicit aspects, as stated in [Tubishat et al., 2018]. An analysis of implicit aspect extraction methodologies was presented by the authors, who looked at the topic from a variety of angles.

2.3 The gathering of information and the selection of features

2.3.1 Data Gathering

As seen in figure 2.2, the process of gathering information from the internet may be achieved using a wide range of techniques. These techniques include online scraping, social media, news channels, e-commerce websites, forums, weblogs, and a multitude of other websites. When doing sentiment analysis, the first step is to gather data, which is the beginning of the process.

- Web scraping is the act of automatically getting data from websites, which may hold a substantial number of important information such as product details. This information may be obtained via the process of web scraping. The process is referred to as ”Internet scraping," which is the word used to describe it. On the other hand, this information may be used for a variety of reasons, one of which is the examination of behavioural states that are characterised by psychological states. Among the many internet scraping programs or services that are available for free download, ParseHub is only one example among many others [Birjali et al., 2021].

- The process of creating data or annotations may be outsourced via the use of crowdsourcing technique. It is possible to create a significant quantity of data in a short period of time, which may be of tremendous benefit. This can be accomplished rather quickly. Amazon Mechanical Turk is a well-known site that is used for crowdsourcing, as pointed out by [Birjali et al., 2021] in his detailed article.

Punctuation marks, which are often commonly referred to as exclamation marks, serve this function particularly when they are used to emphasis the enthusiasm of a positive or negative message. The question mark and the apostrophe are two more punctuation marks that are used in the same manner as is described above.

phrases that are used in slang, include ”lol” and ”rofl,” amongst others. They are a method that is used rather often when one is aiming to infuse a sense of humour into a statement. It is fair to assume that a slang phrase in the text gives evidence of sentiment analysis. This is because opinion tweets are of a certain sort, and it is reasonable to conclude that this term provides proof. In order to better convey their meaning, the definition of the slang term should be modified [Wankhade et al., 2022].

The goal of punctuation marks, which is comparable to the function of exclamation marks, is to emphasise the strength of a statement, whether it be a good or negative one. Not only is the apostrophe included in this category, but the question mark and the apostrophe are also included as supplementary punctuation marks [Wankhade et al., 2022].

2.3.2 Data Input

As a consequence of the growth of Web 2.0, a number of different types of data that were previously unavailable are now available. For the purpose of achieving enhanced sentiment categorization, it is conceivable for research domains that concentrate on sentiment analysis to take use of this version. As a result, the input of a SA system is a collection of papers or media files that are in a range of forms [Poria et al., 2017]. Some examples of these formats include the following:

- A plain text file, often known as a TXT file, is a file that contains text that has not been properly formatted.
- Comma-separated values, often known as CSV files, are plain text files that use commas to provide a separation between data.
- Extended Markup Language, abbreviated as HML A hierarchical text format that differentiates content via the use of individualized tags is known as a language file.
- Serialization and transmission of structured data via a network connection are both possible with the help of the JASON data format.
- Markup language, sometimes known as HTML, is a language that is used to define the structure of a web page.

2.4 Data Pre-processing

When taken as a whole, the information that is compiled from a wide range of sources, most notably social media, is not organised. Furthermore, it is probable that the unprocessed form of this data has a significant quantity of noise, in addition to a considerable number of typographical and grammatical problems [Liu, 2022]. Taking this into consideration, it is of the utmost importance to clean and pre-process the text before undertaking any type of analysis. Preprocessing is a procedure that is performed with the intention of enhancing analysis and decreasing the complexity of the data that is being entered. Eliminating superfluous words that do not contribute to the overall polarity of the text is the method by which this objective might be accomplished. Articles, prepositions, punctuation, and special characters are all examples of words that fall under this category. In Table 1, you will find a selection of tools that are accessible to the general public and may be utilised for a wide range of pre-processing and natural language processing operations. A comparative analysis of a number of natural language processing toolkits was carried out by [Pinto et al., 2016]. This analysis was conducted within the context of both formal and social media writings. The entire process consists of a variety of actions that are performed on a regular basis, including the following:

- The process of tokenisation Tokens are the smaller components that are created by this process. For example, a document may be broken down into phrases, and a sentence can be broken down into words.
- The removal of stop words refers to the elimination of phrases (such as ”the,” ”for,” and ”under”) that do not often contribute to or improve analysis. As a result, these terms are removed in advance. act of identifying various structural components of a text, including as verbs, nouns, adjectives, and adverbs, is referred to as part-of-speech (POS) tagging. You may also hear this technique referred to as POS tagging.
- Lemmatisation is the process of reducing a word to its most basic form. This process is also known as ”dialectical reduction.” Lemmatisation is a process that is comparable to stemming; however, it preserves word-related information, such as identifiers for parts of speech, in its whole.

The kind of data also plays a role that is being supplied, the pre-processing step could be different from one particular instance. Two examples of the extra processing and cleaning methods that are required for some formats are the extension of abbreviations and the elimination of repeated letters, such as the ”I” in ”liiiiiiike”. Both of these procedures are essential for certain formats. According to what was said before, textual data may be rather noisy; hence, in order to do sentiment analysis more effectively, it is necessary to follow two fundamental processes. This process consists of two steps: the extraction of features and the selection of features, both of which will be detailed in the next paragraph for your comprehension [Birjali et al., 2021].

2.5 Sentiment analysis techniques

In addition to being a dynamic and flourishing topic of research, sentiment analysis is also a subject that has the potential to be used in a wide range of sectors. For this reason, academics are always going through the process of proposing, evaluating, and comparing a wide variety of different approaches. Improvements in the performance of sentiment analysis and the development of solutions to the problems that are encountered in this industry are the goals of this project.

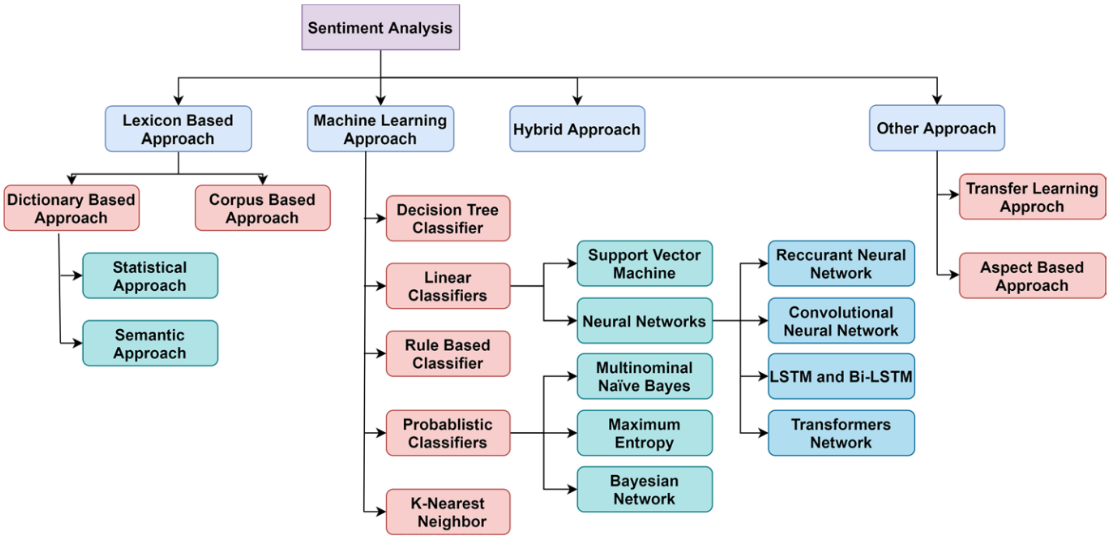

Figure 2.3 Approach of sentiment analysis.

2.5.1 Machine learning approaches

Through the use of both training and testing datasets in combination with machine learning methods, it is possible to achieve the categorization of the polarity of sentiment, which includes categories such as negative, positive, and neutral. There are a variety of categories that may be applied to these approaches, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, to name just a few. These are only some of the many possible classifications. It is necessary to employ the supervised method in circumstances when the problem of the classification is involving a specified collection of classes. On the other hand, in situations when there is a dearth of data that has been labeled, the unsupervised method could present itself as the most successful strategy. Unlabeled datasets that contain a subset of labeled samples are a good candidate for the semi-supervised technique, which may be utilized in such situations. It is the goal of reinforcement learning algorithms to equip the agent with the capacity to interact with its environment in order to maximise the accumulation of rewards. These algorithms make use of methods that incorporate trial and error in order to accomplish this goal. It is feasible for machine learning algorithms to recognise domain-specific patterns from text, which may lead to improved classification results. This is a possibility. In spite of this, these approaches often require enormous training datasets in order to obtain the best possible performance performance. It is of utmost importance to take note of the fact that a classifier that has been trained on a particular dataset does not exhibit the same degree of efficacy when it is applied to a different domain [Pathak et al., 2020].

2.5.1.1 Supervised learning

It is necessary to have training papers that are labeled for supervised techniques, with the labels often indicating the classes (for example, positive, neutral, and negative). Linear, probabilistic, rule-based, and decision tree classification techniques are the four types that fall under the umbrella of supervised classification technique [Yusof et al., 2015]. Following this, the remaining subsections will offer a quick description and comparison of the most common supervised classification algorithms that are used for sentiment analysis.

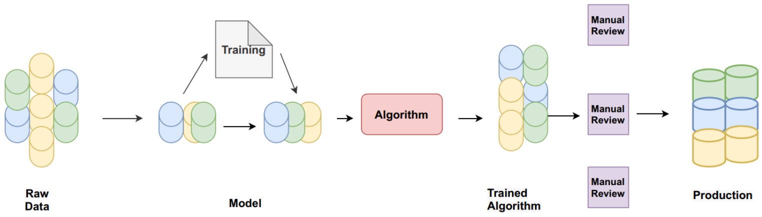

Figure 2.4 Sentiment procedure for supervised machine learning [Wankhade et al., 2022].

- When it comes to the topic of text categorisation, Naïve Bayes (NB) is a fundamental classifier that is often utilised as one of the strategies that is employed the most frequently.Within the framework of the model, the Bayes Theorem is utilised, and the Bag of Words feature extraction approach is utilised in conjunction with it. Because of this, the location of a word within the text is not taken into consideration, and the presence of a certain word is not dependent on the presence of any other words. Consequently, this is the case. With the application of Bayes' theorem, Naïve Bayes assigns a document d to the category c that maximises the probability of P(c/d) in the most effective manner.

Table 2.1 SV MaA Neural Networks Comparison with One Another

Classifier	**Advantages**	**Disadvantages**
SVM	When operating in high dimensional spaces, efficiently and reliably. In comparison to other machine learning algorithms, it is simpler to train and achieves a high level of accuracy. memory-efficient as a result of the advantages that kernel mapping to high-dimensional feature spaces provides. When operating in high-dimensional spaces, efficiently and reliably. In comparison to other machine learning algorithms, it is simpler to train and achieves a high level of accuracy. memory-efficient as a result of the advantages that kernel mapping to high-dimensional feature spaces provides.	In situations when the amount of features greatly exceeds the number of samples, suboptimal performance is more likely to occur. It is very necessary to choose a kernel function that is an appropriate fit. Because there is no probabilistic reason for the noncategorical, there is a limited amount of interpret ability of the data.
ANN	Capable of managing complex relationships among variables and achieving superior generalization, especially with noisy data. Efficient for high-dimensional problems with rapid execution times.	Theoretically complex and challenging to implement. Requires substantial memory and significant training time compared to other methods. May necessitate a large dataset for optimal performance.

$$P(c \mid d) = \frac{p(c)p(d \mid c)}{p(d)} \tag{2.1}$$

In this equation, the term ”p(c)” refers to the prior category of the probability c, ”p(d/c)” represents the conditional likelihood of document d being classified into category c, and ”p(d)” represents the prior of the document to be classified into category d. All of these terms are used interchangeably. A method known as Naïve Bayes is capable of computing the posterior probability of a class by utilising the word distribution that is present within the text. In order to achieve this goal, it is necessary to make the assumption that particular feature needs are not reliant on one another. According to one possible interpretation, the equation that was discussed before might be rewritten as follows:

$$P(c \mid d) = \frac{p(c)\, p(w_1 \mid c) \, * \, \ldots \, * \, p(w_n \mid c)}{p(d)} \tag{2.2}$$

The Naïve Bayes method was utilized in a multitude of studies as a classification software. [Hasan et al., 2015] has developed a classifier that utilities the Naïve Bayes approach to classify opinions expressed in both English and Bangla. This classifier achieves a remarkable level of accuracy. In addition, they utilized their classifiers to conduct an analysis of a number of random evaluations and tweets, and they achieved remarkable outcomes in the majority of the cases.

Rule-based approach Taking this into consideration, the classifiers that are utilised in this technique are dependant on a certain set of criteria in order to carry out the process of sentiment categorisation. As an illustration, let's consider a rule that can be stated as LHS → RHS. Disjunctive Normal Form (DNF) is used to articulate the feature set, while the left-hand side (LHS) indicates the antecedent of the rule or a collection of conditions that apply to the feature set. At the opposite end of the spectrum, the right-hand side (RHS) of the rule is a representation of the conclusion or outcome (class label) of the rule, which is dependant upon the left-hand side (LHS) being met [Tung, 2009]. The efficacy of rule-based classifiers is equivalent to that of decision trees, and they are able to categorise new examples in a very short amount of time. An additional advantage of utilising the rule-based method is that it allows for the prevention of over-fitting, which is a problem that might arise. If, on the other hand, there are an excessive number of regulations, then the interpretation of those rules becomes extremely challenging and laborious. Furthermore, it has a performance that is below average when dealing with noisy data. This is a significant limitation.

In the field of classification, rule-based classification refers to a strategy that employs IF-THEN rules with the purpose of predicting the outcomes of classes [Tung, 2009]. The classifiers that are used in this method are thus dependent on a predetermined set of criteria in order to carry out the process of sentiment categorization. LHS → RHS is one possible expression for a rule.In Disjunctive Normal Form (DNF), the left-hand side (LHS) represents the antecedent of the rule or a set of conditions related to the feature set. On the other hand, the right-hand side (RHS) indicates the conclusion or result (class label) of the rule based on the fulfillment of the label on the left-hand side (LHS) [Sankar and Subramaniyaswamy, 2017].

Classifiers that are based on rules are able to quickly categorise new cases, and their efficiency is equivalent to that of decision trees. The ability of the rule-based approach to avoid overfitting is yet another advantage of using this methodology. In spite of this, its interpretation becomes difficult and complicated when there are a great number of regulations to consider. Furthermore, when it is presented with noisy data, it displays performance that is not enough.

2.5.1.2 Semi-supervised learning

In situations when it is difficult to collect labeled data, semi-supervised learning (SSL) techniques are used. However, in contrast to unsupervised approaches, this strategy makes use of a limited quantity of initial labeled training data in order to have an effect on the process of feature learning. As a consequence of this, it is situated somewhere in the center of the continuum that separates supervised and unsupervised methods. Utilizing SSL techniques allows for the efficient use of large amounts of inexpensive unlabeled data in a cost-effective manner. These approaches save a considerable amount of time and effort while also producing a classifier that is capable of strong generalization with a greater quantity of data that has been labeled. [Zhu et al., 2013]. [Hussain and Cambria, 2018] offered a technique to learning that is semi-supervised and was designed for the goal of analyzing Big Social Data. The use of support vector machines (SVM) and random projection scaling in combination with one another is what contributes to the realization of this possibility. It seems that this semi-supervised model has the potential to considerably increase the performance of some natural language processing tasks, such as sentiment analysis, based on the data that has been collected. Research that was carried out not too long ago on SSL-based sentiment analysis may be broken down into five main categories: generative, co-training, self-training, graph-based, and multi-view learning. [Xia et al., 2015].

Generative approach In order for this method to be able to compute the parameters of each distribution, there must be at least one data point identified as belonging to each category. [Han et al., 2020]. This technique is predicated on the premise that data from a variety of categories correspond to a variety of distributions. Following the training of the model for each class, a generative model will generate distributions for the inputs. Bayes' theorem will then be utilised in order to make a predictive prediction about the label (class) of a test input. After the model has been trained for each possible class, this step is carried out. [Mesnil et al., 2014] This article presents a straightforward yet reliable ensemble approach for sentiment analysis. A tree complementing technique and theoretically baseline models are both incorporated within this strategy. The usage of a generative process was really implemented in one of them. When the entire system is applied to the dataset that is comprised of IMDB movie reviews, it achieves a performance that is unparalleled and at the cutting edge of industry standards. The fact that this is the case suggests that ensemble learning may be utilised in circumstances that are either semi-supervised or in which there is no supervision that is present.

Self-training approach. The concept of self-training may be broken down into two distinct stages. A little amount of labelled data is used to train the classifier in the beginning stages of its training. In the succeeding step, the trained classifier is used to classify unlabeled data, therefore including the samples with the highest level of confidence into the initial training set as new labeled data [Gao et al., 2014]. Iterative execution will be used to carry out the final phase, which will include the newly tagged data. After that, the model that was produced is evaluated by making use of the test data. The approach in question has been used widely within the field of sentiment analysis, as seen by the citation [Hajmohammadi et al., 2015]. It was presented by [He and Zhou, 2011] that a new framework was developed by using a self-training approach. This methodology gets information from labeled features rather than labeled instances itself. As a result of the outcomes of the experiment, it may be concluded that their approach outperformed various recognized methodologies.

2.5.1.3 Reinforcement learning

One of the approaches to machine learning is called reward learning, which is sometimes referred to as reinforcement learning (RL). In this approach, an agent is rewarded in the subsequent time step depending on the evaluation of the activity that it has completed in the time step before it. Reinforcement learning algorithms make use of methods that include trial and error in order to enable the interaction of the agent with its environment in order to maximize the cumulative rewards. This makes it possible for the algorithms to maximize the cumulative rewards. [Li et al., 2020].Among the many challenges that have been addressed via the use of reinforcement learning, robotic control is one of the issues that has been addressed. The majority of its uses, on the other hand, have been in the gaming business. [Wan and Gao, 2015]. On the other hand, despite the fact that it has the capability to tackle complicated tasks, particularly with the incorporation of Neural Networks, the application of this technology to solve problems associated with sentiment analysis is still rather limited. One of the most significant benefits of this approach is that it is fairly comparable to the way in which people learn, which is a characteristic that is greatly sought after in the field of study pertaining to sentiment analysis. throughout the training phase, mistakes that were made throughout the decision-making process are corrected through the use of reinforcement learning, which makes use of previously learnt experiences in order to enhance decision-making and move closer to optimality. In the other way, the process of constructing the model for reinforcement learning might be difficult to do. Furthermore, reinforcement learning necessitates a substantial quantity of data as well as a substantial amount of processing work when it is implemented.

Convolutional neural networks (CNN) Computer vision was the principal use of this design, which is a feed-forward that is a subset of neural networks. Nonetheless, Recent research has shown that it is successful in a number of different areas, including natural language processing and re-commender systems, among others. A convolutional neural network (CNN) consists of three types or layer: an input layer, an output layer, and a hidden layer made up of several convolutional, pooling, normalization, and fully connected layers. All convolutional neural networks (CNNs) have these three layers. Convolutional layers filter inputs (such as word embeddings in text sentiment classification) to extract features. The resolution of the features is decreased by pooling layers, which makes feature recognition more robust against noise and changes of a small magnitude. To provide better convergence during training, the normalization layer standardizes the output of the previous layer and works in tandem with the fully linked layers to perform the classification job. The area of sentiment analysis has seen a rise in the use of Convolutional Neural Networks, or CNNs, in the past several years. [Zhang and Wallace, 2015] presented a CNN model for sentiment analysis that has gained widespread recognition by the scientific community. The author conducted an evaluation of a CNN model that was built using pre-trained word2vec for the goal of categorizing sentiments at the sentence level. In addition to demonstrating that pre-trained word embeddings have the potential to serve as excellent features for natural language processing applications that make use of deep learning, the model outperformed other strategies.

2.5.2 Lexicon-based approach

Two fundamental methodologies are employed for sentiment analysis. One of these techniques is the Lexicon-Based approach, commonly known as the knowledge-based approach. A lexical resource termed an opinion lexicon is essential; it is a predetermined compilation of words that associates terms with their semantic orientation, categorizing them as negative or positive according to assigned ratings. [Hu and Liu, 2004]. It is possible for a score to represent a fundamental polarity value, such as plus one, minus one, or zero, which correspond to words that are positive, negative, or neutral, respectively, or a value that shows the degree or intensity of feeling conveyed by the individual. Both of these values are feasible. The procedure that is utilized in order to discover the ultimate orientation of the composition is the computation of the semantic orientation values of the component words that make up a text. The text is disassembled into its component pieces, which can be either individual words or microphrases, and the lexicon's emotion values are assigned to each of these parts. The microphrases and individual words include the components. One method that may be utilized in order to determine the overarching feeling that is communicated by a piece of literature is the utilization of a formula or algorithm, such as the summing and averaging method.

When it comes to doing sentiment analysis, the lexicon-based strategy is a very effective method that can be utilised at both the sentence and feature levels. Taking into consideration the fact that it does not require any training data, it is conceivable to categorise it as an unsupervised technique. On the other hand, the most significant issue with this strategy is that it is dependent on the domain. This is because words may have several meanings and interpretations; hence, a statement that is judged to be successful in one domain could not have the same connotation in another domain. This is because words can have multiple meanings and interpretations. For instance, when we analyse the word ”small” in connection with the words ”The TV screen is too small” and ”This camera is very small,” the word ”small” in the first sentence assumes a negative meaning due to the fact that people have a tendency to favour screens that are bigger. However, the word ”small” has a positive meaning in the second phrase, which implies that a tiny camera is suitable for transportation. This is because the word ”small” conveys a positive connotation. On the other hand, the deployment of a lexicon adaption strategy or the construction of a sentiment lexicon that is specific to a domain are also viable answers to this problem. [Sanagar and Gupta, 2020a] proposed a strategy for changing an emotion language that is unique to a certain genre of writing. This unique technique takes use of unlabelled data in order to construct sentiment lexicons for both the source domain and the destination domain. This is in contrast to standard implementations of adaptation, which are dependent on data that has been labelled. As demonstrated by the research carried out by [Sanagar and Gupta, 2020b], transfer learning strategies have the potential to be utilised in order to acquire new lexicons that are specialised to a certain domain. A methodology for unsupervised sentiment lexicon learning that is adaptable to new domains within the same genre was proposed by the authors.

Following the acquisition of polarity seed words from corpora spanning many source domains, the information that is particular to the genre is then transferred to the target domains. The performance of the lexicon-based technique is worse in comparison to the performance of the machine learning approach when a significant training dataset is available. In the following paragraphs, we will discuss the three basic approaches to the process of constructing and annotating sentiment lexicons.

2.5.2.1 Manual approach

Through the use of the manual technique, human intervention is required in order to annotate the vocabulary. Within the process of developing sentiment lexicons, there are two stages: the first is the compilation of a list of words that carry a certain feeling, and the second is the assignment of sentiment labels to these words. Despite the fact that this process is often laborious, costly, and time-consuming, it has the potential to produce a vocabulary that is reliable and consistent. In order to speed up this process, it is possible to develop and deploy an automated method. In order to limit the number of errors that occur, a manual approach is used as a benchmarking process. A great number of lexicons have been compiled by the use of physical labor. At the same time as [Wilson, 2005] produced the MPQA Subjectivity Lexicon, [Taboada et al., 2011] established the Semantic Orientation Calculator (SO-CAL). Both of these lexicons are dependent on collections of negators and intensifiers that have been manually selected.

Crowdsourcing and gamification are two more methods that researchers could use. The technique of bringing together a group of people to work towards a common goal via the use of online platforms is known as crowdsourcing. The term ”gamification,” on the other hand, describes the process of incorporating components of games into situations that are not games. Tower of Babel is a game that was designed by [Hong et al., 2013] with the intention of encouraging players to attach emotion polarity to words in order to establish a sentiment lexicon.

2.5.2.2 Dictionary-based approach

This method is based on the idea that words that are synonymous have the same sensation polarity, but those that are antonymous have the opposite degree of polarity. Established dictionaries, such as WordNet9 [Miller et al., 1990] or thesauri

[Mohammad et al., 2009], are used in the construction of the sentiment lexicons that are utilized in this technique. Manual compilation is used to create a collection of initial seed words that have a preset orientation. In the second step, the vocabulary is expanded by investigating synonyms and antonyms with the use of new lexical resources. Until there are no more words to be found, the newly discovered words are gradually added to the list that is already in existence [140]. There is the possibility of doing an additional manual review in order to correct and remove errors. SentiWordNet 3.0 is a well-known lexicon that was constructed by [Baccianella et al., 2010]. This innovative lexicon was created by the automatic annotation of all synsets in WordNet 3. A method for constructing a thesaurus lexicon was provided by [Wankhade et al., 2022] who suggested using three online dictionaries as the tools. Table 8 provides a list of several lexicons that may be used to supplement an original seed. The acquisition of polarity lexicons was the subject of a survey that was carried out by [Sanagar and Gupta, 2016]. The polarity lexicon was investigated by the writers from two different points of view. The first step that they took was to develop the first techniques for developing a polarity lexicon in the first aspect. With regard to the second component, the authors presented relevant information on the open-source polarity lexicon that is available to the public. The end of the article outlined the challenges that are still being faced in the field of research as well as potential paths that may be taken in order to improve polarity lexicons.

The inability of dictionary-based and all lexicon-based techniques to recognize sentiment words with domain-specific connotations is the key problem with these approaches. As a result, these methodologies are not ideal for context and domain-specific classification when it comes to identifying sentiment words. In addition, the compilation of dependency rules is a difficult and labor-intensive process. On the other hand, this method is not computationally intensive as long as there is no dataset training involved. Furthermore, it is an efficient strategy for rapidly constructing a lexicon that includes a significant number of sentiment words and their orientations.

2.5.2.3 Corpus-based approach

Techniques that are based on a corpus, as opposed to those that are based on a dictionary, make use of syntactic or co-occurrence patterns in order to identify new emotion words that have their preferred orientation within a large corpus. Additionally, linguistic constraints on connectives (such as AND, OR, and BUT) are used in order to recognize additional emotion terms. When a conjunction is used to link two adjectives, such as in the phrase ”simple AND easy," the orientation of the adjectives is often the same. In addition to maintaining consistency in emotion, rules may be developed for these connectives; nevertheless, it is possible that these rules will not always be consistent in practice. Methods such as clustering may be used after the operation has been completed in order to generate sets of emotion words, which may include both positive and negative phrases [Liu and Liu, 2011]. [Hatzivassiloglou, 1997] were the ones who originally presented this method to the public. The writers chose words that appeared in the pattern W1 and W2 and had the same orientation in order to widen the initial collection of often recurring adjectives with their orientation. This was done in order to broaden the scope of the collection. The researchers used a network containing words as vertices and their pairings as edges, as well as a log-linear model, in order to identify whether or not two conjoined adjectives had opposing orientations and to categories them into positive and negative terms. Corpus-based techniques are straightforward; but, in order to recognize the polarity of words and the emotions they convey in text, they need a large dataset [Agarwal et al., 2016]. Numerous strategies that are based on corpora are often categorized as either statistical or semantic approaches [Vyas and Uma, 2019], as will be detailed in the subsequent subsections.

Statistical approach

Semantic approach In contrast to the previous method, the ontology-based approach measures word similarity and gives the same emotion value to semantically comparable phrases [Araque et al., 2017]. Typically, this technique searches sentiment dictionaries for synonyms, antonyms, and related terms to expand a vocabulary and analyses sentiment, as shown by [Zhang et al., 2012]. The authors developed Weakness Finder, an expert system that analyses Chinese reviews to identify product weaknesses using a statistical and semantic methodology. [Dong and Dong, 2006] lexicon was utilized to determine word similarity. Experimental findings showed that the suggested expert system performed well.

2.6 Sentiment analysis challenges

2.6.1 The detection of sarcasm

Sarcasm may be defined as "the act of speaking in a manner that is intended to ridicule someone or communicate displeasure." This is one definition of sarcasm out of many. In the

Macmillan English Dictionary, this term may be found. Sarcasm may also be defined as the act of conveying the opposite of what one is actually attempting to say. This is another definition of sarcasm.

2.6.2 Negation handling

In the field of sentiment analysis, it is essential to properly manage negation phrases like not, neither, nor, and so on since these terms have the potential to reverse the polarity of a sentence. For instance, the phrase ”This movie is good.” is considered to be a positive statement, but the statement ”The movie is not good.” is considered to be a negative statement [Wankhade et al., 2022].

Regrettably, in certain methods, negation words are not included because they are included in Stop-Word lists or because they are implicitly eliminated due to their neutral emotion value in a lexicon. This does not have any effect on the final polarity. When compared to other techniques, which do not omit negation concepts, this way of thinking is different. On the other hand, carrying out this function by reversing the polarity is challenging due to the fact that negation words may be inserted in a phrase without having any impact on the sensation that is communicated by the text.

According to [Lazib et al., 2020], In order to accomplish the task of negation scope detection, it is recommended that a hybrid neural network that makes use of a syntactic route approach be utilised. The CNN model is utilised to extract pertinent syntactic characteristics between the token and the cue along the shortest syntactic path in both constituency and dependency parse trees. On the other hand, the Bi-LSTM model is utilised to gain contextual representation throughout the entirety of the sentence in both forward and backward orientations. The bidirectional LSTM and CNN are both incorporated with this technology. It is common practice to employ both of these models in tandem with one another. 90.82 percent out of a possible 100 was the F-score that their model was able to obtain.

2.6.3 Spam detection

The detection of spam is a phase that is of utmost significance in the process of sentiment analysis. The opinions that are published on the internet have a significant impact on the decisions that customers make regarding their purchases; thus, spam and fake reviews have the ability to damage the reputations of companies and affect the impressions that customers have regarding products, services, enterprises, or other affiliations. [Cardoso et al., 2018]. A tremendous obstacle is presented by the absence of obvious contrasts between the assessments, which makes it difficult to develop a spam detection system that is able to differentiate fake reviews from a large number of legitimate reviews. A approach for detecting spam that was developed by [Saumya and Singh, 2018] makes efficient use of three characteristics: the mood of the review and the comments that accompany it, content-based variables, and rating deviation. In this method, the review is classified as either spam or non-spam based on the collection of comment data.

2.6.4 Low-resource languages

The bulk of research in the field of sentiment analysis has focused on the English language or other languages that have a sufficient number of linguistic resources, such as sentiment lexicons and labelled text corpora. This is because the English language is the medium through which sentiment analysis is conducted. For the purpose of sentiment analysis, the most common approach is to employ supervised learning algorithms. On the other hand, these approaches are highly dependent on linguistic resources, which may be rather costly to obtain for languages that are not as widely spoken [Ren et al., 2014]. When it comes to linguistic resources, languages that are considered to be low-resource languages (or under-resourced languages) are those that are lacking in resources. There are a few different approaches that may be taken to solve this issue: The process of creating a language resource from the ground up by using unsupervised, semi-supervised, and transfer learning approaches.

REFERENCES

Agarwal et al., 2016. Agarwal, B., Mittal, N., Agarwal, B., and Mittal, N. (2016). Semantic orientation-based approach for sentiment analysis. *Prominent feature extraction for sentiment analysis*, pages 77–88.

Alqaryouti et al., 2020. Alqaryouti, O., Siyam, N., Monem, A. A., and Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*, 20(1/2):142–161.

Araque et al., 2017. Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Baccianella et al., 2010. Baccianella, S., Esuli, A., Sebastiani, F., et al. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta.

Bhatia et al., 2015. Bhatia, P., Ji, Y., and Eisenstein, J. (2015). Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

Birjali et al., 2021. Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Cambria et al., 2017. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A., et al. (2017). *A practical guide to sentiment analysis*, volume 5. Springer.

Cardoso et al., 2018. Cardoso, E. F., Silva, R. M., and Almeida, T. A. (2018). Towards automatic filtering of fake reviews. *Neurocomputing*, 309:106–116.

De Saa and Ranathunga, 2020. De Saa, E. and Ranathunga, L. (2020). Self-reflective and introspective feature model for hate content detection in sinhala youtube videos. In *2020 From Innovation to Impact (FITI)*, volume 1, pages 1–6. IEEE.

Do et al., 2019. Do, H. H., Prasad, P. W., Maag, A., and Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Dong and Dong, 2006. Dong, Z. and Dong, Q. (2006). *Hownet and the computation of meaning (with Cd-rom)*. World Scientific.

Flek, 2020. Flek, L. (2020). Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7828–7838.

Fredriksen-Goldsen and Kim, 2017. Fredriksen-Goldsen, K. I. and Kim, H.-J. (2017). The science of conducting research with lgbt older adults-an introduction to aging with pride: National health, aging, and sexuality/gender study (nhas).

Gao et al., 2014. Gao, W., Li, S., Xue, Y., Wang, M., and Zhou, G. (2014). Semi-supervised sentiment classification with self-training on feature subspaces. In *Chinese Lexical Semantics: 15th Workshop, CLSW 2014, Macao, China, June 9–12, 2014, Revised Selected Papers 15*, pages 231–239. Springer.

Gope et al., 2022. Gope, J. C., Tabassum, T., Mabrur, M. M., Yu, K., and Arifuzzaman, M. (2022). Sentiment analysis of amazon product reviews using machine learning and deep learning models. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pages 1–6. IEEE.

Hajmohammadi et al., 2015. Hajmohammadi, M. S., Ibrahim, R., Selamat, A., and Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information sciences*, 317:67–77.

Han et al., 2020. Han, Y., Liu, Y., and Jin, Z. (2020). Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications*, 32:5117–5129.

Hasan et al., 2015. Hasan, K. A., Sabuj, M. S., and Afrin, Z. (2015). Opinion mining using naive bayes. In *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 511–514. IEEE.

Hatzivassiloglou, 1997. Hatzivassiloglou, V. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*.

He and Zhou, 2011. He, Y. and Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616.

Hong et al., 2013. Hong, Y., Kwak, H., Baek, Y., and Moon, S. (2013). Tower of babel: A crowdsourcing game building sentiment lexicons for resource-scarce languages. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 549–556.

Hu and Liu, 2004. Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Hussain and Cambria, 2018. Hussain, A. and Cambria, E. (2018). Semi-supervised learning for big social data analysis. *Neurocomputing*, 275:1662–1673.

Indurkhya and Damerau, 2010. Indurkhya, N. and Damerau, F. J. (2010). *Handbook of natural language processing*. Chapman and Hall/CRC.

Lazib et al., 2020. Lazib, L., Qin, B., Zhao, Y., Zhang, W., and Liu, T. (2020). A syntactic path-based hybrid neural network for negation scope detection. *Frontiers of computer science*, 14:84–94.

Li et al., 2020. Li, Y., Fang, Y., and Akhtar, Z. (2020). Accelerating deep reinforcement learning model for game strategy. *Neurocomputing*, 408:157–168.

Liu, 2022. Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

Liu and Liu, 2011. Liu, B. and Liu, B. (2011). Opinion mining and sentiment analysis. *Web data mining: exploring hyperlinks, contents, and usage data*, pages 459–526.

Mesnil et al., 2014. Mesnil, G., Mikolov, T., Ranzato, M., and Bengio, Y. (2014). Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*.

Miller et al., 1990. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Mohammad et al., 2009. Mohammad, S., Dunne, C., and Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 599–608.

Pathak et al., 2020. Pathak, A. R., Agarwal, B., Pandey, M., and Rautaray, S. (2020). Application of deep learning approaches for sentiment analysis. *Deep learning-based approaches for sentiment analysis*, pages 1–31.

Pinto et al., 2016. Pinto, A., Gonçalo Oliveira, H., and Oliveira Alves, A. (2016). Comparing the performance of different nlp toolkits in formal and social media text. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)(2016)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Poria et al., 2017. Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125.

Rajat et al., 2021. Rajat, R., Jaroli, P., Kumar, N., and Kaushal, R. K. (2021). A sentiment analysis of amazon review data using machine learning model. In *2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA)*, pages 1–6. IEEE.

Rathor et al., 2018. Rathor, A. S., Agarwal, A., and Dimri, P. (2018). Comparative study of machine learning approaches for amazon reviews. *Procedia computer science*, 132:1552–1561.

Ren et al., 2014. Ren, Y., Kaji, N., Yoshinaga, N., and Kitsuregawa, M. (2014). Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. *IEICE TRANSACTIONS on Information and Systems*, 97(4):790–797.

Ripa et al., 2021. Ripa, S. P., Islam, F., and Arifuzzaman, M. (2021). The emergence threat of phishing attack and the detection techniques using machine learning models. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pages 1–6. IEEE.

Sanagar and Gupta, 2016. Sanagar, S. and Gupta, D. (2016). Roadmap for polarity lexicon learning and resources: A survey. *Intelligent Systems Technologies and Applications 2016*, pages 647–663.

Sanagar and Gupta, 2020a. Sanagar, S. and Gupta, D. (2020a). Automated genre-based multi-domain sentiment lexicon adaptation using unlabeled data. *Journal of Intelligent & Fuzzy Systems*, 38(5):6223–6234.

Sanagar and Gupta, 2020b. Sanagar, S. and Gupta, D. (2020b). Unsupervised genre-based multidomain sentiment lexicon learning using corpus-generated polarity seed words. *IEEE Access*, 8:118050–118071.

Sankar and Subramaniyaswamy, 2017. Sankar, H. and Subramaniyaswamy, V. (2017). Investigating sentiment analysis using machine learning approach. In *2017 International conference on intelligent sustainable systems (ICISS)*, pages 87–92. IEEE.

Saumya and Singh, 2018. Saumya, S. and Singh, J. P. (2018). Detection of spam reviews: a sentiment analysis approach. *Csi Transactions on ICT*, 6(2):137–148.

Saunders, 2021. Saunders, D. (2021). *Domain adaptation for neural machine translation*. PhD thesis.

Taboada et al., 2011. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Thet et al., 2010. Thet, T. T., Na, J.-C., and Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.

Tubishat et al., 2018. Tubishat, M., Idris, N., and Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing & Management*, 54(4):545–563.

Tung, 2009. Tung, A. K. (2009). Rule-based classification.

Vyas and Uma, 2019. Vyas, V. and Uma, V. (2019). Approaches to sentiment analysis on product reviews. In *Sentiment Analysis and Knowledge Discovery in Contemporary Business*, pages 15–30. IGI global.

Wan and Gao, 2015. Wan, Y. and Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1318–1325. IEEE.

Wankhade et al., 2022. Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Wilson, 2005. Wilson, T. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*.

Xia et al., 2015. Xia, R., Wang, C., Dai, X., and Li, T. (2015). Co-training for semi-supervised sentiment classification based on dual-view bags-of-words representation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* *and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1054–1063.

Yusof et al., 2015. Yusof, N. N., Mohamed, A., and Abdul-Rahman, S. (2015). Reviewing classification approaches in sentiment analysis. In *Soft Computing in Data Science: First International Conference, SCDS 2015, Putrajaya, Malaysia, September 2-3, 2015, Proceedings 1*, pages 43–53. Springer.

Zhang et al., 2012. Zhang, W., Xu, H., and Wan, W. (2012). Weakness finder: Find product weakness from chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, 39(11):10283–10291.

Zhang and Wallace, 2015. Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhao et al., 2017. Zhao, Z., Rao, G., and Feng, Z. (2017). Dfds: a domain-independent framework for document-level sentiment analysis based on rst. In *Web and Big Data: First International Joint Conference, APWeb-WAIM 2017, Beijing, China, July 7–9, 2017, Proceedings, Part I 1*, pages 297–310. Springer.

Zhu et al., 2013. Zhu, S., Xu, B., Zheng, D., and Zhao, T. (2013). Chinese microblog sentiment analysis based on semi-supervised learning. In *Semantic web and web science*, pages 325–331. Springer.