Master's Thesis

A Study on Subjectivity-oriented Polarity Classification

Dong Na

Supervisor Kiyoaki Shirai

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2023

Abstract

Polarity classification in reviews determines sentiment and uses Nature Language Model(NLP) and Machine Learning(ML). Reviews are labeled positive, neutral, or negative. Accurate classification is crucial for applications such as recommender systems, content creation, and marketing.

In many current approaches, the subjectivity of sentences is often ignored, which can lead to inaccurate or inconsistent results. This is because the sentiment of a sentence may vary greatly depending on the perspective of the person expressing it. Various methods for classifying the sentiment or polarity of movie reviews are discussed in many existing studies, each with its own advantages and limitations. But one of the main challenges is to consider the subjectivity of polarity classification.

To address the challenge mentioned above, the goal of this study is to propose two methods for document-level polarity classification that take into account the subjectivity of sentences to a large extent. Our approach is based on the idea that the polarity of a document is strongly influenced by its level of subjectivity. Reviews with strong subjectivity are more likely to express personal opinions, and such reviews have an important place in the polarity classification, while reviews with weak subjectivity are more likely to provide objective information, and such reviews tend to be less important in the polarity classification. By considering the weight of subjectivity in sentences, we aim to improve the accuracy of polarity classification.

We propose Polarity Classification by Subjectivity Weighted Voting (PCBWV) and Polarity Classification by Pre-trained Language Model with Subjectivity Filtering (PCPLM-SF). In PCBWV, first, after preprocessing which includes tokenization, lemmatization, and removing stop words, we split the IMDb review into sentences. Then we use bert-base-uncased and xlnet-base-cased to train the IMDb dataset to get the polarity of each sentence in the given review. Next, We train the xlnet-base-cased model on a subjectivity dataset to get the subjectivity classifier. Then, using this classifier and the softmax function, we train sentences from the IMDb dataset to determine a subjectivity score for each sentence in a movie review. The score takes into account the subjectivity of each sentence, with more weight given to subjective sentences. The overall polarity label of a review is determined by the relative weight of the different sentiment categories of the sentences, while the weighting is based on the subjectivity of each sentence.

On the other hand, in PCPLM-SF, The movie review is split into sentences after preprocessing, as done in the previous model PCSWV. Then train

the xlnet-base-cased model on the subjectivity dataset to get the subjectivity classifier and it is performed on each sentence. If a sentence is classified as objective, it is filtered out. And only the sentences classified as subjective can remain. The subjective sentences that remain are concatenated in the order they appear in the original review to form the pseudo review. bert-base-uncased or xlnet-base-cased is fine-tuned using the training data of the pseudo reviews. Besides, we use the subjectivity-only(S) pseudo dataset as the training dataset, while the subjectivity-only dataset, objectivity-only dataset(O), and the original (containing both subjectivity and objectivity) dataset(S+O) are used as the test dataset.

We present our findings from working with three datasets: the subjectivity movie dataset (5,000 subjective sentences from the Rotten Tomatoes website, 5,000 objective sentences from IMDb plot summaries), the IMDb dataset (50,000 movie reviews, including 25,000 positive reviews and 25,000 negative reviews, as well as 25,000 used for training and 25,000 for testing), and the Amazon dataset (a collection of user reviews posted on the Amazon website, we randomly select 25,000 for training and 25,000 for testing). We describe two methods for polarity classification that take into account sentences with subjectivity and present the results. We also use pre-trained models, such as BERT and XLNet, to obtain better results in the datasets. In our experiment, we use the default parameters of the models.

For the evaluation of the PCBWV method, our results show that the XLNet model is able to achieve the best result of 96% accuracy on the subjectivity movie dataset. When it comes to polarity classification on the IMDb dataset, our method, which uses XLNet and takes into account the subjectivity scores, has an accuracy of 85.3% compared to the accuracy of 82.9% achieved by traditional methods that simply count positive and negative sentences and choose the most frequent polarity class without considering subjectivity.

And for the evaluation of the PCPLM-SF method, we aim to study the impact of subjective reviews on polarity classification (positive or negative sentiment). We use three datasets: S (subjective sentences), O (objective sentences), and S+O (a combination of both subjective and objective sentences). For each training set, we use the three datasets as the test set. The proposed systems, PCPLM-SF-1 and PCPLM-SF-2, use S as the training data and S+O or S as the test data. All other test results are considered as the baseline method. The results indicate that for the Amazon dataset, the PCPLM-SF-1 method which uses S for training and S+O for testing with BERT achieved the highest accuracy of 95.3%. For the IMDb dataset, the PCPLM-SF-2 method using S for training and S+O for testing showed a high accuracy of 98.0%. However, the best result was achieved when both training

and testing were done using S+O, resulting in an accuracy of 99.7%. In the PCPLM-SF method, BERT outperforms XLNet overall. For the IMDb and Amazon datasets, Amazon outperforms IMDb, except when both training and testing sets were S+O.

Finally, we evaluated different methods and shared lessons learned from running this task on two different models. The result indicates that our PCPLM-SF method performs better than the PCBWV method. We also reflect on the open challenge of state-of-the-art systems on IMDb datasets, where pre-trained models are more effective on the original dataset, and we analyze that objective sentences in the IMDb dataset also carry important information that facilitates polarity classification.

Despite the promising results, there are still some open challenges that need to be addressed. The main challenges in sentiment analysis include poor performance on the IMDb dataset by current systems, limited generalization to non-English datasets, and lack of data diversity. Future work should focus on overcoming current limitations and improving performance on challenging datasets. The results of this study provide a foundation for future work in the field of sentiment analysis, which still has much room for improvement and ongoing research.

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Goal . . . 2
- 1.3 Thesis Outline . . . 2

2 Related Work **4**
- 2.1 Sentiment Analysis . . . 4
- 2.2 Polarity Classification . . . 5
- 2.3 Application of Subjectivity Classification in NLP . . . 6
- 2.4 Language Models . . . 8
  - 2.4.1 BERT . . . 8
  - 2.4.2 XLNet . . . 10
- 2.5 Characters of this study . . . 12

3 Proposed Method **13**
- 3.1 Task Definition . . . 13
- 3.2 Motivation for subjectivity-oriented method . . . 14
- 3.3 Polarity Classification by Subjectivity Weighted Voting . . . 16
  - 3.3.1 Data preprocessing . . . 17
  - 3.3.2 Polarity classification of sentence . . . 18
  - 3.3.3 Subjectivity classification of sentence . . . 18
  - 3.3.4 Weighted Voting . . . 20
- 3.4 Polarity Classification by Pre-trained Language Model with Subjectivity Filtering . . . 22
  - 3.4.1 Data preprocessing . . . 22
  - 3.4.2 Subjectivity filtering . . . 23
  - 3.4.3 Fine-tuning . . . 23
  - 3.4.4 Detail algorithm . . . 24

4 Evaluation **27**
4.1 Dataset 27
4.1.1 IMDb Dataset 27
4.1.2 Amazon Review Dataset 30
4.1.3 Subjectivity datasets 32
4.2 Evaluation Criterion 33
4.3 Result of Subjectivity Classification 33
4.4 Results of Subjectivity Weighted Voting 33
4.5 Results of Language Model with Subjectivity Filtering 34

5 Conclusion **38**
5.1 Summary 38
5.2 Future work 39

List of Figures

2.1 Overview of BERT [9] 9
2.2 XLNet 11

3.1 Overview of polarity classification by subjectivity weighted voting (PCSWV) 16
3.2 Comparison between simple voting and PCSWV 21
3.3 Overall of Overview of PCPLM-SF 22

List of Tables

3.1 Example of a review with subjectivity and polarity of each sentence . . . 15
3.2 Example of subjectivity and objectivity scores obtained by BERT . . . 19
3.3 Example of calculation of subjectivity score . . . 19
3.4 Example of review after the polarity and subjectivity classificaton . . . 21
3.5 Example of original review . . . 23
3.6 Example of subjectivity filtering . . . 24
3.7 Example of pseudo review . . . 25

4.1 Statistics and example of IMDb dataset . . . 28
4.2 Example of the original review and the preprocessed review . . . 29
4.3 Examples of reviews in Amazon review dataset[8] . . . 31
4.4 Examples of reviews in subjectivity dataset . . . 32
4.5 Accuracy of subjectivity classification . . . 33
4.6 Accuracy of polarity classification by subjectivity weighted voting and other baselines in IMDb dataset . . . 34
4.7 Accuracy of polarity classification by language models . . . 35
4.8 Example of classification error . . . 36
4.9 Subjectivity of sentences in misclassified review . . . 37

Chapter 1

Introduction

In this chapter, we first explain the background of our research in Section 1.1. Section 1.2 describes the motivation and goal of this work. Finally, the structure of the thesis is given in Section 1.3.

1.1 Background

An increasingly large number of people use the Internet to express their opinions and share their ideas about products and services because of the rapid growth of social networks. In this situation, an unprecedented amount of user-generated data has been produced. It can provide an excellent opportunity for sentiment analysis. Sentiment analysis(SA), also known as opinion mining, is a study that aims at analyzing massive opinions about products, things, services, organizations, etc. It also aims at revealing a trend of reputation of those things. With the rapid development of computer science, SA is now one of the most active research areas in Natural Language Processing (NLP). It occupies an important place not only in computer science, but also in management and social sciences, and even in history, and has attracted the attention of the whole society because of its commercial importance [20].

Polarity classification, a kind of SA, is a task to classify a given text into polarity, i.e. to judge whether a text expresses positive or negative opinions [10]. The solid and growing interest in polarity classification is reflected in the surge of published articles in the area of affective computing, attracting the highest number of researchers in recent years. A text for polarity classification can be a document, sentence, or aspect, but this study focuses on document-level polarity classification. A typical example is the classification of the polarity of movie reviews.

In general, there are two kinds of sentences in user reviews. One is a

subjective sentence that expresses the writer's emotion or opinion, the other is an objective sentence that refers to objective facts. Intuitively, subjective sentences play a more critical role than objective sentences in polarity classification. It would be helpful to filter out objective sentences or put more priority on subjective sentences in order to improve the performance of polarity classification. Subjectivity is also considered in past studies of SA, such as a subjectivity classification task, but it is not paid much attention in polarity classification.

1.2 Goal

The goal of this study is to propose two methods for document-level polarity classification that take into account the subjectivity of sentences to a large extent. As mentioned earlier, an important topic in polarity classification is to consider the subjectivity of sentences. However, if subjectivity information is incorporated in a model of polarity classification, a certain amount of noise may be introduced, resulting in poor performance of the classification model. Therefore, in this study, the first approach is to determine the polarity of an overall document by voting of the polarity of sentences in it where the polarity of the subjective sentences are highly weighted. We also consider objective sentences, but the impact of objective sentences in polarity classification is less considered than that of subjective sentences.

In addition, on the one hand, it has been demonstrated in many prior studies that the pre-trained language model such as Bidirectional Encoder Representations from Transformers (BERT)[9] and XLNet [30] could achieve state-of-the-art results in many NLP tasks. On the other hand, some previous studies of polarity classification showed that filtering out objective sentences and considering only subjective sentences could improve the model to some extent. Therefore, the second approach in this study is to use pre-trained language models after objective sentences are filtered out from a review. We investigate how much the model performance is improved by filtering objective sentences from the reviews in either training data or test data, or both.

1.3 Thesis Outline

The rest of this thesis is organized as follows.

- Chapter 2 discusses related work about sentiment analysis, subjectivity classification, and polarity classification.

- Chapter 3 explains the details of our two proposed methods of the polarity classification considering subjectivity of sentences.
- Chapter 4 reports the results of experiments on two datasets of different domains: IMDb movie review and Amazon review dataset. Furthermore, we also conduct an error analysis and show the major causes of errors.
- Chapter 5 concludes this study and denotes some future work to improve the proposed methods.

Chapter 2

Related Work

This chapter consists of 5 sections. In Section 2.1, we take a brief look at the definition of sentiment analysis, then introduce some related work of it. Section 2.2 introduces the related work for polarity classification. Section 2.3 introduces the application of subjectivity classification in NLP. Section 2.4 introduces two of the most popular language models, BERT and XLNet, and the applications of them in sentiment analysis. Finally, in Section 2.5, we briefly introduce this study, the subjectivity-oriented polarity classification method and clarify the characteristics of this study.

2.1 Sentiment Analysis

Sentiment analysis (SA) is a study to analyze emotional overtones expressed in a text. SA includes many subtasks such as polarity classification, subjectivity classifications, etc.

The purpose of polarity classification is to make positive, negative, and neutral judgments about the reviews. In most cases, there are three categories, positive, neutral and negative. For example, the words "love" and "dislike" belong to different emotional tendencies.

The main purpose of subjectivity classification is to distinguish which parts of the text are objective statements without emotions, and which are subjective descriptions with emotions.

There are two main approaches of sentiment analysis methods: lexicon-based methods and machine learning-based methods [31]. The lexicon-based methods mainly develop a series of sentiment dictionaries, calculate the overall sentiment score of a text based on the number of sentiment words, and then judge the sentiment direction of the text based on the sentiment score [23]. Machine learning-based methods mostly transform SA into a classifi-

cation problem [16]. For example, in polarity classification, an input text is classified into one of the two categories: positive or negative. Manually labeled data is used as training data, and a classification model is obtained from it by a supervised machine-learning process.

Bhaskaran et al. designed a novel Modified Red Deer Algorithm (MRDA) Extreme Learning Machine Sparse Autoencoder (ELMSAE) model for SA [4]. First, as a preprocessing, the proposed MRDA-ELMSAE technique transformed the data into a compatible format. Next, the TF-IDF vectorizer was used to extract features, while the ELMSAE model was used for sentiment classification. Then, the ELMSAE model was optimally parameterized using the MRDA technique. Extensive simulations were performed and the results of the comparative analysis demonstrated the enhanced efficiency of the MRDA-ELMSAE technique relative to other state-of-the-art techniques.

2.2 Polarity Classification

Polarity classification is the task of distinguishing between sentences that express positivity or negativity. Pozzi et al. reported several examples of polarity classification [20].

Polarity classification has been applied to ethical principles. Arunachalam and Sarkar monitored and analyzed several social networks to assess citizens' perceptions of government institutions for several purposes: fine-tuning policies, identifying positive perceptions of best practices, and negative aspects of actions and decisions [2]. In recent years, social networks have emerged as a potential source of information for sentiment analysis in the financial domain. Financial tweets have been investigated for predicting stock market evolution in the short and long term [5].

Vazan et al. proposed a convolutional neural network CNN-based multi-task learning model that can simultaneously aspect categories and their polarity of them to enhance the performance of polarity classification in Persian comments [24]. To alleviate the error problem due to the biased nature and high variance of texts, they used an ensemble learning approach, which means combining several models, to improve the efficiency of the model prediction.

Wattanakitrungroj et al. proposed a low-dimensional vector called V8D for representing text [25]. They did not represent the text by a vector of weights or frequencies of terms that appear in the text, because in the traditional approach, the length of the feature vector was equal to the number of words in the dictionary derived from all possible words in a text collection. The large number of words in the dictionary resulted in a high-dimensional vector representing of the text and introduced a long processing time for

training and testing the text classification model. In their approach, the vectors of the texts were created only from positive and negative phrases, including significant negatives. Four machine learning algorithms for solving classification problems, namely k-Nearest Neighbors, Naïve Bayes classifiers, artificial neural networks, and Support Vector Machines (SVMs), were applied to the classification of opinion texts. The proposed V8D vector was compared with the traditional TF-IDF vector in terms of prediction correctness by experimenting with datasets from eight different domains: IMDb, Amazon, Yelp, Apparel, health, sports, music, and US Airlines. The experimental results showed that opinion text classification with V8D vectors achieved the best efficiency in terms of space usage and processing time.

Kansal et al. overcame the dependence of the Cross Domain Sentiment Classification(CDSC) task on the manual annotation of datasets by applying a polarity detection task [12]. They proposed the CDSC-PDT approach, where a CDSC task precedes a polarity detection task (PDT). In the proposed PDT task the polarity of the comments in the source domain was identified using contextual and relevant information about the words in the document. The PDT was used to automatically annotate the comments with the polarity. This automatically labeled dataset was used to train the classifier for the classification in the target domain. The experimental results demonstrated that the proposed approach achieved comparable performance to traditional CDSC models without manual labeling of documents in any of the domains (source or target). Thus it saved human intervention and was also a time-saving and inexpensive process, unlike the traditional CDSC methods.

2.3 Application of Subjectivity Classification in NLP

The subjectivity of the sentences has been considered in polarity classification in a few previous studies. Pang and Lee proposed a method to extract only subjective sentences from a document by using a minimum cut framework, then classify the polarity of the extracted document by naive Bayes model [17]. Their results of experiments demonstrated that the document consisting of only subjective sentences was not only shorter but also more effective than the original document, namely the accuracy of the polarity classification was significantly improved.

Sindhu et al. proposed a similar method consisting of subjectivity classification and polarity classification [22]. The subjectivity classification was

performed to filter out objective sentences, then the polarity classification was carried out using only subjective sentences.

The subjectivity was also considered in a wide variety of sentiment analysis, e.g. extraction of aspect and opinion words. Kamal proposed a method to extract pairs of features (aspects) and opinions by combining supervised machine learning and rule-based approaches [11]. First, supervised machine learning methods were used to classify subjective and objective sentences in the datasets. Next, a rule-based approach based on a linguistic and semantic analysis of the texts was used to mine feature-opinion pairs from the subjective sentences retained by the first step.

Subjectivity classification is used for not only polarity classification but also other NLP tasks. Ellen et al. described an Information Extraction(IE) system that used a subjective sentence classifier as a filter [21]. They constructed a classifier that judges whether a sentence is subjective or not by combining a rule-based method and Naive Bayes model. Two different strategies were proposed to use the subjectivity classifier for filtering out incorrect extracted information: an aggressive strategy of discarding all extracted information found in subjectivity sentences and a more sophisticated strategy of selectively discarding extracted information. MUC-4 terrorism dataset was used to conduct experiments. Their results showed that indiscriminately filtering extracted information in subjective sentences was overly aggressive, but the more selective filtering strategy improved the precision of IE with minimal loss in the recall.

Zhao et al. proposed a new approach to address bias in the toxic comment classification (TCC) task by exploiting the concept of the subjectivity level of comments and the presence of identity terms [32]. Identity terms are words or terms that refer to specific groups of people, such as "Muslim", "black", "woman", and "Democrat". Such identity terms were often associated with false positive bias. This approach was mainly based on an additional focus on the level of subjectivity of comments where identity words appeared. Three different models considering the above features were proposed:

1. SS-BERT: Subjectivity and identity words (subdentity-sensitive BERT, where "subdentity" denotes "subjectivity" and "identity")
2. SO-BERT: Subjectivity-only BERT
3. SS-BERT+SOC: Subjectivity and identity words with sampling and occlusion (SOC) built on the SS-BERT method.

The subjectivity scores were generated using TextBlob separately to facilitate bias analysis, i.e. the analysis of non-toxic reviews with biased

identity terms such as “Muslim” and “Black”. The results of the experiments showed that combining subjectivity with the presence of identity terms was more informative.

2.4 Language Models

2.4.1 BERT

Recently, BERT[9] has been applied for many tasks of natural language processing (NLP) and often achieved state-of-the-art results. Figure 2.1 shows the architecture of BERT.

There are two main tasks in BERT, pre-training and fine-tuning.

- **Pre-training** (shown on the left side of Figure 2.1): There are two sub-tasks, mask language model (MLM) and next-sentence-prediction (NSP).

  In the MLM, 15% of the tokens are randomly chosen. Then, 10% of the chosen words are replaced with other words, 10% of them are not replaced, and the remaining 80% are replaced with [MASK]. This way the model learns to understand the context of the words in the sentence and also the relationships between the words.

  The next task, NSP, trains a model to predict the next sentence in a given text by using the context of the previous sentence to understand the context and coherence of the text. This is useful in various NLP tasks such as text generation, language understanding, and dialogue systems. NSP performs better in document-level corpora because it can have the ability to abstract continuous long sequence features.

- **Fine-tuning** (shown on the right side of Figure 2.1): It is a technique used to further train a pre-trained model on a new task or dataset. It uses a smaller dataset than that used for pre-training and updates the model parameters, which allows the model to be adapted to a new task or dataset and often improves performance. For the classification task, BERT directly takes the final hidden state $C \in R^H$ of the first [CLS] token, adds a layer of weights $W \in R^{K \times H}$ and then uses softmax to predict the label probability.

Inspired by the success of pre-trained language models, Pota et al. proposed a two-step approach for polarity classification of tweets posted on Twitter [19]. First, tweets including special tokens of Twitter were converted into

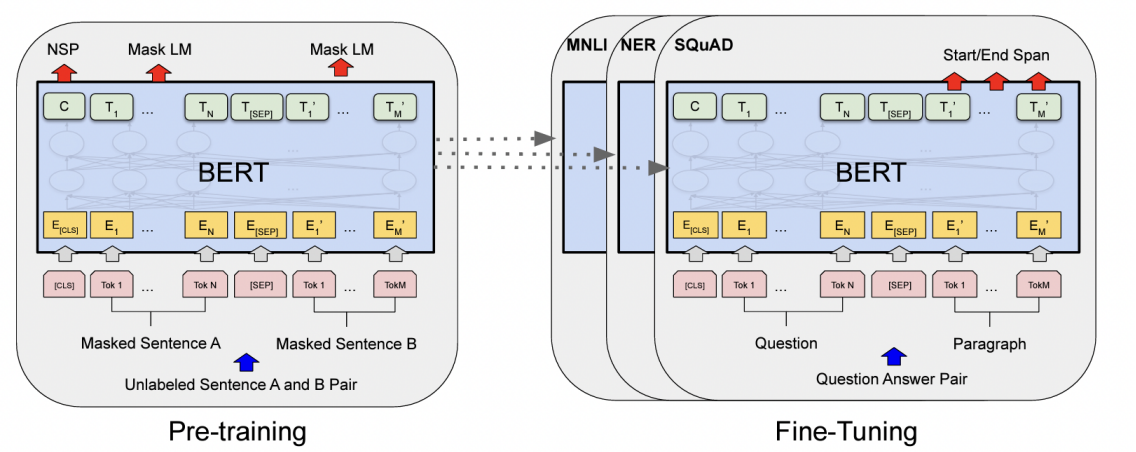

Figure 2.1: Overview of BERT [9]

plain texts using a language-independent or easily adaptable process for different languages. Second, the polarity of the converted tweets was classified using the language model BERT.

BERT has been applied not only polarity classification but also many other NLP tasks. Chen et al. proposed a BERT-based local feature convolutional network (LFCN) model to address the drawbacks of long, informative, and complex structure of Chinese Web news texts, thus improving the accuracy of Chinese long text classification [6]. LFCN consisted of four modules. First, they proposed a method called Dynamic LEAD-n (DLn), which extracted short texts from long texts based on the traditional LEAD summary algorithm to solve the problem that BERT could not accept a long text due to restriction on the maximum input sequence length. Second module was the Text-Text Encoder (TTE) module that used the pre-trained language model of BERT to obtain the sentence-level feature vector representation of news text, where global features were captured by using an attention mechanism that could identify relevant words in the text. Third, they proposed a CNN-based local feature convolution (LFC) module to capture local features in the text, such as key phrases. Finally, feature vectors generated by different modules at several different periods were fused and used to predict the category of news texts. The experimental results showed that the proposed method was effective for Chinese news long text classification.

Kowsher et al. proposed Bangla-BERT, a monolingual BERT model for the Bangla language [13]. They pre-trained BERT on the largest Bangla language model dataset, BanglaLM. Although the available data for pre-training of BERT was limited compared to resource intensive language such

as English, Bangla-BERT achieved the highest results and outperformed multilingual BERT and other previous studies in binary language classification, multi-label extraction, and named entity recognition. Moreover, the model was evaluated by transfer learning based on hybrid deep learning models such as Long Short-Term Memory(LSTM), CNN, and Conditional Random Field(CRF) in the Named Entity Recognition(NER) task. Several experiments were carried out to evaluate Bangla-BERT by using benchmark datasets such as Banfakenews, sentiment analysis of Bengali news reviews, and cross-lingual sentiment analysis of Bengali. They confirmed that Bangla-BERT surpassed all previous state-of-the-art results.

Phan et al. proposed a novel CNN for aspect-level sentiment analysis(ALSA) on the BERT-GCN model [18]. In this paper, first, they addressed the drawback of the model being limited to a few (two or three) layers of graph convolution network (GCN) due to the disappearance of gradients by adding convolutional layers of CNN model after the GCN layers. By combining a BERT and a bidirectional long short-term memory (BiLSTM) model, further useful information about the hidden context between words could be considered. In the BERT-GCN model, firstly, the words in the sentence were converted into vectors using BERT. Second, a contextualized word representation was created on the word vector based on BiLSTM. Third, the GCN model with multiple convolutional layers was used to extract the important features and represent them on the contextualized word representation. Finally, the CNN model was used to classify the polarity of the aspect using the features obtained by GCN. Experiments on three benchmark datasets showed that the combination of BERT and GCN improved the performance of previous context-based GCN methods on ALSA.

2.4.2 XLNet

Yang et al. proposed a generalized autoregressive pre-training method, XLNet, that overturned the era of BERT based language models[30]. In XLNet, bidirectional contexts are learned by maximizing the expected likelihood of all permutations of the factorization order. Thus XLNet integrates ideas from the state-of-the-art autoregressive model Transformer-XL into the pre-training. Due to the autoregressive formulation of XLNet, which overcomes the disadvantage of BERT that ignores dependencies between masked positions. XLNet outperforms BERT on 20 tasks in a comparable experimental setting.

Figure 2.2 shows the architecture of XLNet. It describes the flow of Two-Stream Self-Attention, where a sequence of words in order 1234 is input, and the initialization $h_i^{(0)} = e(x_i)$, $e(x)$ is the value of the word embedding,

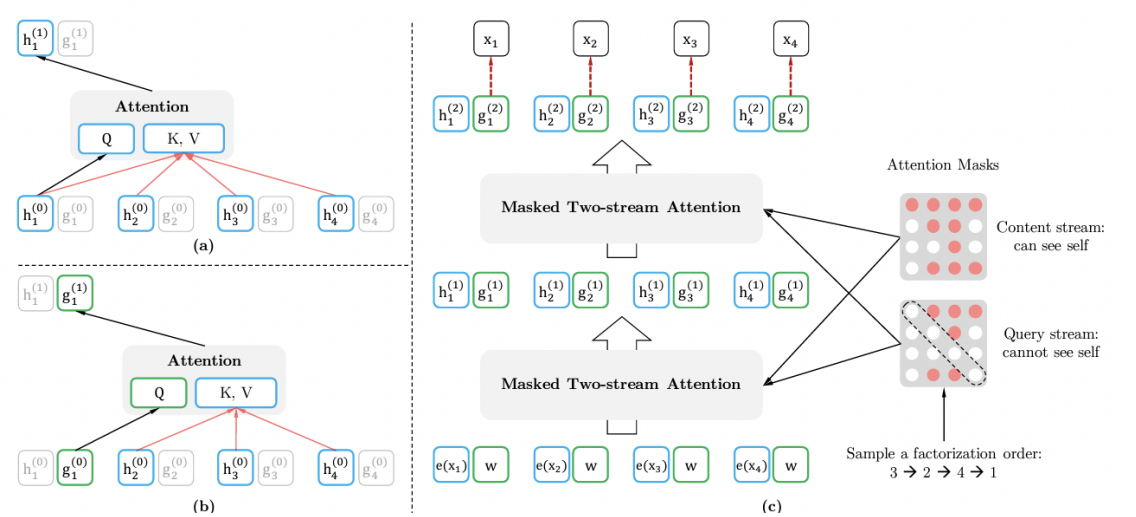

Figure 2.2: XLNet

and $g_i^{(0)} = w$, where w is a trainable parameter. For the word x1, which is "rearranged" to 3241 by multiplying it by the attention masks matrix (as shown on the right side of Figure 2.2), so that x1 is at the end and x1 can see its top (x3, x2, x4, x1 for the Content Stream; x3, x2, x4, x1 for the Query Stream is x3, x2, x4), and then to calculate the prediction probability by repeating multiple layers of Two-Stream Self-Attention to obtain $h_i^{(0)}$ and $g_i^{(0)}$, respectively, as shown in Equation (2.1).

$$P_\theta(X_{z_t} = x | x_{z<t}) = \frac{exp(e(x)^T g_{zt}^{(M)})}{\sum_{x'} exp(e(x')^T g_{zt}^{(M)})} \tag{2.1}$$

Xu et al. proposed a parallel hybrid sentiment analysis network EXLNnet-BG-Att-CNN with integrated sentiment word features to address the problem of poor attention to some sentiment words and difficulty in capturing long-distance dependencies between sentences [28]. In this study, the basic emotion dictionary HowNet, Chinese sentiment word database developed by Dalian University of Technology, and Taiwan University emotion dictionary were combined with additional extensions of relevant words and emotions of them to obtain a more comprehensive dictionary. Then, the sentiment word selection segmentation algorithm (DicSentencePieceSelect) was proposed. After the segmentation by this algorithm, the authors used XLNet to encode the vectors processed by the sentence and dictionary word selection algorithms separately, and merged them to obtain the deep semantic features of the text. Then, these feature vectors were fed into the parallel networks of

BiGRU-ATT(Bi-directional Gated Recurrent Unit-Attention) and TextCNN respectively with a double attention mechanism. Finally, the output vectors of those networks were fused and an activation-pooling layer was used to avoid overfitting. This method could take into account the advantages of global features of text sequences and further extract local features to enhance the semantic representation of the text. The accuracy of EXLNet-BG-Att-CNN was 96.05%, which was higher compared with several existing models.

XLNet was also applied for the NER task. Yan et al. proposed a new neural network model to improve the effectiveness of NER by using pre-trained XLNet, Bi-LSTM and CRF [29]. Using the CoNLL-2003 English dataset and the WNUT-2017 dataset, the pre-trained XLNet model was used to extract sentence features, and then the classical NER neural network model was combined with the obtained features. The results of the experiments showed that XLNet-BiLSTM-CRF obtained state-of-the-art results in the NER task

2.5 Characters of this study

This study shared the basic idea with Pang's method[17] and Sindhu's method[22]: subjective sentences are more important in polarity classification. Although the objective sentences are just ignored in their methods, we suppose that the objective sentences have less but also useful information for polarity classification. Therefore, this thesis investigates the way to use both subjective and objective sentences with priority on the former.

This thesis also explores how to use pre-trained language models for document-level polarity classification. Especially, we investigate how subjective and objective sentences can be handled to improve the performance of polarity classification by the state-of-the-art language models.

Chapter 3

Proposed Method

This chapter explains the details of two proposed subjectivity-oriented polarity classification methods. Section 3.1 describes our task definition. Section 3.2 explains the motivation for our proposed method. Section 3.3 describes our first proposed method, Polarity Classification by Subjectivity Weighted Voting (PCSWV) and Section 3.4 describes our second proposed approach, Polarity Classification by Pre-trained Language Model with Subjectivity Filtering (PCPLM-SF) in details.

3.1 Task Definition

With more and more people posting reviews on the Internet, how to effectively classify user reviews in terms of the polarity has become one of the hot issues. There are several kinds of sentiment analysis or polarity classification: document-level, sentence-level, and aspect-level. The document-level sentiment analysis is a task to classify the polarity of a document such a user review into "positive", "negative", or "neutral". The sentence-level sentiment analysis focuses on individual sentences in a document. Its goal is the classification of a sentence. The aspect-level sentiment analysis is a task to identify the polarity of an aspect of a product or a service. For example, a PC has many aspects such as "CPU", "memory", "keyboard", "price", "design" and so on. The polarity of such aspects is identified in the aspect-level sentiment analysis.

This thesis focuses on the task of document-level sentiment analysis or polarity classification. Especially, the documents to be classified into their polarity are user reviews consisting of several sentences. On Web, since many people write user reviews about various things such as movies and products, the document-level polarity classification of user reviews is rather important.

Another subtask considered in this study is the subjectivity classification. It is a task to classify whether a sentence is subjective or objective. The result of the subjectivity classification is often one of two classes: "subjective" or "objective". That is, the subjectivity classification is essentially a binary classification problem. The reason why subjectivity classification is required in the proposed method will be explained in the next section.

3.2 Motivation for subjectivity-oriented method

This section explains the motivation why the subjectivity is considered for the document-level polarity classification.

In user reviews, people often elaborate many sentences in a single review, and these sentences carry different aspects of content. A simple example of a review is shown below.

> It is so bad that the tickets are sold out! But the movie is really interesting!

It is relatively easy for humans to recognize that this is a positive review. But it is sometimes confusing for a computational model, because there are both positive word "interesting" and negative word "bad". On the one hand, the first sentence just describes the fact, so it can be regarded as an objective sentence. Since it does not express the user's sentiment, the negative word "bad" in it may be independent to the polarity of the overall review. On the other hand, the second sentence describes what the user thought after watching the movie, so it can be recognized as a subjective sentence. Thus, the positive word "interesting" may be consistent with the polarity of the review. If the subjectivity is not considered, a polarity classification model could not judge which the sentiment word, "bad" or "interesting", gives more influence on the document-level polarity. If a model takes the subjectivity of sentences into account, it can distinguish more important sentiment words in a review.

Table 3.1 shows another example of a movie review. There are 7 sentences in this review, and there are both subjective and objective sentences. The objective sentences just explain facts, while the subjective sentences express the user's emotion or opinion. On the other hand, each sentence sometimes expresses the user's sentiment or polarity, as indicated by "Positive" and "Negative" in the third column. When the polarity of the overall document (review) is determined, such polarity of sentences should be considered. However, intuitively, the sentiment of the subjective sentences seems more important and effective feature than that of the objective sentences.

Table 3.1: Example of a review with subjectivity and polarity of each sentence

Sentences in a review	Subjectivity	Polarity
1. A wonderful little production.	Subjective	Positive
2. The filming technique is very unassuming-very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece.	Objective	Neural
3. The actors are extremely well chosen-Michael Sheen not only "has got all the polarity" but he has all the voices down pat too!	Objective	Neural
4. You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth watching but it is a terrifically written and performed piece.	Subjective	Positive
5. A masterful production about one of the great masters of comedy and his life.	Objective	Positive
6. The realism really comes home with the little things: the fantasy of the guard which, rather than using the traditional 'dream' techniques remains solid and then disappears.	Objective	Neural
7. It plays on our knowledge and senses, particularly with the scenes concerning Orton and Halliwell, and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well made.	Objective	Neural

3.3 Polarity Classification by Subjectivity Weighted Voting

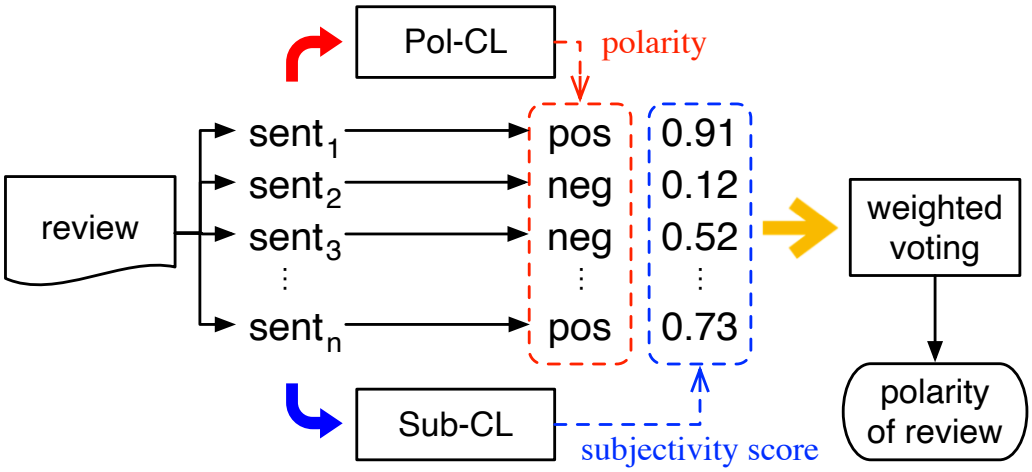

Figure 3.1: Overview of polarity classification by subjectivity weighted voting (PCSWV)

Our first method, named polarity classification by subjectivity weighted voting(PCSWV), determines the polarity on a given document by voting on the polarity of each sentence where the subjective sentences are more heavily considered. The overview of PCSWV is shown in Figure 3.1. It is implemented by the following steps.

1. We split a given review into sentences by using an ending symbol such as a period or exclamation mark as a delimiter.

2. The polarity of each sentence is identified by the polarity classifier, Pol-CL, such as "pos(positive)" and "negative(negative)" in Figure 3.1.

3. The subjectivity of each sentence is identified by the subjectivity classifier, Sub-CL. The subjectivity score that represents the intensity of the subjectivity is obtained, such as 0.91 and 0.12 in Figure 3.1.

4. We sum the subjectivity scores with the same polarity, and the polarity with the higher score is chosen as the polarity of this review.

The details of each step will be explained in the following subsections.

3.3.1 Data preprocessing

We use the NLTK(Natural Language Toolkit) library to preprocess the review to Tokenization, lemmatization, and removing stop words.

- Tokenization is the process to cut sentences or paragraphs into a sequence of tokens, which is defined according to the need of a need of users. Here, tokens can be either words, characters, or subwords. Word tokens are used in our thesis.

  For example, the sentence "I love this movie!" is converted to the following list of words by tokenization. Note that "movie!" are split into two words: "movie" and "!".

  > Sentence: I love this movie!
  > t = "I", "love", "this", "movie", "!"

- Lemmatisation is the process of removing prefixes and suffixes from words to get the root word. For example, words in a plural form, progressive tense, past participle are converted into lemma. So inflected words can be analyzed as a single item, and be identified by the word's lemma, or dictionary form [7].

  For example, the sentence "I am so surprised about the movies!" is converted to the following sequence of lemma by lemmatization.

  > Sentence: I am so surprised about the movies!
  > l = I be so surprise about the movie!

- Stop word is a word that does not has a distinct meaning and is not an effective feature for many NLP tasks. English words such as "the", "a", "an", "in", numbers, mathematical characters like "+", "×", and punctuation marks which are used particularly frequently, are typical stop words. Removing these words reduces the amount of the data, as well as usually improves the performance of the models. In our thesis, we use the "English stop words list" in NLTK library.

  For example, the sequence of the words "s" is obtained by removing the stop words in the sentence "I like watching movies! So I watch it." as follows.

  > Sentence: I like watching movies! So I watch it.
  > s: like watching movies watch

After those pre-processing, the review sentences become more concise and better for the polarity classification.

3.3.2 Polarity classification of sentence

This subsection explains how to classify the polarity of each sentence in a review. The polarity classifier, Pol-CL in Figure 3.1, is obtained by supervised machine learning. It is not the document-level but the sentence-level polarity classifier that identifies the polarity of a sentence in a review. Since the goal of this study is the document-level polarity classification, we suppose that a collection of documents (i.e. user reviews) labeled with the gold polarity is available. To train the sentence-level polarity classifier, we split a review into sentences, then automatically assign the same polarity label for each sentence as that of the review. That is, all sentences in a positive (or negative) review are classified as positive (or negative) sentences. Then, this dataset is used as the training data for the sentence-level polarity classifier.

Our study uses publicly available pre-trained models, BERT and XLNet, to train the polarity classifier. We use "bert-base-uncased"[3] for the BERT model, and use "xlnet-base-cased"[27] for the XLNet model. We use default hyperparameters for both BERT and XLNet models.

3.3.3 Subjectivity classification of sentence

The subjectivity classifier, Sub-CL in Figure 3.1, which judges whether an input sentence is subjective or objective, is also obtained by supervised machine learning. We suppose that a dataset of sentences annotated with the subjectivity labels ("subjective" or "objective") is available. Sub-CL is trained from such a subjectivity dataset. Similar to the polarity classification described in the previous subsection, publicly available pre-trained models, BERT and XLNet, are used to train the subjectivity classifier. Specifically, bert-bsae-uncased[3] is used as the pre-trained model of BERT, while xlnet-base-cased[27] is used as that of XLNet. Those pre-trained language models are fine-tuned using the subjectivity dataset. All hyperparameters are set as the default values.

Next, the probability of the classification predicted by the classifier is used as the subjectivity score of the sentence. The detail procedures are as follows.

1. Sentences in a review are classified by BERT or XLNet. As a result, we can get the output of the classification as well as two scores of each subjectivity class. Table 3.2 shows an example. "Index" shows an index of a sentence, indicating that the review consists of 6 sentences. If the objective score is greater than the subjective score, the sentence will be judged as objective, otherwise subjective.

Table 3.2: Example of subjectivity and objectivity scores obtained by BERT

Index	Prediction	Objective score	Subjective score
0	objective	1.7931842	−2.0638645
1	objective	2.124772	−1.8798144
2	objective	2.5113213	−2.9229922
3	subjective	−2.786232	3.1906624
4	objective	2.822895	−3.3431668
5	subjective	−2.5008123	2.8342867

2. The SoftMax function, as shown in Equation (3.1), is used to calculate the subjectivity score.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{3.1}$$

Where z_i are the elements of the input vector and can take any real value. K means the real numbers of values. The softmax function is used to normalize the subjectivity and objectivity scores so that they become a value between 0 and 1. Table 3.5 shows an example of calculation of the subjectivity score of the sentences shown in Table 3.4. "Nor. Obj. Score" and "Nor. Sub. Score" are the normalized objective and subjective scores obtained by softmax function, respectively. In this study, the subjectivity score of the sentence is defined as the normalized subjective score.

To sum up, the subjectivity score is a value between 0 and 1, where 1 means a completely subjective sentence and 0 means a completely objective sentence. Thus it expresses the intensity of the subjectivity of the sentence.

Table 3.3: Example of calculation of subjectivity score

index	Nor. Obj. Score	Nor. Sub. Score	Subjectivity score
0	0.9793	0.0207	0.0207
1	0.9821	0.0179	0.0179
2	0.9957	0.0043	0.0043
3	0.0025	0.9975	0.9975
4	0.9979	0.0021	0.0021
5	0.0048	0.9952	0.9952

3.3.4 Weighted Voting

Finally, the polarity of the overall review is determined by the weighted voting of the polarity classes of the sentences, where the weight is defined as the subjectivity score of each sentence.

The detailed algorithm of PCSWV is shown in Algorithm 1. The input of the model is a review R consisting of multiple sentences. For each sentence, we get the subjectivity score $Score_{sub}$ by the BERT or XLNet trained by the subjectivity dataset. Also, the polarity of each sentence, $Polarity$, is identified by the polarity classifier. For each polarity class (positive and negative), we add up the subjectivity scores of the sentences with the same polarity label, and store it as Sub_{pos} and Sub_{neg}. Finally, we compare them to determine the final polarity of the review.

Algorithm 1: Algorithm of PCSWV

Input: Review R
Output: Polarity of the review $Polarity(R)$.

Initialize $Sub_{pos} = 0, Sub_{neg} = 0$
for *each sentence s in R* **do**
    $Score_{sub} \leftarrow$ Sub-CL(s)
    $Polarity \leftarrow$ Pol-CL(s)
    if $Polarity = positive$ **then**
        $Sub_{pos} += Score_{sub}$;
    else
        $Sub_{neg} += Score_{sub}$;
if $Sub_{pos} > Sub_{neg}$ **then**
    $Polarity(R) \leftarrow positive$
else
    $Polarity(R) \leftarrow negative$
final ;
return $Polarity(R)$;

Table 3.4 shows an example of a review after the polarity and subjectivity classification of all sentences is performed. There are 3 sentences in this review, one is a negative sentence and two are positive sentences. Then, the subjectivity scores are summed for each polarity class: $Sub_{pos} = 0.77+0.95 = 1.72$ and $Sub_{neg} = 0.13$. Finally, the polarity of this review is judged as positive.

A simple way to classify the polarity of a document based on the re-

sults of the sentence-level polarity classification is to count the number of the positive and negative sentences and choose the most frequent polarity class without considering the subjectivity scores. However, we believe such a simple voting may often cause classification errors.

An illustrative example is shown in Figure 3.2. This example review, whose gold label is "positive", consists of five sentences. The simple voting wrongly classifies it as "negative", since the number of negative sentences is more than that of the positive sentences. On the other hand, our PCSWV classifies it as "positive", since the sum of the subjectivity scores for positive sentences is greater than that of negative sentences.

Table 3.4: Example of review after the polarity and subjectivity classificaton

Review	Polarity	Subj.Score
1. It is so a pity that I didn't buy the tickets.	Negative	0.13
2. The movie is so interesting!	Positive	0.77
3. I love this movie!	Positive	0.95

Note: *Subj.Score* means Subjectivity Score.

review	polarity	score
① This a fantastic movie about three prisoners who become famous.	positive	0.647
② One of the actors is George Clooney and I'm not a fan but this role is not bad.	negative	0.265
③ Another good thing about the movie is the soundtrack (The man of constant sorrow).	negative	0.454
④ I recommend this movie to everybody.	positive	0.782
⑤ Greetings Bart.	negative	0.544

Simple voting:
$Count_{pos} = 2 < Count_{neg} = 3 \Rightarrow$ negative

PCSWV:
$Score_{pos} = 1.429 > Score_{neg} = 1.263 \Rightarrow$ positive

Figure 3.2: Comparison between simple voting and PCSWV

3.4 Polarity Classification by Pre-trained Language Model with Subjectivity Filtering

Our second method, named Polarity Classification by Pre-trained Language Model with Subjectivity Filtering(PCPLM-SF), mainly relies on the pre-trained language model. We use two common language models: BERT and XLNet. In addition, we incorporate a filtering mechanism to use only subjective sentences for polarity classification.

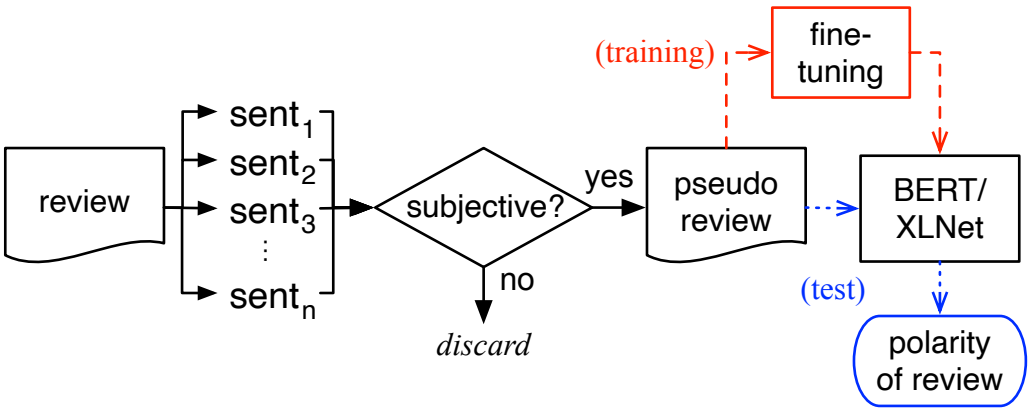

Figure 3.3: Overall of Overview of PCPLM-SF

The overview of PCPLM-SF is shown in Figure 3.3. First, we also split a review into sentences as in the previous model PCSWV. Next, we perform subjectivity classification for each sentence. If a sentence is classified as objective, it is filtered out, while a subjective sentence is kept. By concatenating the remained subjective sentences in the order of the original review, we obtain a pseudo review that consists of only subjective sentences. Finally, BERT or XLNet is fine-tuned using the training data of pseudo reviews. When the polarity of an unseen review is classified by the trained BERT or XLNet, a pseudo review is made by the same procedures and used as the input of the model.

The following subsections describe the details of the steps in PCPLM-SF.

3.4.1 Data preprocessing

Tokenization, lemmatization, and removing stop words are performed as preprocessing. It is the same as the one described in Subsection 3.3.1

3.4.2 Subjectivity filtering

We use the same method for subjectivity classification described in Subsection 3.3.3. Each sentence is classified whether it is subjective or objective. Then the objective sentences are discarded and only subjective sentences are retained. We call a set of remaining subjective sentences as "pseudo review".

Here is an example of the subjectivity filtering. Table 3.5 shows an example of the review, which is the same as one in Figure 3.2, and its gold polarity label. Table 3.6 shows the process of the subjectivity filtering for this review. The table (a) shows the sentences in the original review and the results of the subjectivity classification for each sentence. The table (b) shows a set of the subjective sentences after the filtering process. For your informaiton, the table (c) shows a set of the objective sentences only. In the experiment in Chapter 4, it is also used for comparison with the proposed method. Finally, the subjective sentences in Table 3.6 (b) are concatenated into the pseudo review as shown in Table 3.7.

Table 3.5: Example of original review

Original review: This is a fantastic movie about three prisoners who become famous. One of the actors is George Clooney and I'm not a fan but this role is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommend this movie to everybody. Greetings Bart **Sentiment: Positive**

3.4.3 Fine-tuning

After obtaining the set of pseudo reviews, the pre-trained language models are fine-tuned using it as the training data. Similar to the sentence-level polarity classifier described in Subsection 3.3.2 and the subjectivity classifier described in Section 3.3.3, BERT and XLNet are used as the pre-training model. The name of the pre-trained model and the setting of the hyperparameters are the same as described in these subsections.

Then, the pre-trained BERT or XLNet are fine-tuned using the set of pseudo reviews consisting of the subjective sentences only. The input of BERT and XLNet is supposed to be a single sentence. To classify a document (psuedo review) by BERT and XLNet, we simply input a pseudo review as a single sentence. To speed up the pre-training, the input of BERT and XLNet models is limited to 512 tokens. It is usually enough for most of the

Table 3.6: Example of subjectivity filtering

Sentence	Subjectivity
1. This a fantastic movie about three prisoners who become famous.	Subjective
2. One of the actors is George Clooney and I'm not a fan but this role is not bad.	Objective
3. Another good thing about the movie is the soundtrack (The man of constant sorrow).	Subjective
4. I recommend this movie to everybody.	Subjective
5. Greetings Bart	Objective

(a) Original review

Sentence	Subjectivity
1. This a fantastic movie about three prisoners who become famous.	Subjective
3. Another good thing about the movie is the soundtrack (The man of constant sorrow).	Subjective
4. I recommend this movie to everybody.	Subjective

(b) Review of subjective sentences only

Sentence	Subjectivity
2. One of the actors is George Clooney and I'm not a fan but this role is not bad.	Objective
5. Greetings Bart	Objective

(c) Review of objective sentences only

reviews but there are several reviews that are longer than 512 tokens. We just truncate them to the first 512 tokens.

3.4.4 Detail algorithm

As a summary, the pseudo code of PCPLM-SF is shown in Algorithm 2.

SubjectivityFiltering is the procedure to obtain a list of subjective sentences S from a given review R. For each sentence s in R, the subjectivity of it is identified (line 22). If it is classified as "subjective", it is appended to the end of the list S (line 24). Finally, the list S is returned as a pseudo review.

Training is the procedure to train the document-level polarity classifier M from a review dataset C. A new training data T is constructed by

Table 3.7: Example of pseudo review

Pseudo review: This is a fantastic movie about three prisoners who become famous. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommend this movie to everybody. **Sentiment: Positive**

converting each review R_i in C into $PseudoReview_i$ using the procedure $SubjectivityFiltering$ (line 6 and 7). Then, BERT or XLNet is fine-tuned with T to get the polarity classification model M (line 8).

$Test$ is the procedure to classify the polarity of an unseen review R. The same filtering procedure is used to convert R into the pseudo review (line 13), then its polarity is classified by M (line 14). As a variation of the proposed method, we use the original review, which includes both the subjective and objective sentences, as an input of the classification model M that is trained only on subjectivity sentences (line 15).

Algorithm 2: Algorithm of PCPLM-SF

Procedure: $Training(C)$
Input: review dataset C
Output: polarity classification model M
Initialize $T = \emptyset$
for *each review* R_i *in* C **do**
    $PseudoReview_i \leftarrow SubjectivityFiltering(R_i)$;
    $T \leftarrow T \cup \{PseudoReview_i\}$
$M \leftarrow Fine\text{-}Tuning(\text{BERT}, T)$ or $Fine\text{-}Tuning(\text{XLNet}, T)$
Return M

Procedure: $Test(R, M)$
Input: review R, polarity classification model M (BERT or XLNet)
Output: polarity label P
$PseudoReview \leftarrow SubjectivityFiltering(R)$
$P \leftarrow Classify(M, PseudoReview)$
$\langle\ P \leftarrow Classify(M, R)\ \rangle$
Return P

Procedure: $SubjectivityFiltering(R)$
Input: review R
Output: set of subjective sentences S_{subj}
Initialize $S = [\,]$; $O = [\,]$
for *each sentence* s *in* R **do**
    $y \leftarrow SentenceLevelSubjectivityClassifier(s)$
    if $y = subjective$ **then**
        $S \leftarrow append(s, S)$
    else
        $O \leftarrow append(s, O)$
Return S

Chapter 4

Evaluation

This chapter reports the detail of several experiments to evaluate our proposed methods. Section 4.1 describes three datasets used in our experiment. Section 4.2 outlines our evaluation criterion. Section 4.3 presents the result of subjectivity classification. Section 4.4 evaluates our first method PCSWV. Section 4.5 evaluates our second method PCPLM-SF and reports the results of our error analysis.

4.1 Dataset

Two datasets are used for the experiments. One is IMDb dataset, the other is Amazon review dataset. The details of them are introduced in the following subsections.

4.1.1 IMDb Dataset

IMDb Review Dataset [14] is a collection of 50,000 movie reviews. Table 4.1 shows the statistics of it including the total and unique number of the reviews and the sentiment labels. It also shows an excerpt of the example movie review and its annotated sentiment (polarity). Each review is annotated with binary polarity labels: positive or negative. In this experiment, 25K reviews are used as the training data, while the rest of 25K reviews are used as the test data.

Table 4.2 shows an example of the review in IMDb as well as the text after the pre-processing described in Subsection 3.3.1.

Table 4.1: Statistics and example of IMDb dataset

	review	sentiment
count	50000	2
unique	49582	2
example	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive

Table 4.2: Example of the original review and the preprocessed review

Original review	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me.The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word.It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away.I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.
After preprocessing	one review mention watch oz episod hook right exactli happen first thing struck oz brutal unflinch scene violenc set right word go trust show faint heart timid show pull punch regard drug sex violenc hardcor classic use word call oz nicknam given oswald maximum secur state penitentari focus mainli emerald citi experiment section prison cell glass front face inward privaci high agenda em citi home mani aryan muslim gangsta latino christian italian irish scuffl death stare dodgi deal shadi agreement never far away would say main appeal show due fact goe show dare forget pretti pictur paint mainstream audienc forget charm forget romanc oz mess around first episod ever saw struck nasti surreal say readi watch develop tast oz got accustom high level graphic violenc violenc injustic crook guard sold nickel inmat kill order get away well manner middle class inmat turn prison bitch due lack street skill prison experi watch oz may becom comfort uncomfort view that get touch darker side

4.1.2 Amazon Review Dataset

Amazon review dataset [15] is a collection of user reviews posted to the EC website Amazon. There are 35 million reviews for 18 years, up to March 2013. Each data includes a product name, user information, ratings, and a user review as plain text. The reviews with ratings 4 or 5 are used as positive reviews, while 1 or 2 as negative reviews. We use the same number of reviews as the IMDb dataset, i.e. we randomly choose 25K reviews as the training data and another 25K reviews as the test data.

Table 4.3 shows examples of reviews in Amazon review dataset including the polarity label, the title of the review, and the first several words of the review.

Table 4.3: Examples of reviews in Amazon review dataset[8]

label	**title**	**content**
1 (positive)	"Stuning even for the non-gamer"	"This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid...."
1 (positive)	"The best soundtrack ever to anything."	"I'm reading a lot of reviews saying that this is the best 'game soundtrack' and I figured that I'd write a review to disagree a bit. This in my opinino is Yasunori Mitsuda's ultimate masterpiece..."
1 (positive)	"Amazing!"	"This soundtrack is my favorite music of all time, hands down. The intense sadness of "Prisoners of Fate" (which means all the more if you've played the game) and the hope in "A Distant Promise" and "Girl who Stole the Star" have been an important inspiration to me ..."
1 (positive)	"Excellent Soundtrack"	"I truly like this soundtrack and I enjoy video game music. I have played this game and most of the music on here. I enjoy and it's truly relaxing and peaceful. On disk one..."
1 (positive)	"Remember, Pull Your Jaw Off The Floor After Hearing it"	"If you've played the game, you know how divine the music is! Every single song tells a story of the game, it's that good! ..."
1 (positive)	"an absolute masterpiece"	"I am quite sure any of you actually taking the time to read this have played the game at least once, and heard at least a few of the tracks here..."
0 (negative)	"Buyer beware"	"This is a self-published book, and if you want to know why–read a few paragraphs!"

4.1.3 Subjectivity datasets

Subjectivity datasets [1] is used to train the subjectivity classifier. It includes 5,000 subjective sentences excerpted from the movie review website (Rotten Tomatoes) and 5,000 objective sentences excerpted from IMDb plot summaries. This dataset assumes that all snippets from the Rotten Tomatoes pages are subjective, while all sentences from IMDb plot summaries are objective. Note that it is mostly true, but plot summaries can occasionally contain subjective sentences that are mislabeled as objective. That is, the subjectivity labels of the sentences and snippets were automatically assigned. Each line in the file of this dataset corresponds to a single sentence or snippet, and all sentences or snippets are downcased. Table 4.4 shows some examples of the subjectify dataset.

Table 4.4: Examples of reviews in subjectivity dataset

id	sentence	label
0	the movie begins in the past when a young boy named sam attempts to save celebs from a hunter.	subjective
1	emerging from the human psyche and showing characteristics of abstract expressionism, minimalism, and russian constructivism, graffiti removal has secured its place in the history of modern art while being created by artists who are unconscious of their artistic achievements.	subjective
2	spurning her mother's insistence that she get on with her life, mary is thrown out of the house, rejected by joe, and expelled from school as she grows larger with the child.	subjective
3	amitabh can't believe the board of directors and his mind is filled with revenge what better revenge than robbing the bank himself, ironic as it may sound?	subjective
4	she, among others excentricities, talks to a small rock, gertrude , as if she was alive.	subjective
5	this gives the girls a fair chance of pulling the wool over their eyes using their sexiness to poach any last vestige of common sense the dons might have had.	subjective

Note that the subjectivity classifier trained from this dataset is used for the experiment of the polarity classification using the aforementioned two polarity datasets. The domain of the IMDb dataset is the same as the subjectivity dataset (i.e. the movie review), but that of the Amazon dataset is different. It may degrade the classification performance on the Amazon dataset.

4.2 Evaluation Criterion

Accuracy, a measure of observational error, is used as evaluation Criterion [26]. Equation (4.1) shows the definition of the accuracy, where TP, TN, FP, and FN mean True Positive, True Negative, False Positive and False Nagative, respectively.

It is a commonly used criterion for classification tasks. The accuracy is used to evaluate how accurately the models can predict positive or negative reviews in the test set of datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

4.3 Result of Subjectivity Classification

First, we analyze the performance of subjectivity classification. The subjectivity dataset is randomly divided into 50% training data and 50% test data, then the accuracy of the trained classifiers on the test data is measured. Three classifiers are compared: Support Vector Machine (SVM) using bag-of-words features, BERT and XLNet. Table 4.5 shows the accuracy of these classifiers. It is found that XLNet performs the best and its accuracy is 0.96. In the rest of the experiments, this XLNet model is used as the sentence-level subjectivity classifier.

Table 4.5: Accuracy of subjectivity classification

Model	SVM	BERT	XLNet
Accuracy	0.76	0.94	**0.96**

4.4 Results of Subjectivity Weighted Voting

We evaluate the method using the subjectivity-weighted voting, PC-SWV, explained in Subsection 3.3. The following three methods of document-

level polarity classification are compared.

- **Sub. Only Voting:** the objective sentences are removed and the polarity is determined by voting of the polarity of the subjective sentences. It is a similar approach of previous work [15, 21] that filtered out objective sentences.
- **Simple Voting:** the polarity is determined by voting without considering the subjectivity scores.
- **Weighted Voting:** our proposed method PCSWV

Table 4.6 reveals the accuracy on the IMDb dataset when BERT and XLNet are used as the base sentence-level polarity classifier.

Table 4.6: Accuracy of polarity classification by subjectivity weighted voting and other baselines in IMDb dataset

Method	BERT	XLNet
Sub. Only Voting	0.749	0.667
Simple Voting	0.813	0.829
Weighted Voting(PCSWV)	0.816	**0.853**

"Sub. Only Voting" is obviously worse than the other methods, indicating that it is not good to totally ignore objective sentences. Our method outperforms the "Simple Voting" baseline for both BERT and XLNet, however, the improvement of the BERT model is rather small. On the other hand, significant improvement is found for XLNet, and our method with XLNet achieves the highest accuracy of 0.853.

We did not conduct the same experiment on the Amazon dataset, since the accuracy of PCSWV was much worse than our second method, PCPLM-SF, as will be reported in the next section.

4.5 Results of Language Model with Subjectivity Filtering

We evaluate the proposed method using the pre-trained language models with the subjectivity filtering explained in Subsection 3.4. In this method, only the subjective sentences are used for training the polarity classifier. For comparison, we also evaluate the method using all (both subjective and objective) sentences and only objective sentences in each training and test

data. Table 4.7 shows the accuracy of the IMDb and Amazon datasets. The system using S+O (subjective and objective sentences) as the training and test data is the baseline that simply applies BERT or XLNet without filtering ("Baseline" in Table 4.7), while the systems using S (subjective sentences) as the training data and S+O or S as the test data are our proposed systems ("PCPLM-SF-1" or "PCPLM-SF-2"). The best system among ones trained from the same training data is indicated in bold.

Table 4.7: Accuracy of polarity classification by language models

	Training	Test	IMDb		Amazon	
			BERT	XLNet	BERT	XLNet
(Baseline)	S+O	S+O	**0.997**	0.975	**0.939**	0.938
		S	0.749	0.701	0.920	0.928
		O	0.668	0.601	0.806	0.803
(PCPLM-SF-1)	S	S+O	0.886	0.819	**0.953**	0.924
(PCPLM-SF-1)		S	**0.980**	0.962	0.918	0.900
		O	0.663	0.616	0.800	0.764
	O	S+O	0.859	0.638	**0.924**	0.892
		S	0.743	0.669	0.894	0.868
		O	**0.963**	0.646	0.799	0.765

As for the IMDb dataset, BERT always achieves better accuracy than XLNet. When the settings (S+O, S, or O) of the training and test data are the same, the accuracy becomes the highest. It seems reasonable because the classifiers are fine-tuned using the training data obtained by the same filtering strategy as the test data. It is found that the accuracy is low when only objective sentences are used as the test data, except for BERT using O as the training data. It indicates that the subjective sentences are more informative for polarity classification. However, the baseline achieves the best accuracy, 0.997. Thus the filtering of the objective sentences is not effective in the IMDb dataset.

As for the Amazon dataset, BERT is slightly better than XLNet but they are almost comparable. Comparing the settings of the test data, the systems using subjective and objective sentences (S+O) are always the highest, following only S and only O. In the test data, the subjective sentences seem more effective than the objective sentences, but the latter also includes some useful information. The best system is PCPLM-SF-1, one of our proposed methods, where S is the training data and S+O is the test data. It indicates that the removal of the objective sentences from the training data is effective to improve the quality of the polarity classifier using BERT. Meanwhile, the

systems trained from only the objective sentences perform poorly.

The baseline is the best in the IMDb dataset, while our PCPLM-SF-1 is the best in the Amazon dataset. It may indicate that the objective sentences are less informative in the Amazon dataset than in the IMDb dataset, so the subjectivity filtering works well only in the Amazon dataset.

Finally, it is found that the accuracy of methods using the pre-trained language model (Table 4.7) is much better than that of the voting methods (shown in Table 4.6) on the IMDb dataset. Those results prove that the pre-trained language model is powerful and effective for the polarity classification as reported in many previous papers on various NLP tasks.

Error Analysis We carried out error analysis to investigate the major causes of errors of our proposed method PCPLM-SF. We found many cases that even the objective sentences carry some polarity information. So completely ignoring the objective sentences in our approach sometimes leads to inaccurate classification. Table 4.8 is an example of classification error, which means the PCPLM-SF method fails to classify the polarity of this review, while the baseline method, which uses the original reviews as the training data, can successfully classify it.

Table 4.8: Example of classification error

Review: Phil the Alien is one of those quirky films where the humor is based on the oddness of everything rather than actual punchlines. At first, it was very odd and pretty funny but as the movie progressed I didn't find the jokes or oddness funny anymore. It's a low-budget film (that's never a problem in itself), and there were some pretty interesting characters, but eventually, I just lost interest. I imagine this film would appeal to a stoner who is currently partaking. For something similar but better try "Brother from another planet" **Gold Label: Negative** **Baseline: Negative** **PCPLM-SF: Positive**

Table 4.8 shows individual sentences in this review, the predicted subjectivity labels, and flags whether the sentences are filtered out or not. In the retained subjective sentences after filtering out the objective sentences, the sentiment words such as "pretty", "funny", and "interesting" may indicate the positive sentiment of the user. Since our method PCPLM-SF only considered the subjective sentences, such the positive words strongly influenced the polarity classification, causing misclassification of this review as posi-

tive. On the other hand, the sentiment words or phrases such as "oddness", "punchlines", and "but better try" may carry the negative sentiment. The baseline model without the subjectivity filtering could consider those sentiment words and correctly classified the review as negative, but our method could not due to the removal of the objective sentences. To sum, it is found that the polarity information carried by contextual objective sentences is helpful to make a correct judgment on the polarity classification of a review.

Table 4.9: Subjectivity of sentences in misclassified review

Sentences	Subjectivity	Is filtered out
Phil the Alien is one of those quirky films where the humor is based on the oddness of everything rather than actual punchlines.	Objective	Yes
At first, it was very odd and pretty funny but as the movie progressed I didn't find the jokes or oddness funny anymore.	Subjective	No
It's a low-budget film (that's never a problem in itself), and there were some pretty interesting characters, but eventually, I just lost interest. I imagine this film would appeal to a stoner who is currently partaking.	Subjective	No
For something similar but better try "Brother from another planet".	Objective	Yes

Chapter 5

Conclusion

5.1 Summary

This thesis focused on the polarity classification of user reviews, where subjectivity was highly considered. Since objective sentences could not perform well in polarity classification, we considered subjective sentences more important in the review. This thesis proposed two methods of document-level polarity classification, PCSWV(Polarity Classification by Subjectivity Weighted Voting) and PCPLM-SF(Polarity Classification by Pre-trained Language Model with Subjectivity Filtering), which considered the subjectivity of the sentences.

PCSWV estimated the subjectivity score for each sentence in one review and determined the polarity of the review by the sum of the subjectivity scores for each polarity class. The advantage of this method was that all sentences in the review were considered and the polarity of the review was determined based on the subjectivity score obtained by the subjectivity classification. The disadvantage of this method was that we should still handle all sentences in the review, which might need much computational cost especially for a long review. Also, some noisy sentences that were independent of the polarity of an overall review might remain.

Therefore, we proposed the second method PCPLM-SF. In this model, we classified the subjectivity of the sentences in each review, filtered out the objective sentences and kept only the subjective sentences as the pseudo review. Then, a set of the pseudo reviews was used as the training data for fine-tuning of two pre-trained language models, BERT and XLNet. When an unseen review was classified, the pseudo review consisting of subjective sentences or the original review consisting of both subjective and objective sentences was fed into the fine-tuned BERT or XLNet. We supposed that

the above subjectivity filtering for construction of the training data could improve the polarity classification performance of BERT and XLNet. In addition, unlike PCSWV, PCPLM-SF could relatively easily handle long reviews, since the reviews were shortened by removing the objective sentences.

The results of our experiments showed that the subjective sentences had an important place in the polarity classification. Among the two methods proposed in this paper, PCPLM-SF achieved the best result on the Amazon dataset, reaching the accuracy of 95.3%. However, PCPLM-SF did not perform the best on the IMDb dataset, indicating that filtering out objective sentences did not have much effect on the IMDb dataset. However, the classifiers trained from only objective comments performed the worst in both the IMDb dataset and the Amazon dataset. Therefore, we believed that subjective sentences played a more important role in polarity classification than objective sentences.

5.2 Future work

Although our subjectivity-oriented approach improved the performance of polarity classification to some extent, filtering the objective sentences in the training data was not effective for the IMDb dataset against our expectation. In addition, through the error analysis, we found that objective sentences sometimes conveyed the sentiment of the user and could be effective features to classify the polarity of the overall review. Therefore, there is much room to improve our method. One of the difficult problems in polarity classification is irony. Since the literal and genuine meanings and polarity are different in ironic expressions, it is rather hard for a computational model, even for a human, to identify the polarity of it. In the future, we will explore a powerful subjectivity-oriented approach that can effectively classify not only ordinary reviews but also reviews including irony.

Bibliography

[1] Subjectivity datasets, (Accessed on Dec. 2022). https://www.cs.cornell.edu/people/pabo/movie-review-data/.

[2] Ravi Arunachalam and Sandipan Sarkar. The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP)*, pages 23–28, 2013.

[3] https://huggingface.co/bert-base-uncased, (accessed in Jan. 2023).

[4] R Bhaskaran, S Saravanan, M Kavitha, C Jeyalakshmi, Seifedine Kadry, Hafiz Tayyab Rauf, and Reem Alkhammash. Intelligent machine learning with metaheuristics based sentiment analysis and classification. *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, 44(1):235–247, 2023.

[5] Mohamed Reda Bouadjenek, Scott Sanner, and Ga Wu. A user-centric analysis of social media for stock market prediction. *ACM Trans. Web*, 2022.

[6] Xinying Chen, Peimin Cong, and Shuo Lv. A long-text classification method of Chinese news based on BERT and CNN. *IEEE Access*, 10:34046–34057, 2022.

[7] Collin. lemmatize. https://www.collinsdictionary.com/dictionary/english/lemmatize.

[8] https://huggingface.co/datasets/amazon_polarity, (accessed in Jan. 2023).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.

[11] Ahmad Kamal. Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources. *arXiv preprint arXiv:1312.6962*, 2013.

[12] Nancy Kansal, Lipika Goel, and Sonam Gupta. Cross-domain sentiment classification initiated with polarity detection task. *EAI Endorsed Transactions on Scalable Information Systems*, 8(30):e1–e1, 2021.

[13] M Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870, 2022.

[14] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011. `https://ai.stanford.edu/~amaas/data/sentiment/index.html`.

[15] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172, 2013.

[16] MS Neethu and R Rajasree. Sentiment analysis in twitter using machine learning techniques. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–5. IEEE, 2013.

[17] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

[18] Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. Aspect-level sentiment analysis using CNN over BERT-GCN. *IEEE Access*, 10:110402–110409, 2022.

[19] Marco Pota, Mirko Ventura, Rosario Catelli, and Massimo Esposito. An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian. *Sensors*, 21(1):133, 2020.

[20] Federico Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.

[21] Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *AAAI*, pages 1106–1111, 2005.

[22] C Sindhu, Binoy Sasmal, Rahul Gupta, and J Prathipa. Subjectivity detection for sentiment analysis on Twitter data. In *Artificial intelligence techniques for advanced computing applications*, pages 467–476. Springer, 2021.

[23] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[24] Milad Vazan, Fatemeh Sadat Masoumi, and Sepideh Saeedi Majd. A deep convolutional neural networks based multi-task ensemble model for aspect and polarity classification in Persian reviews, 2022. arXiv, 10.48550/arXiv.2201.06313.

[25] Niwan Wattanakitrungroj, Nichapat Pinpo, and Sasiporn Tongman. Sentiment polarity classification using minimal feature vectors and machine learning algorithms. In *The 12th International Conference on Advances in Information Technology*, pages 1–8, 2021.

[26] Wikipedia. Accuracy and precision. `https://en.wikipedia.org/wiki/Accuracy_and_precision`.

[27] `https://huggingface.co/xlnet-base-cased`, (accessed in Jan. 2023).

[28] ZhiZhan Xu, YiKui Liao, and SiQi Zhan. Xlnet parallel hybrid network sentiment analysis based on sentiment word augmentation. In *2022 7th International Conference on Computer and Communication Systems (ICCCS)*, pages 63–68. IEEE, 2022.

[29] Rongen Yan, Xue Jiang, and Depeng Dang. Named entity recognition by using XLNet-BiLSTM-CRF. *Neural Processing Letters*, 53(5):3339–3356, 2021.

[30] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[31] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89:1–8, 2011.

[32] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. Ss-bert: Mitigating identity terms bias in toxic comment classification by utilising the notion of" subjectivity" and" identity terms". *arXiv preprint arXiv:2109.02691*, 2021.