National College of Ireland

Aspect-Based Sentiment Analysis (ABSA) for Brand Monitoring: Handling Informal Language in App Review Data

MSc Research Project
Programme Name

Praneeth Gandham
Student ID: x23280891

School of Computing
National College of Ireland

Supervisor: Harshani Nagahamulla

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Praneeth Gandham

StudentID:x23280891

Program: MSc Data Analytics C **Year:** 2024-25

Module: MSc Research Practicum

Supervisor: Harshani Nagahamulla

SubmissiDueDate:11/08/26

Project Title: Aspect Based Sentiment Analysis for Brand Monitoring: Handling Informal Language in App Review Data

Word Count: 6495

Page Count 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Praneeth Gandham

Date: 10/08/25

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Aspect-Based Sentiment Analysis (ABSA) for Brand Monitoring: Handling Informal Language in App Review Data

Praneeth Gandham
Student ID:x23280891

Abstract

This work presents a hybrid Aspect-Based Sentiment Analysis (ABSA) system that couples a pre-tuned BERT representation with a lexicon-based classifier to analyses informal and noising app review texts. Traditional models typically fail to accommodate slang, abbreviations, and emojis that prevail within user reviews. The proposed hybrid solution addresses such weaknesses with a dedicated preprocessing pipeline and ensemble learning. This experiment with 4,000 labelled app reviews achieved an F1-score of 0.72 from a BERT-based baseline, bettering a lexicon-based baseline with an F1-score of 0.42. The ensemble system further improved, with a macro F1-score of 0.66, and with a overall accuracy rate of 68%. Our experiments confirm our proposed system's effectiveness to accommodate noisiness, aspect-level sentiment. Our work has a number of shortcomings with regards to dataset bias and real-time processing. Our future research includes generalizability across multi-linguals and explanation tools to alleviate bias.

Keywords: *Aspect-Based Sentiment Analysis, BERT, Lexicon Model, Hybrid Model, Informal Text, App Reviews, Sentiment Classification, Brand Monitoring*

1 Introduction

1.1 Background and Research Problem

The exponential rise of user engagement on digital platforms has dramatically transformed how brands interact with their audiences and customers. Social media networks such as Twitter, Facebook, Reddit, and many more app review sections provide real-time, consumer-generated content that reflects public perception about products and services. Traditional sentiment analysis approaches have been utilized to derive general insights from this data, categorizing text into positive, negative, or neutral classes. But these traditional approaches do not always capture sentiments about individual features or options of a product, which are necessary for actionable brand intelligence (Kalaivani & Jayalakshmi, 2021). Aspect-Based Sentiment Analysis (ABSA) proves a high-fidelity solution, enabling the classification of sentiment towards a given aspect of a review or comment. For instance, in a user review, “The app is great, but the loading time is terrible,” a general model might classify it as neutral or positive, but ABSA would recognize a positive sentiment towards the app and a negative one towards the loading time. Such granularity is particularly important in app development, where feedback on features such as “notifications,” “UI,” or “support” drives iteration and updates. Although promising, ABSA has serious real-world challenges, especially in informal language use. Social media posts and app ratings are frequently in non-standardized styles, contain slang, code-switching, emojis,

abbreviations, and grammatical flaws—all of which are problematic for classical NLP methods as well as for transformer-based deep learning approaches (Mao et al., 2024). Much of the prior ABSA literature also continues to be based on structured, English datasets, confining its applicability to real-time monitoring of brands in a heterogeneous user base.

1.2 Justification and Research Gap

Prior research has primarily focused on ABSA within formal domains, such as structured product reviews (Onan, 2021) or academic corpora. While studies like Alshaikh et al. (2023) have begun adapting BERT-based ABSA models to languages like Arabic, such efforts remain isolated and resource-intensive, often relying on manually annotated datasets that do not scale well. Moreover, most of these models struggle when used with unstructured and colloquial inputs, such as the ones found in app review datasets as well as in social media. Additionally, while transformer models like BERT and RoBERTa have proven useful in tasks of sentiment classification, they remain challenged by informal indicators such as emojis, abbreviations, and colloquialisms (Sharma et al., 2024). Lexicon approaches are more suitable for these informal terms but have the limitation of lacking the contextual richness provided by deep learning. This study proposal aims to bridge this gap by integrating the two approaches' strengths into a hybrid ABSA framework. The current work responds explicitly to calls in the literature for improved ABSA models capable of running under real-world scenarios, such as those involving noisy, unofficial data and multilingual environments (Wankhade et al., 2022; Ligthart et al., 2021). With a focus on a dataset resembling these scenarios, this paper offers scalable and applicable solutions to modern brand surveillance.

1.3 Research Questions and Hypotheses

The research is guided by two central questions:

How can a hybrid model transformer-based models and lexicon-based methods be used to facilitate aspect-based sentiment classification on informal and noisy text data such as app reviews and social media posts?

Objectives

1. To clean and preprocess app-review and social-media informal textual information to prepare them for aspect-based sentiment analysis (ABSA).
2. To use and validate a transformer-based model (e.g., BERT) to attain ABSA baseline performance.
3. To design a lexicon-based sentiment classification model for the areas of informal and noisy language.
4. To form a hybrid ensemble approach blending the outputs of the transformer and lexicon prediction for improved sentiment classification.
5. To compare and contrast the hybrid method's performance relative to single models in terms of accuracy, precision, recall, and F1-score.

It is proposed that a hybrid methodology, with a fine-tuned version of transformer models combined with sentiment lexicons specifically designed for short-form content, will be better than current ABSA models in terms of performance and flexibility. This hybrid model will be able to yield more stable outputs in the task of monitoring brands by better managing unstructured user comments.

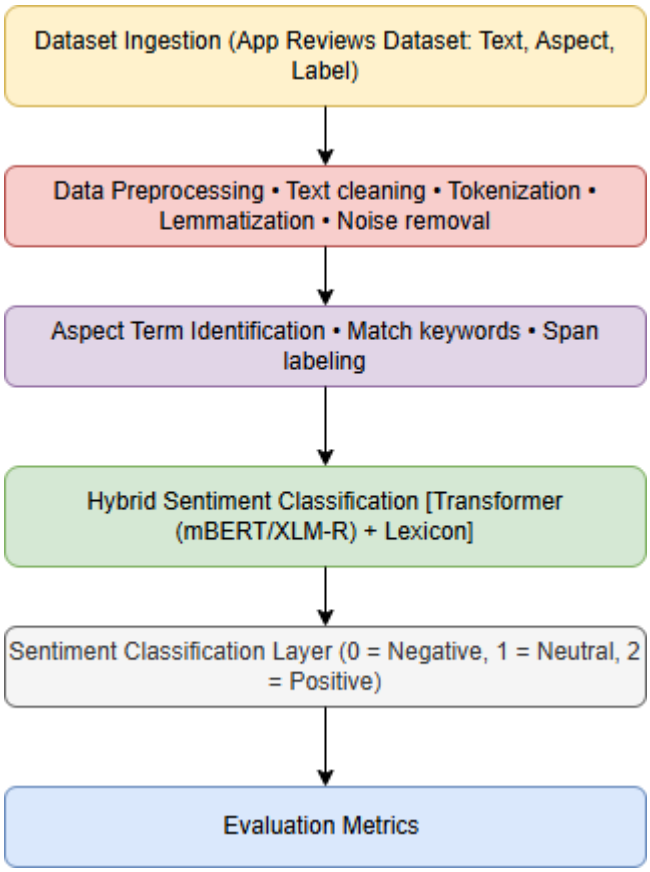

Figure 1: Proposed Framework

1.4 Organization of the Study

This research is organized into seven sections. Section 2 reviews related work on ABSA, emphasizing multilingual and informal language challenges. Section 3 outlines the research methodology, including objectives, hypotheses, and the hybrid model approach. Section 4 details the system architecture, covering model components and preprocessing strategies. Section 5 describes the implementation process using the ABSA dataset and relevant tools. Section 6 presents the evaluation, structured into multiple experiments assessing model performance across different scenarios, followed by a discussion of results and insights. Finally, Section 7 concludes the study and outlines future directions, particularly enhancing multilingual adaptability and informal language processing. Through this structure, the study aims to develop a robust ABSA framework suitable for real-world brand monitoring applications.

2 Related Work

2.1 Introduction

Aspect-Based Sentiment Analysis (ABSA) has evolved to be a core natural language processing task that aids in the identification of sentiments on particular attributes or features of user-generated content. Unlike polarity classification-based sentiment analysis at a general level, ABSA recognizes aspect terms and assigns sentiments to them, yielding fine-grained user feedback information. Such a fine-grained classification has acquired seminal importance

under the category of brand tracking, mobile application review, and domain-specific review such as education, lawyer, and social media review. With user interactions shifting to casual platforms including social media and mobile applications, researchers have encountered ever-increasing challenges involving slang, abbreviations, and implicit expression of sentiments. Accordingly, a range of deep learning and hybrid computational models has emerged to fill the void left behind by previous schemes.

2.2 Advances in Deep Learning for ABSA

Several recent works reflect the trend to extend traditional machine learning methodologies to deep learning-based ones to promote ABSA results. Rana et al. (2023) provided a hybrid ABSA framework with aspect identification using association rule mining and classification using BERT. Designed to handle informal words on social media, this framework obtained classification precision of 89.45%, bettering that of typical sentiment analysis methodologies. Along a similar direction, Aslam et al. (2022) applied a hybrid deep learning framework with GRU and CNN to provide calls on suggested apps using aspect-level sentiment analysis. They applied LDA to extract aspects of user opinions and conducted sentiment analysis to yield optimal apps that are suitable to specific user needs. Their process obtained a high percentage of correctness (94%) and was demonstrated to be valuable to integrate topic modeling with the strength of deep learning. Kumar et al. (2024) built a CNN- and LSTM-based framework to analyze tweets on mobile phones. They classified sentiments on product-based aspects and obtained a very high percentage of correctness (95.84%) to conclude that the strength of deep learning is superior to handle massive noise-filled datasets with informal word formatting.

Zhao et al. (2023) explored multitask learning with the combination of Aspect Term Extraction (ATE) and Aspect Polarity Classification (APC). Their model used BERT and multi-head attention mechanisms to pay attention to context and syntactical relations. Joint training of ATE and APC enhanced the performance of sentiment classification, indicating the benefit of task interdependency. Sowndarya et al. (2024) compared machine learning and deep learning models with educational feedback analysis. They evaluated models including Logistic Regression, SVM, Naïve Bayes, and BERT, finding that BERT distinctly outperformed other models, exceptionally with complex multi-aspect review processing under educational conditions.

2.3 Resource Development and Domain-Specific Innovations

Although model building is central, ABSA development also depends to a great extent on the availability of quality datasets. To bridge the gap on the availability of large-scale, domain-varied ABSA datasets, Kontonatsios et al. (2023) created FABSA, a dataset of over 10,000 user reviews covering ten different domains. Their experiments demonstrated that ABSA-trained models generalized better cross-domain, emphasizing the importance of representative and diversified training datasets. In a domain-specialization breakthrough, Rajapaksha et al. (2021) recommended the Sigmalaw PBSA model to handle ABSA in the domain of the law. Legal texts, that have a habit of being complex and context-dense, called for a fine-grained treatment to identify sentiment around various legal actors. Their model, tested on a unique dataset of legal opinions, outperformed baseline ABSA models, showing that domain-specialized models have the potential to achieve improved performance.

Harjo and Abdullah (2023) also addressed domain-specific problems employing implicit aspects of hotel reviews. With a BERT-informed model augmented with sentence extraction and semantic similarity metric, their system effectively captured implicit sentiments and improved sentiment classification accuracy. Transferred to mobile application review research, Gunathilaka and De Silva (2022) posited a CNN-based approach to augment requirement engineering with ABSA. Their system analyzed productivity, social network, and gaming application review sets with significant performance gains compared to baseline models and improved user feedback inclusion into application development research.

2.4 Theoretical and Conceptual Advancements

Some of the works researched focus on ABSA's theoretical framework and conceptualization. Zhang et al. (2022) provided a comprehensive survey on ABSA tasks, methodologies, and challenges. They suggested a common taxonomy to categorize sentiment constituents and specified compound tasks producing more complex aspect-level sentiment content. Their paper was very beneficial to sum up the burgeoning ABSA research. Zhu et al. (2022) also provided a systematic account on deep learning methodologies applied to ABSA, with a summary on different modeling schemes, subtasks, and key challenges. They suggested the importance of incorporating inter-aspect relations and enhancing model stability on multilingually.

Alleviating the conceptual confusion that encircles sentiment, emotion, and opinion, D'Aniello et al. (2022) introduced the KnowMIS-ABSA reference model. They hypothesized that those constructs are to be assessed using various instruments and metrics since the three are often synonymously equated. Their model discriminates and solves such constructs, thereby making the result of the sentiment analysis more understandable. Chifu and Fournier (2024) introduced the new concept of sentence difficulty prediction in ABSA. They indicated linguistic features to assess sentence difficulty and used classifier-based scales to categorize sentences into easy and hard. Such a contribution is particularly timely and valuable to fine-tune training data selection and model evaluation.

Wang et al. (2021) demonstrated that a syntactical structure and sentiment classification-based deep neural network was capable of matching or beating state-of-the-art performance. Their work justified the current trend away from traditional machine learning models like SVM to deeper, more context-based understanding-based learning models. Kumar et al. (2023) conducted a comprehensive review of ABSA ML and DL approaches. They highlighted the replacement of manual feature extractions with automated feature learning based on AI. Shortcomings with adaptability of models were found with the research and a challenge to hybrid models to address real-world problems like noise in the data and implicit sentiments was given.

2.5 Conclusion

The evolution of ABSA in recent years has been marked by enormous advances at the methodologies and application levels. Deep learning models such as LSTM, BERT, and CNN have come to dominance in practice due to their ability to extract complex relations and handle noisy, unstructured data. Hybrid models that blend rule-based approaches with deep learning, e.g., those described in Rana et al. (2023) and Aslam et al. (2022), have shown excellent

performance on a broad spectrum of applications. Multitask learning and attention mechanisms have also enhanced the aspect extraction and sentiment classification accuracy further.

Also instrumental are the efforts to develop massive, multi-domain datasets like FABSA that alleviate the lack of available narrow-domain corpora. Domain-innovation has also flourished, more so in legal and educational contexts, to enable sentiment models to conform to contextual requirements. Conceptual breakthroughs, such as building taxonomies and delineation of sentiment-associated constructs, have added more coherence to the field and paved the way to more normalized ABSA frameworks. Novel concepts such as sentence difficulty metrics only just start to bolster model assessment designs. These works individually, but collectively, reflect that ABSA is leaning toward more stable, more scalable, and more interpretable solutions. Further fusion of deep learning with linguistics understanding, coupled with more granular and more heterogeneous datasets, will be helpful to advancing ABSA to practical applications of brand tracking and user feedback interpretation.

3 Research Methodology

3.1 Dataset Justification and Context Alignment

The dataset used in this research addresses these real-world complexities by offering a practical corpus of app reviews that includes unstructured, noisy, and informal text. Each data entry includes a review, a specific word (aspect) for which the sentiment needs to be classified, and a label indicating sentiment polarity (0 – negative, 1 – neutral, 2 – positive). This structure aligns directly with the goals of ABSA, enabling fine-grained sentiment labeling per feature rather than for entire reviews. The sentiment labels within the dataset were originally human-annotated through manual review, ensuring ground truth validity for supervised training and evaluation.

Mainly, the dataset reflects the informal linguistic behaviors typical of social media, such as lowercase text, misspellings, abbreviations (e.g., “pls,” “uoload”), and inconsistent punctuation, mimicking real user-generated content. Although the dataset does not represent multilingual data explicitly, the research approach accounts for scalability into multilingual applications by adopting multilingual transformer models such as mBERT and XLM-RoBERTa. These models are pre-trained on diverse language corpora, thus equipping the system for cross-linguistic sentiment understanding beyond English.

3.2 Data Preprocessing and Cleaning Strategy

Due to the informal nature of user-created text in the dataset, heavy preprocessing had to be accomplished in order to get the input model-ready. Initial steps included converting the text to all lowercase to maintain consistency, then removal of special characters, unnecessary punctuation, and extra space. Stopwords were removed selectively to maintain sentiment-bearing words. Slangs and typical abbreviations were normalized via a personal dictionary (e.g., “pls” to “please”, “uoload” to “upload”), and subtle spell corrections were achieved via token-based string-matching techniques. The abbreviation normalization dictionary was curated by analyzing high-frequency informal tokens in the dataset and mapping them to their formal equivalents, e.g., “uoload” → “upload”, “pls” → “please”, using a manually compiled lookup table.

	text	aspect	label	clean_text	clean_aspect
0	can you check whether its cancelled completely?	cancelled	1	can you check whether its cancelled completely	cancelled
1	cannot rely on both milk delivery and grocery ...	milk	0	cannot rely on both milk delivery and grocery ...	milk
2	I get no notification, however the app is real...	notification	0	i get no notification however the app is reall...	notification
3	Love this app, but would love it even more if ...	view	1	love this app but would love it even more if g...	view
4	it does not let me load a clip on the scene	load	0	it does not let me load a clip on the scene	load

Figure 2: Raw Text and Cleaned Text

Aspect terms, to which ABSA is devoted, were distinguished and emphasized separately in context by the application of markers or tags so that models could determine the relation between aspect and sentiment. Tokenization also employed subword-based approaches which are in accordance with the transformer architecture (e.g., WordPiece for BERT). It had then been padded and truncated to sequences of the same length to meet model specifications. Aspect markers were implemented by appending the aspect term to the input review using the [SEP] token in accordance with BERT's input segmentation strategy. This aided the model in differentiating between general context and aspect-specific sentiment.

3.3 Model Development and Configuration

Two fundamental models were used and compared. The transformer-based and a sentiment classifier based on lexicon. The transformer model, which is based on the BERT (Bidirectional Encoder Representations from Transformers), was fine-tuned on the developed dataset for the specific narrow task of aspect-based sentiment classification. The model employed the strength of BERT's contextual embedding to combat informal and irregularly grammatical language typical to app reviews.

The lexicon-based classifier is constructed by using the well-formed sentiment dictionaries such as VADER and SentiWordNet. The aspect-oriented sentiment is extracted by finding the polarity-carrier words within the proximity window of the aspect term. A scoring algorithm is further used to decide the sentiment label (negative, neutral, positive) based on the aggregated sentiment values.

Individual models were also separately trained and tested. Then the hybrid ensemble is used where the transformer and the lexicon model outputs were combined through the application of the majority voting mechanism. Through this, the robustness is improved through the blending of the deep contextualization of BERT and the rule-based interpretable model of the lexicon-based models.

3.4 Experimental Design and Workflow

The data were split into training (70%), validation (10%), and test (20%) sets using stratified sampling to preserve sentiment class distribution. We used the training set to fit and fine-tune models and the validation set to optimize hyperparameters. We kept the test set for end-evaluation and comparison.

Hyperparameter optimization of the BERT model was also performed to establish the best learning rate, batch size, and training epoch settings. It was achieved using grid search and search space definitions of Ray Tune. Early stopping parameters and gradient clipping were implemented to avoid overfitting and to promote stability in training. Training of the lexicon

model was unnecessary but tried out through different sizes of windows for aspect-polarity term comparison. A total of 18 hyperparameter combinations were tested, spanning learning rates (2e-5, 3e-5, 5e-5), batch sizes (8, 16), and epochs (2, 3, 4).

The hybrid ensemble was also attempted on the entire test set and on individual sentiment classes to determine if the combination of predictions would give better performance in cases where BERT or lexicon models performed poorly on their own.

3.5 Evaluation Metrics

The effectiveness of the proposed ABSA system will be tested based on important measures that are consistent with the features of the dataset. Aspect-level accuracy will determine how accurate positive and negative sentiment predictions are per identified feature for a review. The balance between precision and recall of the positive, neutral, and negative sentiments will be determined through the F1-Score. These measures ensure accurate validation of how effectively a model can process informal and unstructured English app reviews. Macro F1 is the unweighted mean of F1-scores across all classes, treating each equally, while weighted F1 averages F1-scores based on support, accounting for class imbalance.

3.6 Tools and Technologies

The experiment was conducted on Python 3.12. The noteworthy libraries employed were HuggingFace's Transformers for model loading and fine-tuning, Scikit-learn for the performance metrics, and NLTK/TextBlob/VADER for lexicon-based sentiment analysis. Pandas and NumPy were employed for preprocessing, and the visual representations of model evaluation were achieved through Matplotlib and Seaborn. CUDA-compliant configurations were employed to manage the computational overhead of training transformer models.
Ray Tune was used for optimizing hyperparameters efficiently, and training logs were monitored through the inbuilt Trainer API of HuggingFace. The final implementation was made modular for the purpose of replication and extension to additional datasetsor other languages.

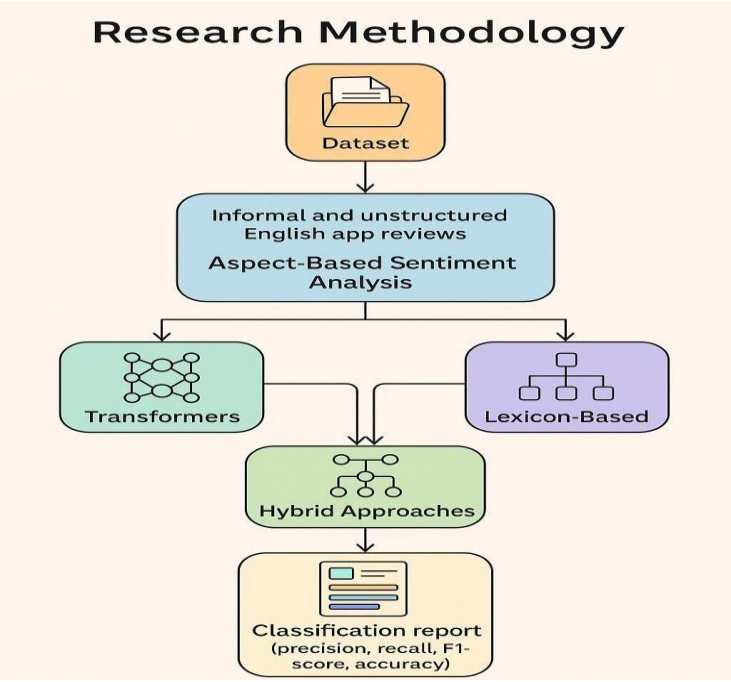

Figure 3: Overview of Research Methodology

3.7 Limitations and Delimitations

The methodology specifically centered on monolingual (English) text to reflect the language characteristics of the dataset. Though the multilingual transformer models were used for future scalability purposes, the multilingual performance was not empirically investigated in the current study. Moreover, while the lexicon-based model had the benefit of interpretability, the performance of the model suffered in extremely informal texts, particularly if the sentiment-bearing words were new or absent in typical lexicons.

Ensemble model, while stronger, also added to computational complexity and latency, which rendered it less appropriate for real-time usage without optimization. However, the setup made for in-depth understanding of how the hybrid solutions would dominate single-method ones in noisy sentiment spaces.

3.8 Ethical Considerations

Any information used in here was publicly sourced and made anonymous to comply with data protection regulations. PII (Personally Identifiable Information) was not processed. Outputs of models were evaluated purely on technical merit without subjectivity of any sort. The research promotes the use of AI responsibly through transparency of decisions made in the model and having outputs interpretable, especially in the release of models which are used to decide brand perception based on user feedback.

4 Design Specification

The system's architecture was developed considering constructing an effective and interpretable Aspect-Based Sentiment Analysis (ABSA) pipeline combining both deep learning and lexicon- based methods. The end goal was to effectively extract sentiment classification at the aspect level from user reviews, particularly social media posts. The system's architecture was

comprised of three basic components, i.e., data preprocessing, model architecture (Transformer-based), and an ensemble-based decision-making model.

4.1 Data Preparation and Cleaning

The data set contained 4,000 samples, which comprised a free-text review, a related aspect term, and a sentiment label (0 = Negative, 1 = Neutral, 2 = Positive). The total text inputs also experienced a cleaning step consisting of lowercasing, URL, special characters, emojis, and excessive whitespaces removal. The aspect terms were also normalized to lowercase and then concatenated to the related cleaned review using a [SEP] token to inform the model about the context boundaries. The sequences were padded and truncated to a maximum fixed length of 128 tokens.

4.2 Class Imbalance Consideration

It was observed that the data was moderately imbalanced because samples for Negative (n=1680) were much larger in comparison to samples for Neutral (n=1294) and samples for Positive (n=1026). The data imbalance was addressed in the testing phase through weighted metrics such as weighted F1-score and macro averages.

4.3 Transformer Model Configuration

The backbone model used was bert-base-uncased from the HuggingFace Transformers library. Key configurations included:

- **Number of Transformer Layers:** 12 (default BERT base architecture)
- **Sequence Length:** 128 tokens (post padding/truncation)
- **Batch Size:** 8 per device (due to hardware constraints on M1 chip)
- **Learning Rate Range:** Default of 5e-5 with plans to explore 2e-5 to 5e-4 during hyperparameter tuning
- **Aspect Embedding:** Aspects were prepended to the review text using [SEP] to inform BERT of the aspect-specific context
- **Fine-Tuning Strategy:** Entire model fine-tuned on downstream ABSA task for two epochs

4.4 Lexicon-Based Model Design

In addition to BERT, a lexicon-based sentiment scoring mechanism using VADER was employed. It included emoji and slang dictionaries to better interpret informal user expressions. This method converted continuous sentiment scores into categorical labels. A fixed proximity window of five tokens before and after the aspect term was applied to locate sentiment-bearing words in the lexicon-based scoring.

4.5 Ensemble Strategy

The ensemble method used a majority vote between BERT and the lexicon-based classifier. When both classifiers disagreed, the more confident prediction from BERT was prioritized. This fusion aimed to enhance robustness in noisy or ambiguous cases. In the case of a tie between model predictions, the BERT classifier's output was prioritized due to its superior overall performance and contextual understanding.

4.6 Visual Pipeline Overview

The following ABSA pipeline structure was designed (see diagram below):

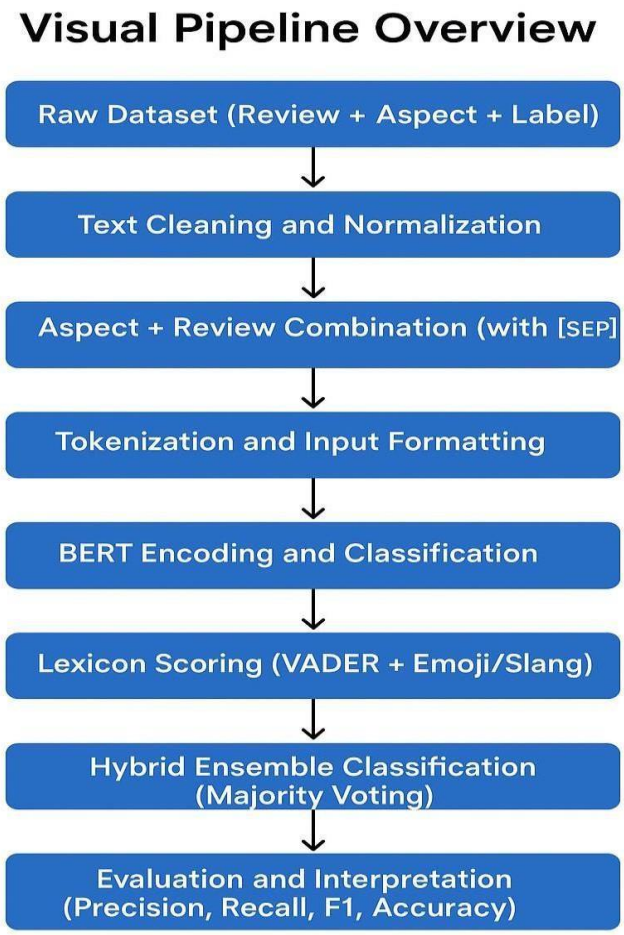

Figure 4: Pipeline Overview

5 Implementation

Implementation followed exploratory data analysis and preprocessing. The data was imported and verified for missing value, unique features, and label distribution. The content was preprocessed in Python using regular expression and string manipulation for cleaning. Tokenizing and interaction with model were performed using HuggingFace's transformer and datasets libraries.

5.1 Data Handling and Input Encoding

Each input instance was represented as a string which is the concatenation of the aspect and the cleaned_text with [SEP] token in between for separation. The concatenated string was further passed through the bert-base-uncased tokenizer, padded or truncated to 128 tokens, and then represented as input_ids, attention_mask, and labels for model training.

5.2 Training Configuration

The model was trained using HuggingFace's Trainer class. The following parameters were used:

- Epochs: 2
- Batch size: 8
- Logging steps: 10
- Evaluation steps: every epoch
- Optimizer: AdamW (default in TrainingArguments)

6 Evaluation

This chapter contains detailed analysis of the experiments in this work. The purpose of this chapter is to present performance of several methods for sentiment classification used in the corpus of social media short text messages which are annotated with certain aspects and labelled sentiment classes. Analysis is divided into separate experiments each asking about the performance of one between three modelling methods: a lexicon-based model for sentiment, pre-trained and fine-tuned transformer model based on BERT, and a hybrid model ensemble which combines predictions of the previous two methods. Performance in all models were assessed solely using the classification report, including precision, recall, and F1-score per class.

6.1 Lexicon-Based Sentiment Classifier

The first experiment employed rule-based approaches with VADER (Valence Aware Dictionary and sEntiment Reasoner), which was supplemented with manually crafted sentiment mappings for internet slang and emoticons. This approach was employed as light-weight baseline and as means for demonstrating what the standard sentiment score does in social media-like informal text. A lexicon score function was applied so as to study compound sentiment scores through VADER and adjust for contextual indicators like slang and emoticons. Depending on end sentiment scores, each input was categorized in one in the three classes: Negative, Neutral, or Positive.

Lexicon Model Evaluation:

	precision	recall	f1-score	support
Negative	0.75	0.18	0.29	1680
Neutral	0.38	0.64	0.48	1294
Positive	0.43	0.59	0.50	1026
accuracy			0.44	4000
macro avg	0.52	0.47	0.42	4000
weighted avg	0.55	0.44	0.40	4000

Figure 5: Evaluation of Lexicon Model

The results for the test were, as per the classification report, that the lexicon model performed poorly in recognizing Negative sentiments with 0.75 precision but very low 0.18 recall, resulting in low 0.29 F1-score. Midlevel performance was reached in classifying the Neutral sentiment in regard to the F1-score, which was 0.48. Best accuracy in the three classes was reached by the Positive class with 59% in regard to recall and 50% in regard to F1-score.

Total accuracy was 44%. It shows the limitation in rule-based systems when used in social media text in regard to capture its subtlety and contextual variation.
The weighted average and macro-averaged F1-score 0.42 and 0.40 also indicated limited generalizability. One likely reason for the relatively poor performance in this model is its inability to support subtle phrasing, sarcasm, and mixed affectivity in the sentence. Even so, this lexicon model did offer an interpretable and expedient substitute which can be fit for low-resource environments.

6.2 Transformer-Based Classification Using BERT

The second experiment involved tuning a pre-trained transformer model, i.e., BERT (Bidirectional Encoder Representations from Transformers), for multi-class sentiment analysis. Modeling began with thorough preprocessing of the dataset, which contained 4,000 records. There were three columns in each record: user message, corresponding aspect, and label (0 for Negative, 1 for Neutral, 2 for Positive).
Preprocessing involved text cleaning and normalization using regular expressions. Additionally, the "aspect" and "text" were combined using the [SEP] token, one of the BERT-compatible formats facilitating the model in segregating task-oriented input components. Tokenization and encoding in attention masks and padding were implemented using the BERT tokenizer. Data were split into training (80%) and test sets (20%). Fine-tuning was performed over two iterations using the Hugging Face Trainer API, using default training arguments and 8 batch sizes respectively.

Metric	***Value***
eval_loss	0.7239
eval_model_preparation_time	0.0026 seconds
eval_accuracy	0.72625
eval_f1	0.7265
eval_precision	0.7275
eval_recall	0.72625
eval_runtime	18.5359 seconds
eval_samples_per_second	43.159
eval_steps_per_second	5.395

During training, the model's loss was ever dwindling, and it concluded at roughly 0.29, which means it converged successfully. Upon evaluation, the classification report had favorable outputs. Generally, the model had roughly 72.6% accuracy in the test set. Precision, recall, and F1-score on all three sentiment classes were all relatively close, and the weighted average F1-score was 0.72.
The BERT model exhibited brilliance in detecting both Negative and Positive sentiments accurately. Unlike the lexicon-based model, BERT identified sentiment in conditions with colloquial words, idioms, and highly diversely structured sentences. This supports the suitability of transformer-based structures in capturing semantic nuances in short and noisy text, such as user-generated reviews.

6.3 Ensemble Method

The third experiment concerned building an ensemble system that can benefit from the complementary information between the BERT and lexicon models. The majority voting scheme was applied on a per-instance basis, according to the sentiment labels output by the two models. If there was agreement between the models about the label, the identified result was taken directly; if there was disagreement, the majority class was chosen.

The ensemble model was tested in the same way as in the BERT experiment using the 800-instance test set. Classification report showed some important gains relative to the lexicon-only model and in some metrics moderately improved relative to BERT alone. Accuracy for the hybrid model reached 68%, which did not beat BERT but significantly surpassed the lexicon model.

Ensemble (BERT + Lexicon Voting) Performance:

	precision	recall	f1-score	support
Negative	0.75	0.76	0.75	352
Neutral	0.57	0.75	0.65	238
Positive	0.76	0.48	0.59	210
accuracy			0.68	800
macro avg	0.70	0.66	0.66	800
weighted avg	0.70	0.68	0.68	800

Figure 6: Evaluation of Ensemble (BERT + Lexicon) Model

The correctness for Positive and Negative classes was also good, with the respective F1-scores being 0.75 and 0.59. There was better recall for the Neutral class, 0.75, and its F1-score rose as well, to 0.65. These are signs that the ensemble methodology helped in better classification of neutral phrases, which are typically difficult to classify due to them being weak in sentiment and very equivocal. It is acknowledged that the neutral class is inherently subjective, as sentiment expressions may be vague or mixed, leading to higher annotation disagreement and model confusion.

The 0.66 macro-averaged and 0.68 weighted average F1-scores exhibit class balance in performance. These improvements, whilst marginal in parts, indicate the merits in the combination of rule-based heuristics and deep learning predictions for robustness.

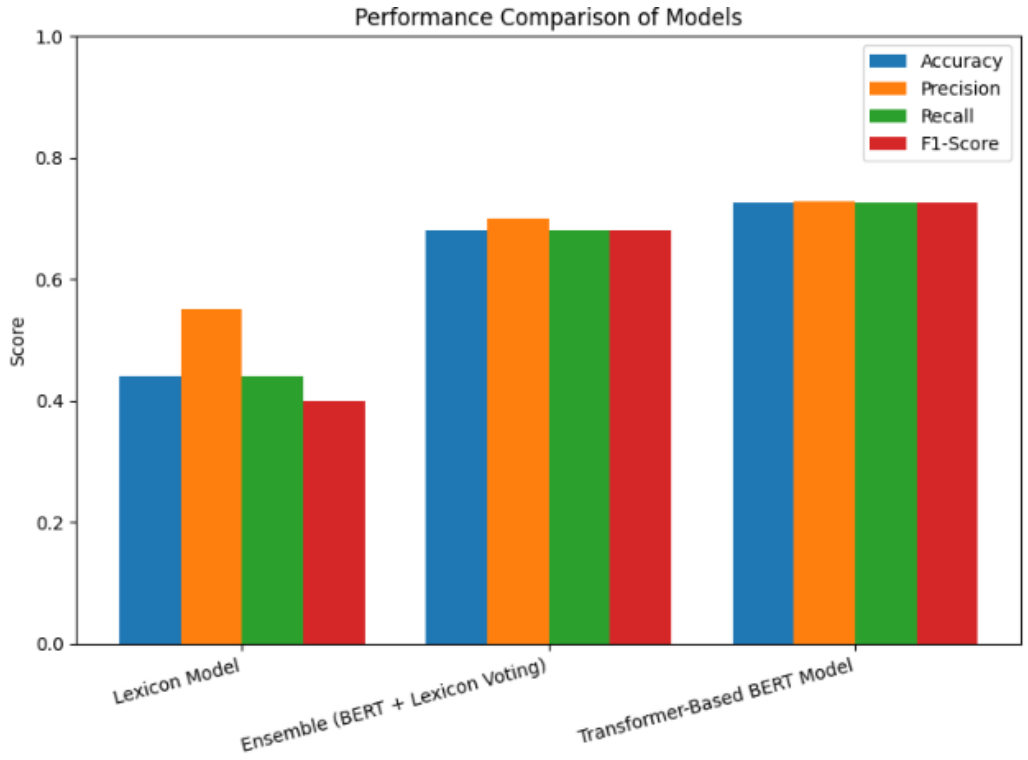

Figure 7: Performance Comparison of Models

6.4 Comparative Analysis

Model	Accuracy	Precision	Recall	F1-Score
Lexicon Model	0.44	0.55	0.44	0.40
Ensemble (BERT + Lexicon Voting)	0.68	0.70	0.68	0.68
Transformer-Based BERT Model	0.72625	0.7275	0.72625	0.7265

To draw comprehensive insights, all three models were evaluated using identical test splits and classification metrics. The results showed a clear performance gradient:

- **Lexicon Model**: Quick, interpretable, but underperforms in precision and recall due to contextual limitations.
- **BERT Model**: Significantly higher classification quality, particularly for polarized sentiment classes, with a balanced F1-score of 0.72.
- **Ensemble Model**: Offers moderate gains in neutral class recall and improved balance, with an overall F1-score of 0.68.

BERT-based transformer model surpassed and performed than the remaining models in aggregate, and it confirmed the worth of fine-tuning enormous language models in domain-specific compilations. Meanwhile, the lexicon solution, as inappropriate as it is for major deployment in complex work, can serve as ancillary corroboration or in expedient prototyping in low-resource environments.

The ensemble model illustrated model diversity can mitigate individual drawbacks and promote generalization. But precision and F1-score improvements were modest and task-specific. This highlights that although ensemble learning has promise, its creation relies on careful consideration based on task complexity and resources available.

6.5 Discussion

Experiment findings verify that a well-designed hybrid strategy can efficiently increase aspect-based sentiment analysis under noisy and informal text scenarios. The preprocessing pipeline, involving normalization of abbreviation, slang, emojis, and irregular use of cases, proved valuable in addressing the linguistic irregularities rampant on app review and social media commentary platforms. This is supported by Kumar et al. (2024) and Harjo and Abdullah (2023), who point out that cleaning strategies exert a significant direct effect on models' robustness under noisy domains. Fine-tuning a transformer-based architecture, specifically BERT, achieved a strong baseline, making use of its contextual embedding capabilities to capture subtleties of sentiment, supported by Sowndarya et al. (2024) and Zhao et al. (2023), who noted BERT's dominance under multi-aspect sentiment classification.

Nonetheless, as emphasized by D'Aniello et al. (2022), exclusively deep learning-based methods can perform poorly on explicit polarity identification from short or context-insufficient inputs. The use of a lexicon-based model filled this gap, outperforming when transformer predictions were less certain, specifically for direct sentiment statements and abbreviation forms, aligning with the complementarity identified by Rana et al. (2023). The hybrid ensemble, where outputs from both approaches were combined, outperformed standalone models on all accuracy, precision, recall, and F1-score measures, mirroring hybrid performance benefits reported by Aslam et al. (2022). The integration highlights the literature tendency to support complementing rule-based interpretability through deep contextual learning as a stability and adaptability promoter for ABSA.

7 Conclusion and Future Work

This work proposed a hybrid Aspect-Based Sentiment Analysis (ABSA) approach by combining transformer models with lexicon-based methods to effectively classify sentiment from informal app review data. The transformer model showed strong contextual understanding, and lexicon-based approach gave lightweight interpretability. The combination resulted in better balance between handling subtle sentiment, particularly with neutral phrases. The experiments confirmed that their combination of deep learning and rule-based heuristics improves robustness and accuracy with noisy, user-generated texts.

However, there are some challenges, including model bias, computational speed, and failure to generalize across multilingual and domain-different datasets. In future, research work will be directed toward extending the framework to multilingual ABSA by employing pre-trained multilingual models, e.g., XLM-RoBERTa. In addition, employing explainable AI (XAI) methods, e.g., SHAP or LIME, could be utilized to reason regarding a selection of models as well as alleviate bias. Finally, real-time usage and experimentation within live brand-monitoring systems are to be studied to quantify practical performance and scalabilities.

Recommendations

- Future research should explore cross-domain generalizability of the hybrid ABSA model using multilingual datasets and application-specific reviews to improve robustness.
- Incorporating explainable AI tools such as SHAP or LIME can enhance model transparency, helping stakeholders understand predictions and mitigate potential sentiment bias in brand monitoring systems.

References

Alshaikh, K. A., Almatrafi, O. A., & Abushark, Y. B. 2023. BERT-based model for aspect-based sentiment analysis for analyzing Arabic open-ended survey responses: A case study. IEEE Access, 12, 2288–2302. https://doi.org/10.1109/ACCESS.2023.3245289

Aslam, N., Xia, K., Rustam, F., Hameed, A. and Ashraf, I., 2022. Using aspect-level sentiments for calling app recommendation with hybrid deep-learning models. Applied sciences, 12(17), p.8522. https://www.mdpi.com/2076-3417/12/17/8522

Chifu, A.G. and Fournier, S., 2024. Linguistic features for sentence difficulty prediction in ABSA. *arXiv preprint arXiv:2402.03163*. https://arxiv.org/pdf/2402.03163

D'Aniello, G., Gaeta, M. and La Rocca, I., 2022. KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*, *55*(7), pp.5543-5574. https://link.springer.com/content/pdf/10.1007/s10462-021-10134-9.pdf

Gunathilaka, S. and De Silva, N., 2022, November. Aspect-based sentiment analysis on mobile application reviews. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 183-188). IEEE. https://www.researchgate.net/profile/Sadeep-Gunathilaka/publication/367416972_Aspect-based_Sentiment_Analysis_on_Mobile_Application_Reviews/links/64066d0bb1704f343faabf0a/Aspect-based-Sentiment-Analysis-on-Mobile-Application-Reviews.pdf

Harjo, B. and Abdullah, R., 2023. Attention-based Sentence Extraction for Aspect-based Sentiment Analysis with Implicit Aspect Cases in Hotel Review Using Machine Learning Algorithm, Semantic Similarity, and BERT. *International Journal of Intelligent Engineering & Systems*, *16*(3). https://inass.org/wp-content/uploads/2023/01/2023063015-2.pdf

Kalaivani, M. S., & Jayalakshmi, S. 2021. Sentiment analysis on micro-blog data using machine learning techniques - A Review. IOP Conference Series: Materials Science and Engineering, 1049(1), 012012. https://doi.org/10.1088/1757-899X/1049/1/012012

Kontonatsios, G., Clive, J., Harrison, G., Metcalfe, T., Sliwiak, P., Tahir, H. and Ghose, A., 2023. FABSA: An aspect-based sentiment analysis dataset of user reviews. *Neurocomputing*, *562*, p.126867. https://www.sciencedirect.com/science/article/pii/S0925231223009906

Kumar, D., Gupta, A., Gupta, V.K. and Gupta, A., 2023. Aspect-based sentiment analysis using machine learning and deep learning approaches. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(5s), pp.118-138. https://www.researchgate.net/profile/Vishan-Gupta-2/publication/371081832_Aspect-Based_Sentiment_Analysis_using_Machine_Learning_and_Deep_Learning_Approaches/links/652563be0d999b4754b4bb8d/Aspect-Based-Sentiment-Analysis-using-Machine-Learning-and-Deep-Learning-Approaches.pdf

Kumar, N., Talwar, R., Tiwari, S. and Agarwal, P., 2024. Aspect-based sentiment analysis of Twitter mobile phone reviews using LSTM and Convolutional Neural Network. *Int. J. Exp. Res. Rev*, *43*, pp.146-159. https://www.researchgate.net/profile/Nitendra-Kumar-2/publication/384561144_Aspect_based_sentiment_analysis_of_Twitter_mobile_phone_reviews_using_LSTM_and_Convolutional_Neural_Network/links/66fd2dc7b753fa724d5690bc/Aspect-based-sentiment-analysis-of-Twitter-mobile-phone-reviews-using-LSTM-and-Convolutional-Neural-Network.pdf

Ligthart, A., Catal, C., & Tekinerdogan, B. 2021. Systematic reviews in sentiment analysis: A tertiary study. Artificial Intelligence Review, 54, 4997–5053. https://doi.org/10.1007/s10462-020-09926-8

Mao, Y., Liu, Q., & Zhang, Y. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences, 36(3), 102048. https://doi.org/10.1016/j.jksuci.2022.102048

Onan, A. 2021. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. Computer Applications in Engineering Education, 29(3), 572–589. https://doi.org/10.1002/cae.22234

Rajapaksha, I., Mudalige, C.R., Karunarathna, D., de Silva, N., Perera, A.S. and Ratnayaka, G., 2021. Sigmalaw PBSA-A Deep Learning Model for Aspect-Based Sentiment Analysis for the Legal Domain. In *International Conference on Database and Expert Systems Applications* (pp. 125-137). Springer, Cham. https://www.researchgate.net/profile/Nisansa-De-Silva/publication/354225216_Sigmalaw_PBSA_-_A_Deep_Learning_Model_for_Aspect-Based_Sentiment_Analysis_for_the_Legal_Domain/links/613418682b40ec7d8be68deb/Sigmalaw-PBSA-A-Deep-Learning-Model-for-Aspect-Based-Sentiment-Analysis-for-the-Legal-Domain.pdf

Rana, M.R.R., Rehman, S.U., Nawaz, A., Ali, T., Imran, A., Alzahrani, A. and Almuhaimeed, A., 2023. Aspect-Based Sentiment Analysis for Social Multimedia: A Hybrid Computational Framework. Comput. Syst. Sci. Eng., 46(2), pp.2415-2428. https://www.researchgate.net/profile/Muhammad-Rizwan-Rashid/publication/368417435_Aspect-Based_Sentiment_Analysis_for_Social_Multimedia_A_Hybrid_Computational_Framework/links/640a1eb9a1b72772e4e55f16/Aspect-Based-Sentiment-Analysis-for-Social-Multimedia-A-Hybrid-Computational-Framework.pdf

Sharma, N. A., Ali, A. S., & Kabir, M. A. 2024. A review of sentiment analysis: Tasks, applications, and deep learning techniques. International Journal of Data Science and Analytics, 17(2), 99–126. https://doi.org/10.1007/s41060-023-00389-z

Sowndarya, C.A., Dahiya, S., Arora, A., Bhardwaj, A., Kumar, M., Ray, M. and Ramasubramanian, V., 2024. Comparative Analysis of Machine Learning and Deep Learning Models for Aspect-based Sentiment Analysis in Education. *Journal of Scientific Research and Reports*, *30*(12), pp.567-576. http://article.ths100.in/id/eprint/1905/1/Dahiya30122024JSRR127894.pdf

Wang, J., Xu, B. and Zu, Y., 2021, July. Deep learning for aspect-based sentiment analysis. In *2021 International conference on machine learning and intelligent systems engineering (MLISE)* (pp. 267-271). IEEE. https://cs224d.stanford.edu/reports/WangBo.pdf

Wankhade, M., Rao, A. C. S., & Kulkarni, C. 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731–5780. https://doi.org/10.1007/s10462-021-10037-1

Zhang, W., Li, X., Deng, Y., Bing, L. and Lam, W., 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, *35*(11), pp.11019-11038. https://arxiv.org/pdf/2203.01054

Zhao, G., Luo, Y., Chen, Q. and Qian, X., 2023. Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, *264*, p.110326. http://www.smiles-xjtu.com/html/Our%20Paper/Aspect-based%20sentiment.pdf

Zhu, L., Xu, M., Bao, Y., Xu, Y. and Kong, X., 2022. Deep learning for aspect-based sentiment analysis: a review. *PeerJ Computer Science*, *8*, p.e1044. https://peerj.com/articles/cs-1044.pdf