electronics

MDPI

Article

Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models

Abhilash Pathak [1], Sudhanshu Kumar [1,*], Partha Pratim Roy [1] and Byung-Gyu Kim [2,*]

[1] Department of Computer Science & Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India; a_pathak@cs.iitr.ac.in (A.P.); partha@cs.iitr.ac.in (P.P.R.)
[2] Department of IT Engineering, Sookmyung Women's University, Seoul 04310, Korea
* Correspondence: skumar2@cs.iitr.ac.in (S.K.); bg.kim@sookmyung.ac.kr (B.-G.K.); Tel.: +82-2-2077-7293 (B.-G.K.)

Abstract: Sentiment Analysis is becoming an essential task for academics, as well as for commercial companies. However, most current approaches only identify the overall polarity of a sentence, instead of the polarity of each aspect mentioned in the sentence. Aspect-Based Sentiment Analysis (ABSA) identifies the aspects within the given sentence, and the sentiment that was expressed for each aspect. Recently, the use of pre-trained models such as BERT has achieved state-of-the-art results in the field of natural language processing. In this paper, we propose two ensemble models based on multilingual-BERT, namely, mBERT-E-MV and *mBERT-E-AS*. Using different methods, we construct an auxiliary sentence from this aspect and convert the ABSA problem to a sentence-pair classification task. We then fine-tune different pre-trained BERT models and ensemble them for a final prediction based on the proposed model; we achieve new, state-of-the-art results for datasets belonging to different domains in the Hindi language.

Keywords: aspect-based sentiment analysis; BERT; classification; ensemble; Hindi

Citation: Pathak, A.; Kumar, S.; Roy, P.P.; Kim, B.-G. Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models. *Electronics* **2021**, *10*, 2641. https://doi.org/10.3390/electronics10212641

Academic Editor: Stefano Ferilli

Received: 19 August 2021
Accepted: 21 October 2021
Published: 28 October 2021

1. Introduction

Sentiment analysis (SA) [1–4] is one of the most critical tasks in the field of natural language processing (NLP). It is the process of analyzing and summarizing users' opinions and emotions as expressed in a sentence. In simple words, in SA, we determine whether the underlying sentiment in a piece of text is positive, negative, or neutral. This has gained attention in academia and business, particularly when identifying customer satisfaction with products and services, offering valuable feedback on websites such as Flipkart, Amazon, etc., through online reviews.

In the last decade, the popularity of e-commerce websites among consumers has increased tremendously. Users are sharing their experiences online regarding the products and services that they have used. There is a steep increase in the number of online reviews that are being posted daily. These opinions and feedback act as a measure for the goodness of the products and services. Reading all these reviews is time-consuming and analyzing them is practically challenging. Therefore, there is a need for automation to effectively maintain and analyze these reviews. Reviews are used for decision-making by the organizations as well as the consumers. Consumers decide or confirm which products to buy based on the reviews, while organizations tend to improve or develop new products, plan marketing strategies and campaigns, etc.

Early studies in this field were centered only on detecting the overall polarity of a sentence, irrespective of the entities to which they referred (e.g., *mobiles*) and their aspects (*camera*, *display*, etc.). The basic assumption behind this task is that there is a single overall polarity for the whole review sentence. The sentence, however, can include various aspects, e.g., *"This mobile comes with 6.53-inch AMOLED display which is pretty good but the 16MP camera disappoints"*. The polarity of the aspect *'display'* is *Positive*, while the polarity

of the aspect *'camera'* is *Negative*. The goal of Aspect-Based Sentiment Analysis (ABSA) is to consider fine-grained polarity against a particular aspect. This task is especially useful because a consumer may analyze the aggregated sentiment for each aspect of a given product and obtain a more in-depth understanding of its quality.

The ABSA task is analogous to learning the *Aspect Category Detection* subtask and the *Aspect Category Polarity* subtask at the same time.

Aspect Category Detection: This is a multi-label classification problem. In this task, we are provided a set of sentences and a pre-defined set of aspect categories, e.g., (*FOOD*, *PRICE*, *etc.*). The objective is to detect all the aspect categories that are discussed in each sentence. Typically, the aspect terms do not appear in the sentences as words. For instance, in the example, *"Delicious but expensive"*, the adjectives, *'delicious'* and *'expensive'* implies the presence of aspect categories *FOOD* and *PRICE* in the sentence.

Aspect Category Polarity: It is a multi-class classification problem. In this task, given the aspect categories for each review sentence, the objective is to decide the sentiment polarity (*positive*, *negative*, *neutral*, or *conflict*) of each aspect category present in the sentence. The *'conflict'* polarity is assigned to an aspect when it has both the *'positive'* and the *'negative'* sentiments associated with it. For example, in the sentence, *"The pizza was good but the burger was tasteless"*, the aspect *Food* has the *conflict* polarity.

An example of the ABSA task is presented in Table 1. There are two aspects present in the sentence: one is *Food*, and the other is *Service*. The polarity of aspect *Food* is *negative*, while the polarity of *Service* is *positive*.

Table 1. An example for the ABSA task. The first subtask is to identify the two aspects present in the sentence: one is *Food*; the other is *Service*. The second subtask is to predict the polarity of both the aspects. The polarity of *Food* is *negative*, while the polarity of *Service* is *positive*.

Sentence: I liked the service and the staff, but the pizza was below par.	
Aspect Category	Sentiment Polarity
Food	Negative
Service	Positive

Recently, pre-trained language models, such as ELMo [5], OpenAI GPT [6], and BERT [7], have demonstrated their efficacy in solving many natural language-processing problems. BERT has performed exceptionally well on Question Answering (QA), and Natural Language Inference (NLI) tasks [7], both of which are sentence-pair classification tasks. However, direct application of the BERT model does not result in significant improvements in the ABSA task. The authors in [8] assumed that this was due to the unsuitable application of the BERT model.

Before the introduction of Transformers [9] in 2017, language models mainly used RNNs and CNNs to perform NLP tasks. The Transformer is a significant improvement as it doesn't need text to be processed in any predetermined order. Also, Transformers allow training on a massive amount of data in very little time. They are the basis for models like BERT.

Word embedding models like GloVe [10], and word2vec [11] map each word to a vector that tries to represent some aspects of the word's meaning. Word embeddings are useful for many NLP tasks, but some limitations prevent them from being used. There is a limitation to what these word models can capture, as they are not trained on deep modeling tasks, so they cannot effectively represent the negation of words and word combinations. Another significant flaw is that these models ignore the context of the words. For example, the word *'bank'* has different meanings in the sentences *"He opened a new account in the bank"* and *"A dead body was found on the bank of the river"*. However, embedding methods will assign the same vector for the word *'bank'* in both sentences, so a single vector is forced to capture both meanings.

The above drawbacks were motivated by context-based language models that train a neural network to assign a vector to one word, based on either the surrounding context or the entire sentence. For example, the sentence, *"He opened a new account in the bank"*, represents *'account'* based on the word's context. A unidirectional model represents *'account'* based on *"He opened a new"* but not *"in the bank"*. However, a bidirectional contextual model represents *'account'* using the context—*"He opened a new…in the bank"*.

OpenAI GPT, ELMo, and BERT are examples of transfer-learning-based models. OpenAI GPT and ELMo were previous state-of-the-art contextual pre-training methods. OpenAI GPT is unidirectional, based on a unidirectional Transformer. ELMo is shallowly bidirectional. Two LSTMs are trained independently: one is left-to-right, and the other is right-to-left. Then, the learned embeddings are concatenated to generate the features that are used in downstream tasks. Only BERT is deeply bidirectional. In BERT, representations are learned based on both left and right contexts. ELMo is a feature-based approach, while the other two are fine-tuning approaches.

In this paper, we propose two ensemble models based on multilingual BERT, namely, *mBERT-E-MV* and *mBERT-E-AS*. As mBERT can take a single sentence or a pair of sentences as input, we transform ABSA task into a sentence-pair classification task by constructing an auxiliary sentence using aspect. Then, we fine-tune different pre-trained mBERT models for each auxiliary sentence construction method, based on the newly generated task. Finally, we ensemble the models using majority voting and average score for final prediction, and achieve state-of-the-art results on datasets belonging to different domains in the Hindi language.

The main contributions of the paper are as follows:

1. To the best of our knowledge, this is the first time the transfer-learning-based method has been used for aspect-based sentiment analysis in the Indian language;
2. The proposed methodology can be treated as a baseline for solving further problems involving Indian languages.

The rest of the paper is organised as follows. Section 2 summarizes the relevant works in the field of aspect-based sentiment analysis. Section 3 discusses the methodology of the proposed framework. Section 4 presents the datasets used in the experiments and the experimental results. The paper concludes with the derived conclusions and the scope for the future presented in Section 5.

2. Related Work

In prior works for ABSA, methods related to machine learning were dominant [12,13]. They were primarily concentrated on the extraction of hand-crafted lexical and semantic features [14]. The authors [15] proposed sentiment-specific word embedding. Such feature-engineering-based studies require professional-level knowledge in linguistics and have limitations regarding the achievement of the best possible performance. An SVM-based model was proposed in [16], which used word-aspect-association lexicons for sentiment classification. The authors [17] proposed a multi-kernel approach for aspect category detection. Previous aspect-based techniques did not appropriately adapt general lexicons in the context of aspect-based datasets, resulting in a reduced performance. The authors in [18] presented extensions of two lexicon generation methods to handle this problem: one using a genetic algorithm and the other using statistical methods. They combined the generated lexicons with the well-known static lexicons to categorize these aspects into reviews.

Neural networks can dynamically extract features without feature engineering. They can transform the original features into continuous, low-dimensional vectors because of this ability; they have been gaining huge popularity in ABSA. The sentences and aspects were independently modeled using two separate LSTM models in [19]. Then, pooling operation was performed to measure the attention given to the sentences and aspects. In recent years, with the increased use of attention mechanisms in deep learning models, many researchers have incorporated them into RNNs [20–22], CNNs [23], and memory

networks [22,24]. This enables the model to learn various attention distribution levels for different aspects, as well as create attention-based embeddings. The authors [22] proposed the use of delayed, context-aware updates with a memory network. Context-aware embeddings were generated using interaction-based embedding layers in [25]. To handle the complications and increase the expressive power of LSTM, several attention layers were used with LSTM in [20,26]. In [21], Attention-Based LSTM with Aspect Embedding (ATAE-LSTM) was proposed, which focused on identifying the sentiment-carrying words that were relatively correlated with the entity or target.

Most recently, the authors have used transfer-learning-based models. BERT has been used in various papers [27,28] to produce contextualized embeddings for input sentences, which were subsequently used to identify the sentiment for target-aspect pairs. The authors in [29,30] used BERT as the embedding layer, while the authors in [31] used a fine-tuning approach for BERT, with an additional layer acting as the classification layer. BERT was fine-tuned for **Targeted Aspect-Based Sentiment Analysis (TABSA)** in recent works [32,33] by altering the top-most classification layer to include the targets and aspects. Instead of utilizing the top-most classification layer of BERT, the authors in [34] investigated the possibility of using the semantic knowledge present in BERT's intermediate layers to improve BERTs fine-tuning performance. The authors [8] proposed the construction of sentences from the target–aspect pairs, before feeding them to BERT to fully utilize the power of BERT models. However, BERT's input format is limited to a sequence of words that cannot provide more contextual information. To overcome this issue, authors in [35] introduced GBCN, a new method that enhances and controls the BERT representation for ABSA by combining a gating mechanism with context-aware aspect embeddings. The input texts are first fed into the BERT and context-aware embedding layers, resulting in independent BERT representations and refined context-aware embeddings. The most associated information chosen in this context is contained in these refined embeddings. The flow of sentiment information between these context-aware vectors and the output of the BERT encoding layer is then dynamically controlled by employing gating units.

However, these works are mainly carried out in the English language. For Indian languages, most of the existing works aim to classify the sentiments at either the sentence- or at document level. ABSA in Indian languages is still an open challenge, as minimal resources are available, and hardly any significant work has been performed in this field in Indian languages. The authors [36] used different models such as Decision Tree, Naive Bayes, and a sequential minimal optimization implementation of SVM (SMO) to solve the ABSA problem in the Hindi language. They used lexical features such as n-grams, non-contiguous n-grams, and character n-grams, together with a PoS tag and semantic orientation (SO) score [37], for polarity classification.

The author [38] showed the relationship between affective computing and sentiment analysis. The primary tasks of affective computing and sentiment analysis are emotion recognition and polarity detection. They can enhance the customer relationship management and recommendation system abilities, for example, to reveal which features customers enjoy or should be excluded from a recommendations system that received negative feedback. In [39], the authors showed a range of the current approaches and tools for multilingual sentiment analysis. In addition to the challenge of understanding the formal textual content, it is also essential to consider the informal language, which is often coupled with localized slang, to express 'true' feelings.

The authors proposed BabelSenticNet [40], the first multilingual concept-level knowledge base for sentiment analysis. The system was tested on 40 languages, proving the method's robustness and its potential for utility in future research.

The authors [41] proposed an attention-based bidirectional CNN-RNN deep model for sentiment analysis (ABCDM). The effectiveness of ABCDM is evaluated on five reviews and three Twitter datasets. It showed that ABCDM achieves state-of-the-art results for both long-review and short-tweet polarity classification.

In [42], the authors proposed a multi-task ensemble [43] framework of three deep learning models (i.e., CNN, LSTM, and GRU) and a hand-crafted feature representation for the predictions. The experimental results suggest that the proposed multi-task framework outperformed the single-task frameworks in all experiments.

3. Proposed System

We propose two ensemble models, namely, mBERT-Ensemble-Majority Vote (*mBERT-E-MV*) and mBERT-Ensemble-Average Score (*mBERT-E-AS*). Figure 1 shows the workflow for both the ensemble methods. At first, auxiliary sentences are constructed from the aspect information using four auxiliary sentence construction methods, namely Natural Language Inference—Multi (NLI-M), Question Answering—Multi (QA-M), Natural Language Inference—Binary (NLI-B), and Question Answering—Binary (QA-B), which are discussed in the following subsection. Then, the constructed auxiliary sentence and the input review sentence are fed to the WordPiece tokenizer, which breaks the two sentences into a stream of tokens. Segment embeddings and position embeddings are then added to the token embeddings. Then, the generated token sequences are fed to four different mBERT models for fine-tuning. After fine-tuning, each mBERT model predicts the output label and softmax scores for every label. The final step is to aggregate the four predictions and make a final prediction. The two ensemble models are similar to each other except for this aggregation step. The *mBERT-E-MV* model makes the final label prediction based on the majority of the output labels. In contrast, the *mBERT-E-AS* model averages the output softmax scores of the four mBERT models and outputs the label's highest softmax score.

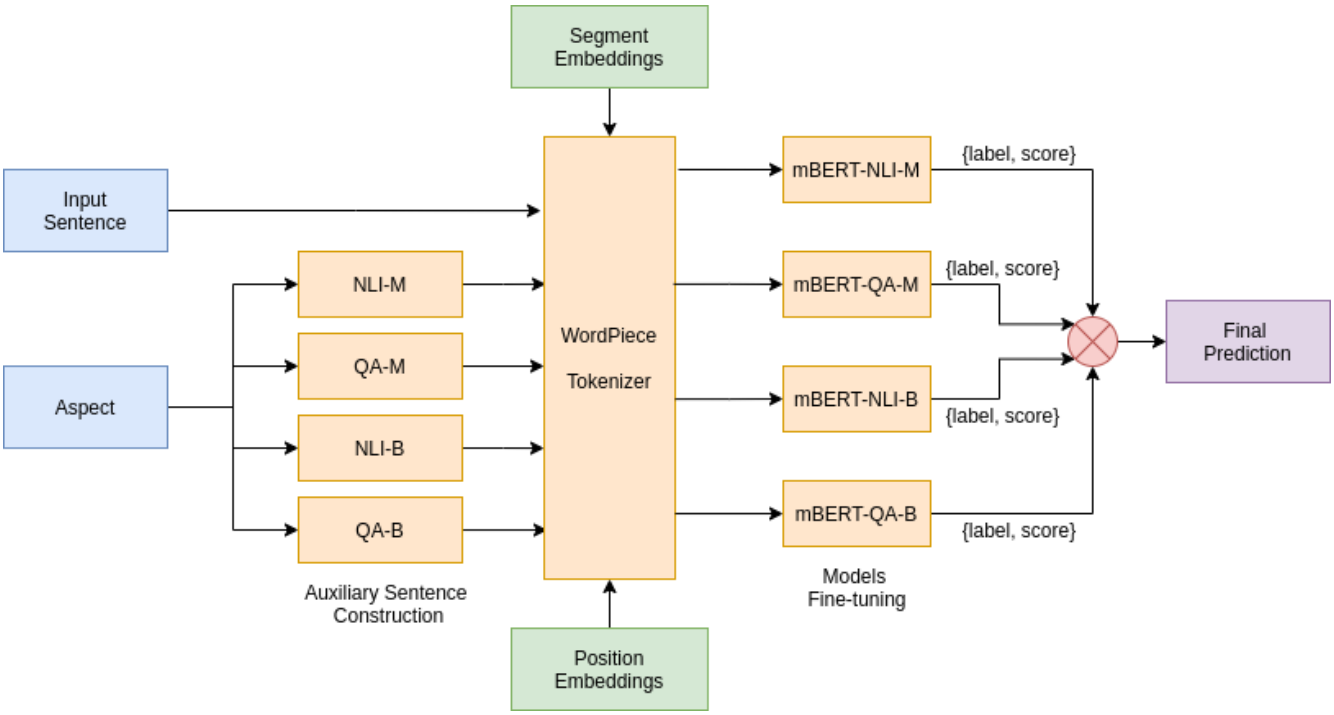

Figure 1. Workflow for *mBERT-E-MV* and *mBERT-E-AS* ensemble models. At first, auxiliary sentences are constructed from the aspect information using four auxiliary sentence construction methods, namely NLI-M, QA-M, NLI-B, and QA-B. Then, the constructed auxiliary sentence and the input review sentence are fed to the WordPiece tokenizer. The final step is to aggregate the four predictions and make a final prediction.

3.1. Auxiliary Sentence Construction

To obtain better results from the mBERT model, we transform the ABSA task into a sentence-pair classification task. Apart from the original input text review, we need to add an auxiliary sentence for each input sentence. We use the following four methods, proposed in [8] to construct an auxiliary sentence.

3.1.1. QA-M

In this method, we generate the sentence as a question using the aspect. The format of the question should be the same for all the auxiliary sentences. For example, if the aspect pair is *price*, then the format of the question generated can be *"what do you think of the price ?"*.

3.1.2. NLI-M

In this method, we do not generate a full standard sentence but a simple pseudo-sentence. The format for this is much simpler. For example, if the aspect is *price*, the auxiliary sentence is: *"price"*.

3.1.3. QA-B

In this method, we also use the label information while creating the auxiliary sentence Therefore, the ABSA problem is temporarily transformed into a binary classification problem *yes*, *no*. For each aspect, three sequences have to be generated. For example, suppose the aspect is *price*, then the three sequences are *"the polarity of the aspect price is positive"*, *"the polarity of the aspect price is negative"*, *"the polarity of the aspect price is none"*. The class of the sequence for which we obtain the highest probability value of *'yes'* is chosen as the predicted category.

3.1.4. NLI-B

This method is similar to QA-B, with the only difference being that we generate pseudo-sentence instead of a standard sentence. For example, if the aspect is *price*, the auxiliary sentences are: *"price—positive"*, *"price—negative"*, and *"price—none"*.

All four methods are summarized in Table 2.

Table 2. The form of auxiliary sentences generated using the auxiliary sentence construction methods and the expected output labels for the newly generated sentence-pair classification task.

Method	Auxiliary Sentence	Output
QA-M	Question without sentiment polarity	Sentiment Polarity
NLI-M	Pseudo-sentence without sentiment polarity	Sentiment Polarity
QA-B	Question with sentiment polarity	{yes, no}
NLI-B	Pseudo-sentence with sentiment polarity	{yes, no}

3.2. Fine-Tuning Pre-Trained mBERT

Since mBERT is already pre-trained on a large corpus, we can now use it to fine-tune the ABSA task. Next, we discuss the input data representation and the process of fine-tuning.

3.2.1. Input Representation

We feed the original review and the constructed auxiliary sentence to the WordPiece embedder that converts the two sentences into a sequence of tokens. A unique classification token ([CLS]) is present at the first position of each sequence, which is used for the classification task. Two separating tokens ([SEP]) are also added: one after the tokens corresponding to the original review, and another after those corresponding to the auxiliary sentence. The first ([SEP]) token acts as the separator for the two sentences, and the second ([SEP]) token signifies the end of the token sequence.

In mBERT, two sentences are fed at a time into the model. So there is a need for segment embeddings that tells the mBERT model how to distinguish between the two inputs in a given pair. Suppose the two sentences are *"My dog is cute"* and *"He likes playing"*. This layer has only two vector representations. All tokens that belong to the first sentence are assigned to the first vector, while all tokens that belong to the second sentence are assigned to the second vector.

mBERT is based on Transformers, which do not encode the sequential information of the input [9]. For input text, *"I do, what I think"*, both *'I'*s should have different vector representations. That is why positional embeddings are required. mBERT learns a vector representation for each position during training. Every word present at the same position has the same position embedding. Hence, for input texts like *"Good job"* and *"Well done"*, both *'Good'* and *'Well'* have same position embeddings. Similarly, both 'job' and 'done' will have the same representation.

The relevant token, segment, and position embeddings are added together to form the input representation for a given token.

3.2.2. Fine-Tuning Procedure

Fine-tuning mBERT is pretty straightforward. The final hidden state corresponding to the ([CLS]) token is considered as the fixed-dimensional pooled representation of the input sequence. We denote this vector as $C \in \mathbb{R}^H$, where H is the size of the hidden state. This is fed to a classification layer with parameter matrix $W \in \mathbb{R}^{K \times H}$, where K denotes the number of labels. Finally, the softmax function $P = \text{softmax}(CW^{\mathrm{T}})$, is used to calculate the probability for each label.

3.3. Ensembling mBERT Models

After fine-tuning, each mBERT model predicts the output label and softmax scores for every label. The final step is to aggregate the four predictions and make a final prediction. The aggregation process uses two methods: one is based on majority voting, the other is based on average scoring. The *mBERT-E-MV* model makes the final label prediction based on the majority of the output labels, while the *mBERT-E-AS* model averages the output softmax scores of the four mBERT models and outputs the label with the highest softmax score.

4. Experiments and Discussion

4.1. Dataset Description

For the aspect category detection and sentiment classification task, there was no dataset available for the Indian languages; therefore, authors in [36] introduced the IIT-Patna Hindi Reviews dataset to facilitate research in this field for Indian languages. They collected user reviews from various Hindi websites and annotated them manually using a pre-defined set of aspect categories. The reviews belong to four different domains, which are discussed below.

4.1.1. Electronics

The *Electronics* domain contains 3614 reviews for different electronic items like *'Laptops'*, *'Tablets'*, *'Mobiles'*, *'Televisions'*, *'Speakers'*, *'Headphones'*, *'Cameras'*, *'Home Appliances'* and *'Smartwatches'*. The six pre-defined aspect categories for this domain are: *'Hardware'*, *'Software'*, *'Design'*, *'Price'*, *'Ease of use'* and *'Miscellaneous'*. For every aspect, a polarity class among *'Positive'*, *'Negative'*, *'Neutral'* and *'Conflict'* is also provided. Table 3 shows the distribution of instances for each aspect and polarity. Some examples of input sentences and output labels for this domain are presented in Table 4.

Table 3. Number of annotated aspects and their sentiments in IIT-Patna Hindi Reviews dataset: Electronics domain

Category	Polarity				
	Positive	**Negative**	**Neutral**	**Conflict**	**Total**
Hardware	700	261	763	73	1797
Software	160	55	149	6	370
Design	305	69	137	13	524
Price	110	31	83	4	228
Ease of use	70	19	30	3	122
Miscellaneous	290	89	173	21	573
Total	1635	524	1335	120	3614

Table 4. Examples of input sentences and output labels from the IIT-Patna Hindi Reviews Dataset for Electronics domain. The label contains the aspect and its polarity, respectively.

Sentence:
हमारे लगातार वीडियो प्लेबैक टेस्ट में एक्सपीरिया टेबलेट जेड की बैटरी लगभग 5 घंटे तक चलती है , जो कि एक मध्यम स्तर का बैटरी परफॉरमेंस है।
Translated:
The Xperia Tablet Z's battery lasted about 5 h in our continuous video playback test, which is a moderate level of battery performance.
Labels:
(Hardware, Neutral)

Sentence:
यदि क्वालिटी की बात करें तो कैमरा बहुत खास नहीं है लेकिन जिस कीमत में यह डिवाइस मिल रहा है उस हिसाब से कैमरा क्वालिटी अच्छी है।
Translated:
If we talk about quality, then the camera is not very special, but given the price at which this device is available, the camera quality is good.
Labels:
(Price, Positive)
(Hardware, Conflict)

4.1.2. Mobile Apps

The *Mobile Apps* domain contains 197 reviews for various mobile apps. The four predefined aspect categories for this domain are: *'Price'*, *'Ease of use'*, *'GUI'* and *'Miscellaneous'*. The aspect *'GUI'* refers to the graphical user interface of the app. For every aspect, a polarity class of *'Positive'*, *'Negative'*, *'Neutral'* and *'Conflict'* is also provided. Table 5 shows the distribution of instances for each aspect and polarity. Some examples of input sentences and output labels for this domain are presented in Table 6.

Table 5. Number of annotated aspects and their sentiments in IIT-Patna Hindi Reviews dataset: Mobile Apps domain

Category	Polarity				
	Positive	**Negative**	**Neutral**	**Conflict**	**Total**
Price	4	0	6	0	10
Ease of use	18	4	3	1	26
GUI	14	5	8	0	27
Miscellaneous	64	13	57	0	134
Total	100	22	74	1	197

Table 6. Examples of input sentences and output labels from the IIT-Patna Hindi Reviews Dataset for Mobile Apps domain. The label contains the aspect and its polarity, respectively.

Sentence: 60 रूपये प्रति माह में लाइव टीवी स्ट्रीमिंग और ऑन-डिमांड सेक्शन में उपलब्ध एक मैसिव मूवी लाइब्रेरी, यह बुरा सौदा नहीं है। **Translated:** 60 rupees per month for live TV streaming and a massive movie library available in the on-demand section, it's not a bad deal. **Labels:** (Price, Neutral) (Miscellaneous, Positive)
Sentence: इस एप्लीकेशन का लेआउट काफी साफ है और डाउनलोड लिंक को सीधे एप्लीकेशन में पेस्ट करने की सुविधा है। **Translated:** The layout of this application is very clean, and there is a facility to paste the download link directly into the application. **Labels:** (GUI, Positive)

4.1.3. Travel

The *Travel* domain contains 565 reviews for different tourist places. The four pre-defined aspect categories for this domain are: *'Place'*, *'Reachability'*, *'Scenery'* and *'Miscellaneous'*. The aspect *'Reachability'* signifies the convenience in reaching the destination. For every aspect, a polarity class among *'Positive'*, *'Negative'*, *'Neutral'* and *'Conflict'* is also provided. Table 7 shows the distribution of instances for each aspect and polarity. Some examples of input sentences and output labels for this domain are presented in Table 8.

4.1.4. Movies

The *Movies* domain contains 878 reviews for different movies. The four pre-defined aspect categories for this domain are: *'Story'*, *'Performance'*, *'Music'* and *'Miscellaneous'*. The aspect *'Performance'* covers various prospects of the movie, such as action, direction, etc. For every aspect, a polarity class among *'Positive'*, *'Negative'*, *'Neutral'* and *'Conflict'* is also provided. Table 9 shows the distribution of instances for each aspect and polarity. Some examples of input sentences and output labels for this domain are presented in Table 10.

Table 7. Number of annotated aspects and their sentiments in IIT-Patna Hindi Reviews dataset: Travel domain

Category	**Polarity**				
	Positive	**Negative**	**Neutral**	**Conflict**	**Total**
Place	195	5	**103**	**1**	304
Reachability	7	9	19	0	35
Scenery	97	1	24	0	122
Miscellaneous	57	6	41	0	104
Total	356	21	187	1	565

Table 8. Examples of input sentences and output labels from the IIT-Patna Hindi Reviews Dataset for Travel domain. The label contains the aspect and its polarity, respectively.

Sentence: साल भर बर्फ से लदे रहने वाले गगन चूमते पर्वत, प्राकृतिक सौंदर्य के मालिक मनाली के पर्यटन स्थल और देव-संस्कृति आदि यहां आने वाले पर्यटकों के दिलों में गहराई तक उतर जाते हैं। **Translated:** The sky kissing mountains, which are covered with snow throughout the year, the tourist places of Manali, the owner of natural beauty and the God-culture etc., get deep into the hearts of the tourists visiting here. **Labels:** (Place, Positive) (Scenery, Positive)
Sentence: इस किले में पहुंचने के लिए एक खड़े और घुमावदार मार्ग से होकर जाना होता है। **Translated:** One has to go through a steep and winding route to reach this fort. **Labels:** (Reachability, Negative)

Table 9. Number of annotated aspects and their sentiments in IIT-Patna Hindi Reviews dataset: Movies domain

Category	Polarity				
	Positive	**Negative**	**Neutral**	**Conflict**	**Total**
Story	6	11	17	1	35
Performance	109	35	95	5	244
Music	14	5	8	0	27
Miscellaneous	30	17	525	0	572
Total	159	68	645	6	878

Table 10. Examples of input sentences and output labels from the IIT-Patna Hindi Reviews Dataset for Movies domain. The label contains the aspect and its polarity, respectively.

Sentence: दिबाकर बनर्जी ने शरदिंदु बनर्जी की सभी 32 कहानियों के अधिकार लेकर उन्हें अपनी फिल्म 'डिटेक्टिव ब्योमकेश बक्शी' में सुविधानुसार इस्तेमाल किया है। **Translated:** Dibakar Banerjee has taken the rights of all 32 stories of Sharadindu Banerjee and used them conveniently in his film 'Detective Byomkesh Bakshi'. **Labels:** (Story, Neutral)
Sentence: यहां तक कि इस फिल्म के कलाकारों की एक्टिंग ऐसी है कि लगता है वे नशे में जरुरत से ज्यादा ऊर्जावान है और उनका उत्साह चिढ़ानेवाला है। **Translated:** Even the acting of the actors of this film is such that it seems that they are more energetic than necessary while intoxicated and their enthusiasm is irritating. **Labels:** (Performance, Negative)

4.2. Result Analysis

For fine-tuning, we use the multilingual-cased BERT-base pre-trained model on Hindi datasets. For the model, the number of transformer blocks and the self-attention heads is

12 each, the size of the hidden layer is 768, and the total number of parameters is 110 M. The dropout probability is set at 0.1 while fine-tuning. The optimizer used is 'Adam' and the activation function is 'gelu'. Table 11 summarizes the different hyperparameters' values used in the experiments.

Table 11. Hyperparameters' values used in the experiments for different datasets.

Datasets	Max Sequence Length	Batch Size	Learning Rate	Training Epochs
IIT-Patna Hindi: Electronics	128	16	5×10^{-6}	14
IIT-Patna Hindi: Mobile Apps	128	16	5×10^{-6}	9
IIT-Patna Hindi: Travel	128	16	5×10^{-6}	8
IIT-Patna Hindi: Movies	128	16	5×10^{-6}	8

All experiments were conducted on an Intel(R) Xeon(R) Gold 5120 CPU @ 2.20 GHz, 96 GB RAM, and NVIDIA Quadro P5000 graphic card with 16 GB memory. The results of the various datasets are presented in the following sections.

For each domain in the IIT-Patna Hindi reviews dataset, a separate mBERT ensemble model is trained. For the experiments, all four datasets are split into training and testing sets in the ratio of 4:1. The results obtained for each domain are presented in the following subsections. We use the results reported in [36] for comparison purposes. The authors in [36] used two techniques, the (i) binary relevance approach and (ii) label powerset approach, to solve the multi-label aspect category detection subtask. In the binary relevance approach, the first n distinct models are build for each n unique label. Then, the final prediction is produced by combining the predictions of the n models. However, in the label powerset approach, each label combination is treated as a unique label. The model is then trained and evaluated on these labels.

For all the datasets, we have compared our ensemble models with the best-performing individual mBERT model.

4.2.1. Electronics

The results obtained for the *Electronics* domain are presented in Table 12. It was observed that both mBERT-E-MV and mBERT-E-AS models achieved much better results than the previous models on both the subtasks. The mBERT-E-MV model achieved the best F1-score on the aspect category detection task while mBERT-E-AS achieved the best accuracy on the aspect polarity classification task.

Table 12. Results of Aspect Category Detection and Aspect Polarity Classification tasks for IIT-Patna Hindi Reviews Dataset: Electronics domain.

Method	Aspect Category Detection						Polarity
	Binary Relevance			Label Powerset			
	P	R	F1	P	R	F1	Accuracy
NB [36]	31.62	37.63	34.37	48.00	45.05	46.46	50.95
DT [36]	49.61	17.28	25.63	31.73	31.73	31.73	54.48
SMO [36]	26.70	46.93	34.03	39.36	44.90	41.94	51.07
Method	Precision		Recall		F1-Score		Accuracy
mBERT-QA-M	80.71		67.11		73.28		65.90
mBERT-E-MV	84.39		66.31		**74.26**		69.95
mBERT-E-AS	84.53		64.82		73.38		**70.49**

4.2.2. Mobile Apps

The results obtained for the *Mobile Apps* domain are presented in Table 13. Both mBERT-E-MV and mBERT-E-AS models achieved better accuracy than the previous models on the aspect polarity classification task. However, the mBERT-E-AS model fails to

surpass the F1-score value obtained by the *Naive Bayes* model for the aspect category detection task. Overall, mBERT-E-MV turns out to be the best performer in both the subtasks.

Table 13. Results of Aspect Category Detection and Aspect Polarity Classification tasks for IIT-Patna Hindi Reviews Dataset: Mobile Apps domain.

Method	**Aspect Category Detection**						**Polarity**
	Binary Relevance			**Label Powerset**			
	P	**R**	**F1**	**P**	**R**	**F1**	**Accuracy**
NB [36]	39.30	46.19	42.47	59.20	54.09	**56.53**	46.78
DT [36]	44.28	41.75	42.97	85.07	24.89	38.51	47.95
SMO [36]	51.73	38.47	44.12	45.77	57.14	50.82	42.10
Method	Precision		Recall		F1-Score		Accuracy
mBERT-QA-M	72.22		31.70		44.06		47.68
mBERT-E-MV	76.92		48.78		**59.70**		**51.22**
mBERT-E-AS	70.83		41.46		52.31		48.78

4.2.3. Travel

The results obtained for the *Travel* domain are presented in Table 14. Both mBERT-E-MV and mBERT-E-AS models achieved much better results than the previous models for both subtasks. The mBERT-E-MV model is the best performer for the aspect category detection task. In contrast, mBERT-E-MV and mBERT-E-AS performs equally well on the aspect polarity classification task. Overall, for this domain, mBERT-E-MV model performed better than the other models.

Table 14. Results of Aspect Category Detection and Aspect Polarity Classification tasks for IIT-Patna Hindi Reviews Dataset: Travel domain

Method	**Aspect Category Detection**						**Polarity**
	Binary Relevance			**Label Powerset**			
	P	**R**	**F1**	**P**	**R**	**F1**	**Accuracy**
NB [36]	26.84	26.88	26.86	20.87	31.90	25.23	56.06
DT [36]	27.98	22.73	25.08	99.82	18.33	30.97	65.20
SMO [36]	25.51	20.67	22.83	15.61	39.55	22.38	60.63
Method	Precision		Recall		F1-Score		Accuracy
mBERT-NLI-M	73.91		48.11		58.28		65.09
mBERT-E-MV	76.32		54.72		**63.74**		**75.47**
mBERT-E-AS	78.46		48.11		59.65		**75.47**

4.2.4. Movies

The results obtained for the *Movies* domain are presented in Table 15. From the table, it can be observed that the mBERT-E-MV and mBERT-E-AS models achieved better results than the previous models for the aspect category detection task. However, they failed to surpass the results obtained by *DT* and *SMO* models in the aspect polarity classification task by a significant amount. Among the ensemble models, mBERT-E-MV performed better in the aspect category detection task while mBERT-E-AS achieved better results for the aspect polarity classification task.

Table 15. Results of Aspect Category Detection and Aspect Polarity Classification tasks for IIT-Patna Hindi Reviews Dataset: Movies domain

Method	Aspect Category Detection						Polarity
	Binary Relevance			Label Powerset			
	P	**R**	**F1**	**P**	**R**	**F1**	**Accuracy**
NB [36]	41.99	65.44	51.15	56.66	63.32	59.81	87.78
DT [36]	47.45	58.12	52.24	64.16	64.38	64.27	**91.62**
SMO [36]	43.78	59.81	50.55	48.60	63.26	54.97	**91.62**
Method	Precision		Recall		F1-Score		Accuracy
mBERT-NLI-M	77.09		77.52		77.31		73.03
mBERT-E-MV	80.70		77.53		**79.08**		78.09
mBERT-E-AS	80.95		76.40		78.61		**79.77**

5. Conclusions and Future Work

This paper proposes two ensemble models based on Multilingual BERT, namely *mBERT-E-MV* and *mBERT-E-AS*. Our proposed models outperformed the existing state-of-the-art models on Hindi datasets. On the III-Patna Hindi Reviews dataset, *mBERT-E-MV* reports F1-scores of 74.26%, 59.70%, 63.74% and 79.08% on the aspect category detection task in Electronics, Mobile Apps, Travel and Movies domains, respectively. It reports accuracies of 69.95%, 51.22%, 75.47% and 78.09% on the aspect polarity classification task for the four respective domains. Similarly, *mBERT-E-AS* reports F1-scores of 73.38%, 52.31%, 59.65% and 78.61% on the aspect category detection task for the respective domains. It reports accuracies of 70.49%, 48.78%, 75.47% and 79.77% on the aspect polarity classification task for the four respective domains.

Overall, BERT-based models performed much better than the other models. This is possible because of the construction of auxiliary sentences from the aspect information, which is analogous to exponentially increasing the dataset. A sentence s_i in the original dataset is transformed into (s_i, a_1), (s_i, a_2),…, (s_i, a_{n_a}) in the sentence pair classification task. The BERT model has an additional advantage in handling sentence pair classification tasks, which is evident from its impressive improvement on the QA and NLI tasks. This improvement comes from both unsupervised Masked Language Modeling (MLM) and the Next Sentence Prediction (NSP) tasks, which are used to pre-train the BERT model [7]. In MLM, a word in a sentence is masked, and then the model is trained to predict which word was masked based on the context of the word. In NSP, the model is trained to predict whether the two input sentences are connected logically/sequentially or whether they are unrelated to each other.

In future work, the proposed system can be applied to other NLP problems. As is evident from the obtained results, there is scope for augmenting the Hindi datasets for further improvements in performance. There is also scope for introducing a dataset for the TABSA task in Indian languages, as there is no dataset available for the same purpose.

Author Contributions: Conceptualization, A.P. and S.K.; methodology, A.P.; writing—original draft preparation A.P. and S.K.; supervision, P.P.R. and B.-G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We have used IIT-Patna Hindi Reviews dataset to facilitate research in the Indian languages.

Conflicts of Interest: The authors declared that they have no conflict of interest to this work.

References

1. Kumar, S.; Yadava, M.; Roy, P.P. Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *Inf. Fusion* **2019**, *52*, 41–52. [CrossRef]

2. Gauba, H.; Kumar, P.; Roy, P.P.; Singh, P.; Dogra, D.P.; Raman, B. Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Netw.* **2017**, *92*, 77–88. [CrossRef] [PubMed]
3. Kumar, S.; Gahalawat, M.; Roy, P.P.; Dogra, D.P.; Kim, B.G. Exploring impact of age and gender on sentiment analysis using machine learning. *Electronics* **2020**, *9*, 374. [CrossRef]
4. Kumar, S.; De, K.; Roy, P.P. Movie recommendation system using sentiment analysis from microblogging data. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 915–923. [CrossRef]
5. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
6. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. *Improving Language Understanding with Unsupervised Learning*; Technical Report; OpenAI: San Francisco, CA, USA, 2018.
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
8. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. *arXiv* **2019**, arXiv:1903.09588.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
10. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
11. Goldberg, Y.; Levy, O. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *arXiv* **2014**, arXiv:1402.3722.
12. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6–7 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 79–86.
13. Wagner, J.; Arora, P.; Vaíllo, S.C.; Barman, U.; Bogdanova, D.; Foster, J.; Tounsi, L. DCU: Aspect-based Polarity Classification for SemEval Task 4. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 223–229.
14. Rao, D.; Ravichandran, D. Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009; pp. 675–682.
15. Vo, D.T.; Zhang, Y. Target-dependent twitter sentiment classification with rich automatic features. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, 25–31 July 2015; Yang, Q., Wooldridge, M., Eds.; AAAI Press: Palo Alto, CA, USA, 2015; pp. 1347–1353.
16. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014; pp. 437–442.
17. Castellucci, G.; Filice, S.; Croce, D.; Basili, R. Unitor: Aspect based sentiment analysis with structured learning. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 761–767.
18. Mowlaei, M.E.; Abadeh, M.S.; Keshavarz, H. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Syst. Appl.* **2020**, *148*, 1–13. [CrossRef]
19. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the International Joint Conferences on Artifical Intelligence(IJCAI), Melbourne, VIC, Australia, 19–25 August 2017; pp. 4068–4074.
20. Ma, Y.; Peng, H.; Cambria, E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5876–5883.
21. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
22. Liu, F.; Cohn, T.; Baldwin, T. Recurrent Entity Networks with Delayed Memory Update for Targeted Aspect-Based Sentiment Analysis. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–8 June 2018; pp. 278–283.
23. Zhang, C.; Li, Q.; Song, D. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4568–4578.
24. Majumder, N.; Poria, S.; Gelbukh, A.; Akhtar, M.S.; Cambria, E.; Ekbal, A. IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2–4 November 2018; pp. 3402–3411.
25. Liang, B.; Du, J.; Xu, R.; Li, B.; Huang, H. Context-aware Embedding for Targeted Aspect-based Sentiment Analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4678–4683.

26. Chen, P.; Sun, Z.; Bing, L.; Yang, W. Recurrent attention network on memory for aspect sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 452–461.
27. Huang, B.; Carley, K.M. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5472–5480.
28. Hu, M.; Zhao, S.; Guo, H.; Cheng, R.; Su, Z. Learning to Detect Opinion Snippet for Aspect-Based Sentiment Analysis. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 970–979.
29. Yu, J.; Jiang, J. Adapting bert for target-oriented multimodal sentiment classification. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10–16 August 2019.
30. Lin, P.; Yang, M.; Lai, J. Deep Mask Memory Network with Semantic Dependency and Context Moment for Aspect Level Sentiment Classification. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10–16 August 2019; pp. 5088–5094.
31. Xu, H.; Liu, B.; Shu, L.; Philip, S.Y. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 6–7 June 2019; pp. 2324–2335.
32. Li, X.; Bing, L.; Zhang, W.; Lam, W. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019), Hong Kong, China, 4 November 2019; pp. 34–41.
33. Rietzler, A.; Stabinger, S.; Opitz, P.; Engl, S. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. *arXiv* **2019**, arXiv:1908.11860.
34. Song, Y.; Wang, J.; Liang, Z.; Liu, Z.; Jiang, T. Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference. *arXiv* **2020**, arXiv:2002.04815.
35. Li, X.; Fu, X.; Xu, G.; Yang, Y.; Wang, J.; Jin, L.; Liu, Q.; Xiang, T. Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access* **2020**, *8*, 46868–46876. [CrossRef]
36. Akhtar, M.S.; Ekbal, A.; Bhattacharyya, P. Aspect based sentiment analysis: Category detection and sentiment classification for Hindi. In *Computational Linguistics and Intelligent Text Processing, CICLing 2016*; Gelbukh, A., Ed.; Springer: Cham, Switzerland, 2016; pp. 246–257.
37. Hatzivassiloglou, V.; McKeown, K. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July 1997; pp. 174–181.
38. Cambria, E. Affective Computing and Sentiment Analysis. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
39. Lo, S.L.; Cambria, E.; Chiong, R.; Cornforth, D. Multilingual sentiment analysis: From formal to informal and scarce resource languages. *Artif. Intell. Rev.* **2017**, *48*, 499–527. [CrossRef]
40. Vilares, D.; Peng, H.; Satapathy, R.; Cambria, E. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1292–1298.
41. Basiri, M.E.; Nemati, S.; Abdar, M.; Cambria, E.; Acharya, U.R. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Gener. Comput. Syst.* **2021**, *115*, 279–294. [CrossRef]
42. Akhtar, S.; Ghosal, D.; Ekbal, A.; Bhattacharyya, P.; Kurohashi, S. All-in-One: Emotion, Sentiment and Intensity Prediction using a Multi-task Ensemble Framework. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
43. Wang, J.; Peng, B.; Zhang, X. Using a stacked residual LSTM model for sentiment intensity prediction. *Neurocomputing* **2018**, *322*, 93–101. [CrossRef]