An Aspect-Based Sentiment Classification Method Using the Pachinko Allocation Model

Yojitha Chilukuri
St. Jude Children's Cancer Research Hospital, Danny Thomas Place, Memphis, USA
chilukuriyojitha@gmail.com

T. Jhansi Rani
Department of CSE, GITAM (Deemed to be) University, Hyderabad, India
jhansi.rani.t@gmail.com

N. Lakshmipathi Anantha
VIT-AP University, Amaravati, Andhra Pradesh, India
anlakshmipathi@gmail.com

Gopisetty Rathnamma
Department of CSE GITAM (Deemed to be) University, Hyderabad, India
rgurram@gitam.edu (corresponding author)

Ulligaddala Srinivasarao
Department of CSE GITAM (Deemed to be) University, Hyderabad, India
ulligaddalasrinu@gmail.com

P. Sowjanya
Department of CSE GITAM (Deemed to be) University, Hyderabad, India
sponnuru@gitam.edu

Nemala Jayasri
MLR Institute of Technology, Dundigal, Hyderabad, Telangana, India
jayasree@mlrit.ac.in

Rakesh Kumar Donthi
Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India
drrakesh2175@gmail.com

B. Krishna Chaitanya
Department of Electrical and Electronics Engineering, Chaitanya Bharathi Institute of Technology, Hyderabad, India
krishnachaitanyab_eee@cbit.ac.in

M. Ramesh
Department of CSE GITAM (Deemed to be) University, Hyderabad, India
rmunipal@gitam.edu

Received: 17 April 2025 | Revised: 7 May 2025, 19 May 2025, and 26 May 2025 | Accepted: 31 May 2025

 DOI: https://doi.org/10.48084/etasr.11555

ABSTRACT

Sentiment analysis is based on extracting features from customer reviews using topic modeling and Latent Dirichlet Allocation (LDA) from large textual data to extract useful information. Different methods have been proposed to adapt LDA to short texts. This study uses a Pachinko Allocation Model (PAM)-based method to examine opinions and extract aspects related to product information, using data augmentation to improve model training. Feature extraction is performed using the TF-IGM and TF-IDF-ICSDF methods, and an opinion lexicon is used to extract sentiment. The experimental results show that the PAM method provides accurate results in extracting aspect sentiments. The proposed method was compared with existing models and evaluated on two datasets for sentiment classification of reviews, achieving better accuracy.

Keywords-feature extraction; PAM; aspect-based sentiment analysis; topic modeling

I. INTRODUCTION

As internet applications expand, the volume of brief messages containing client comments has increased massively. Along these messages, feeling examination and client assessment mining have become important themes using different granularities of sentiment analysis. Comments are classified into negative and positive reviews to assess user attitude. However, in most cases, more than sentiment analysis is required at this level. Aspect-Based Sentiment Analysis (ABSA) should be used for a more comprehensive evaluation, extracting what users say about products, such as the polarity of their feelings and preferences. It is important to note that only the overall sentiment of a text may be discerned by sentiment analysis, as opposed to ABSA. Instead of analyzing a text's overall emotion, ABSA approaches allow linking distinct thoughts with different traits or qualities of a product or service, digging deeper into the content of a text to uncover hidden meanings, offering more specific results.

In [1], classic machine-learning techniques were used to improve sentiment analysis, with One R being superior to Naive Bayes, J-48, and BF tree. In [2], the ABSA challenge was examined in Arabic hotel reviews using deep Recurrent Neural Network (RNN) and SVM classifiers, the latter being better in terms of classification accuracy. In [3], SVM and ANN classifiers were used to analyze extracted features, with the ANN outperforming the SVM by a small margin. In [4], synthetic training samples for ABSA were produced using large language models in low-resource environments.

This study presents an ABSA classification model for product reviews using sentiment analysis and the Pachinko Allocation Model (PAM) to classify sentiments for products and services. As trust is the primary concern in online purchases, a model should rapidly assess user sentiment toward a product and its essential features with little computational effort. This study tested the proposed model on hotel and smartphone phone reviews. Latent Dirichlet Allocation was used to extract topic and product information, and PAM was used for topic modeling to identify features in reviews. PAM can be beneficial in sentiment analysis, as it can reveal complex connections between subjects and sentiment polarity scores, performing very well in sentiment classification. This study evaluates different emotions, improves ABSA performance in product reviews, and efficiently classifies sentiment information.

Several studies on sentiment analysis have been published in the literature, especially in recent years. In [5], natural language processing and sentence syntactic analysis were used to extract aspects, adhering to a set of guidelines based on word PoS tags. In [6], a rule-based strategy was proposed for ABSA, based on greedy search and simulated annealing to find the ideal rules. In [7], a graph data structure was used for ABSA in Chinese social media communications. The aspect words are nodes in the network, and each edge is weighted based on the emotional connection between the two nodes. The graph data structure was combined with these word embeddings. Recently, the use of sentiment analysis techniques, such as word embeddings, has increased. In [8], it was stated that word embeddings in sentiment analysis have improved. This study considered all PoS tags, lexicon-based techniques, word position algorithms, and word embedding algorithms, finding a higher degree of success. In [9], two approaches for sentiment analysis were presented, one using a genetic algorithm and the other based on the frequency of phrases in favorable and unfavorable remarks. In [10], product reviews from several websites were used for experiments.

The Latent Dirichlet Allocation (LDA) method can process large datasets and lengthy documents. It is essential to know how often a particular term appears in the document using LDA. However, LDA is useless for short texts because there are not enough co-occurrence patterns and data. The study of LDA adaptation for brief texts is a hot topic. In [11], an additive regularization method was proposed for sentiment analysis. Long-Short-Term Memory (LSTM)-based strategies have recently risen to prominence in the ABSA challenge. In [12], clinical remarks from ICU patients were used to build a mortality prediction model. In [13], the combination of PDA, LSTM, and CRF was proposed to discover named elements from user-generated social media data that are uncommon or new. In [14], an LDA-GRU hybrid model was proposed for feature extraction and sentiment classification of texts. In [15], a sentiment classification approach was developed based on machine learning methods using Amazon product reviews.

II. BACKGROUND

A. *Pachinko Allocation Model*

A Discrete Acyclic Graph (DAG) structure is used in the probabilistic PAM to show and identify sparse topic correlations [16]. Table I shows the detailed notation in PAM.

TABLE I. NOTATIONS OF PAM

V	Words vocabulary $\{w_1, w_2, w_3, \ldots, w_n\}$
R	The root, a special topic in T
T	A set of topics $\{t_1, t_2, t_3, \ldots, t_s\}$
D	A document, the comments in this study
$g_i(\alpha_i)$	Dirichlet distribution associated with the topic t_i
Z_{wi}	The i^{th} topic sampled for word w
$\theta_{ti}^{(d)}$	Multinomial distribution sampled from topic t_i for document d

PAM can be used to create new documents and search for documents in a database, just like other topic models. The procedure required to create a document in PAM is as follows:

1. Sample $\theta_{t1}^{(d)}, \theta_{t2}^{(d)}, \ldots, \theta_{ts}^{(d)}$ from $g_1(a_1)$, $g_2(a_2)$, $\ldots, g_w(a_s)$, where $\theta_{ti}^{(d)}$ is the multinomial distribution of topic t_i over its children.
2. The topic route is sampled whenever w appears in the text. Since Zw_1 is the root and Zw_2 through Zw_{Lw} are topic nodes in topics T, the Zw of length Lw is sampled. Zw_i is a Zw_{i-1} descendant sampled using a multinomial distribution. The word w is sampled from $\theta_{ZwLw}^{(d)}$.

The combined likelihood of creating document d is established by:

$$P\left(d, z^{(d)}, \frac{\theta^{(d)}}{\alpha}\right) = \prod_{i=1}^{s} P\left(\frac{\theta_{ti}^{(d)}}{\alpha_i}\right) \times \sum_{w}\left(\prod_{i=2}^{Lw} P\left(Z_{wi}/\theta_{ZwLw}^{(d)}\right) P\left(w/\theta_{ZwLw}^{(d)}\right)\right) \quad (1)$$

$$\left(\frac{d}{a}\right) = \int \sum_{i=1}^{s} P\left(\theta_{ti}^{(d)}/\alpha_i\right) \times \prod w \sum_{z_w}\left(\prod_{i=2}^{Lw} P\left(Z_{wi}/\theta_{z_{w(i-1)}}^{(d)}\right) P\left(w/\theta_{zwLw}^{(d)}\right)\right) d\theta^{(d)} \quad (2)$$

B. *Parameter Estimation*

The default setting of 1e-3 was used for the Adam optimizer for the entire model. The batch-size setting and initialization impact the topic coherence, particularly for the smaller smartphone reviews and hotel reviews datasets. The encoder capacity was set using a grid search while switching between the two datasets. In general, for PAM, the encoder capacity should increase as the vocabulary size does.

$$Mean_{ij} = \frac{1}{N_i} \times \sum_d \frac{n_{ij}^{(d)}}{n_i^{(d)}} \quad (3)$$

$$Vari_{ij} = \frac{1}{N_i} \times \sum_d \left(\frac{n_{ij}^{(d)}}{n_i^{(d)}} - mean_{ij}\right)^2 \quad (4)$$

$$m_{ij} = \frac{mean_{ij} \times (1 - mean_{ij})}{var_{ij}} - 1 \quad (5)$$

$$\propto_{ij} = \frac{mean_{ij}}{Exp\left(\frac{\sum_j \log(m_{ij})}{s' - 1}\right)} \quad (6)$$

C. *SentiwordNet*

Polarity was determined using SentiWorldNet 3.0 [17]. This technique is based on the training of a group of ternary classifiers, each of which can determine if a synset is objective, negative, or positive. The classification results of the WordNet synsets are different for each ternary classifier due to differences in the training sets. The normalized percentage of ternary classifiers that have given a synset the corresponding label results in opinion-related ratings for that synset. Each label has a score proportionate to the number of classifiers assigned, but if all ternary classifiers agree to give the same label to a synset, then the label will get the maximum score for that synset. Synsets have scores ranging from -0.75 to 1.0, and their total is 1.0. A synset might have non-zero scores in each of the three classifications, showing that the related words have some of the three viewpoint attributes to some degree, as demonstrated by the synset.

III. PROPOSED METHOD

The datasets were preprocessed, forming a collection of tokenized words. A separate database was maintained for the sentiment analysis of tokenized sentences. PAM is applied to a bag of tokenized terms to produce the subject word distribution. These fundamental elements were recognized using POS criteria and likelihood values. These aspects are ranked according to their probability distribution value on the subject. Aspect classification is aided by adding domain-specific terms as extended aspect terms. In the same way, the emotion score is determined for each of the other aspects. The average sentiment score can be used to determine the sentiment strength of a particular aspect in the review data. These reviews are analyzed as shown in Figure 2 and Algorithm 1.

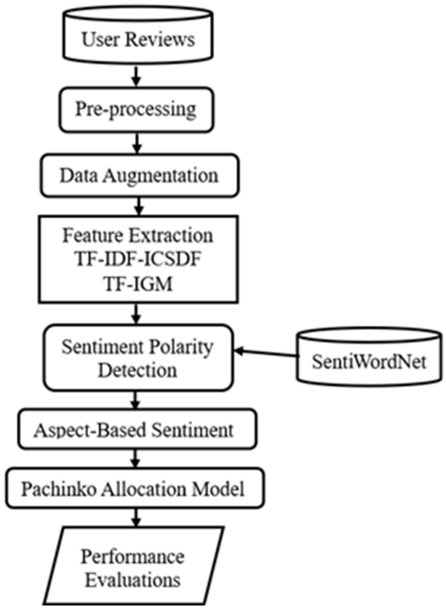

Fig. 1. Flowchart of the proposed method.

```
Algorithm 1: Proposed Method
Input:  Review Data
Output: Reviews per aspect category in
        sentiment
Step 1: Input review corpus
Step 2: Data preprocessing using
        tokenization, stop words, and PoS
```

```
          tagging etc
Step 3: Data augmentation
Step 4: Feature extraction methods: TF-
          IDF-ICSDF, TF-IGM(8)
Step 5: Apply sentiment lexicon algorithm
         (SentiWordnet)
Step 6: Categorize aspects into various
          groups
Step 7: Apply PAM for topic-word
          probability
Step 8: Aspect probability values
```

A. Dataset Description

This study used two datasets from two different fields. The Europe Hotel Reviews dataset contains several fields available, but this study used only a few, specifically those that are Negative, Positive, Neutral, and Reviewer scores [18]. Figure 2 shows two word clouds for the Hotel negative and positive reviews. A smartphone dataset was also used, which included 9716 reviews of different smartphones [19]. Both datasets are freely available. Since ABSA is a major challenge in the field of data analysis, this study used a large amount of reviewer comments to evaluate the proposed model.

Fig. 2. Word cloud of Hotel Negative (left) and Positive (right) reviews

TABLE II. DATASET STATISTICS

Dataset	Total
Hotel reviews	1493
Smartphone reviews	9716

B. Data Preprocessing

Text preprocessing is a three-step process that includes the following: (i) Tokenization is the process of splitting large words into small tokens, (ii) Stop words removal is used to remove unnecessary words such as is, was, were, etc., from the given input dataset, and (iii) PoS tagging is used to identify each word and assign it to some grammatical tags, such as adjective, noun, and verb. PoS can understand each word with various NLP analysis tasks. NLP tools are used to correct types and perform morphological analysis tasks.

C. Data Augmentation with Random Word Replacement

Data augmentation is the process of applying several alterations to current data in order to artificially expand the size and variety of a training dataset [20]. This is especially helpful when the initial dataset is small or when enhancing a model's capacity to generalize to new data is required. Two stages are crucial in the data argumentation framework:

- Random Insertion: This technique chooses a random word in a phrase, and a Natural Language Toolkit (NLTK) is used to discover a synonym and then insert it at a random location. The target phrase should not contain the insertion, which is ensured by substituting a fixed expression for the target before the insertion. Next, the phrase is appended to the training data, and the target expression is substituted once more for the fixed expression.

- Synonym Replacement. This technique is used to find a synonym for a random word in a phrase and substitute it. The same process as above can be used to ensure that this phrase cannot be chosen for the synonym substitution, but a fixed expression can be substituted for the target.

D. Feature Extraction Methods

This part suggests a two-term weighting strategy called TF-IDF-ICSDF and TF-IGM to effectively communicate the ability of terms to differentiate from the vector space model. TF-IDF-ICSDF and TF-IGM are the names of these two weighing systems, respectively. They contain term frequency components from the weighting schemes TF-IDF-ICSDF, TF-IGM, and TF-IGM.

E. TF-IDF-IC-SDF

IC-SDF accounts for the distribution of inter-class documents to calculate weighting values for each term [21]. Calculated using the TF-IDF-IC-SDF term weighting system, this term's weighting value uses class-specific data values (t_i) and the number of texts in class j:

$$W_{TF.IDF.ICSDF}(t_i) = TF(t_i, d_k) * \left(1 + log\left(\frac{D}{d(t_i)}\right)\right) * \left(1 + log\left(\frac{C}{\sum_{j=1}^{C}\frac{df_{ij}}{D_j}}\right)\right) \quad (7)$$

F. TF-IGM

This new supervised term weighting strategy combines the TF and Inverse Gravity Moment (IGM) information of terms as the collection frequency factor. IGM is a statistical model modified for term weighting purposes [22]. There are a few ways to compute inter-class distributions, and the IGM approach is one of the most commonly used. TF-IGM is calculated using:

$$W_{TF-IGM}(t_i) = TF(t_i, d_k) * \left(1+\propto*\frac{f_{i1}}{\sum_{r=1}^{C} f_{ir}*r}\right) \quad (8)$$

The word frequency of the phrase t_i is expressed by the constant f_{ir}. Sorting documents by frequency and then by rank $(r = 1, 2, 3, \ldots, C)$ is the basis of the IGM weighting approach. This equation includes a balance parameter, with a defined range of values of 5.0-9.0, with the anticipated value being 7.0.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This study calculated an overall ABSA average value by averaging the scores for each category. Tables IV and V evaluate the performance of the proposed and different existing methods, such as ANN, SVM, LDA, Random Forest (RF), and ViT, to determine the best method for smartphone and hotel review sentiment evaluations, respectively. The results suggest that the proposed model achieved better results in terms of accuracy. The LDA model is improved using PAM, which uses DAG to incorporate topical relationships. The model can better capture semantic links and generate more coherent themes with this method. PAM may find more significant and cohesive themes in a dataset by considering topic correlations and improving the interpretation of the findings. The Sentiment Analysis (SA) values are classified into 3 classes: Positive (P), Negative (N), and Neutral (Ne).

TABLE III. COMPARISON OF POSITIVE, NEGATIVE, AND NEUTRAL REVIEWS FOR SMARTPHONE REVIEWS

Method	Dataset	SA	A	P	R	F1
PAM	Smart-phone reviews	P	97.09	96.18	98.46	97.30
		N	86.17	86.19	87.43	86.80
		Ne	81.14	80.29	83.37	81.80
ANN		P	89.19	89.14	87.13	88.12
		N	83.07	83.21	85.71	83.42
		Ne	81.51	81.59	83.74	82.65
SVM		P	79.09	79.01	79.46	79.23
		N	73.14	72.19	72.81	72.49
		Ne	76.05	75.12	75.97	75.54
LDA		P	87.45	88.71	89.09	88.89
		N	81.24	82.19	83.76	82.96
		Ne	75.08	72.91	72.81	72.85
RF		P	71.19	72.57	73.87	73.21
		N	78.41	78.09	76.86	77.47
		Ne	75.07	73.67	78.09	74.82
ViT		P	92.21	91.03	93.25	92.12
		N	91.71	89.29	90.31	89.79
		Ne	88.67	87.37	89.47	88.40

TABLE IV. COMPARISON OF POSITIVE, NEGATIVE, AND NEUTRAL HOTEL REVIEWS

Method	Dataset	SA	A	P	R	F
PAM	Hotel reviews	P	95.11	96.51	93.36	94.90
		N	85.21	86.39	83.13	84.72
		Ne	80.15	82.12	79.37	80.72
ANN		P	82.21	84.54	80.34	82.38
		N	82.17	83.21	80.81	81.99
		Ne	81.01	82.79	81.45	83.12
SVM		P	78.09	79.01	77.56	78.27
		N	70.19	71.24	69.51	70.36
		Ne	74.25	75.42	72.47	73.91
LDA		P	88.25	89.31	86.91	89.61
		N	83.24	85.79	80.86	83.25
		Ne	77.58	78.11	75.14	76.59
RF		P	76.11	77.27	74.71	75.96
		N	79.12	80.29	77.16	78.69
		Ne	77.77	78.09	75.01	76.51
VIT		P	93.29	90.35	93.47	91.88
		N	91.67	89.92	90.15	90.03
		Ne	86.91	85.91	87.68	86.78

The performance of the classifiers is evaluated using precision measurements. A better precision score means that the data is classified more accurately and has fewer false positives. Recall is used to evaluate the thoroughness or sensitivity of a classifier. The data is classified with fewer false negatives when the recall value is higher. Figures 3 and 4 show the results of different models on the two datasets. Figure 5 shows the ROC curve and AUC for the different models tested on the Hotel reviews dataset.

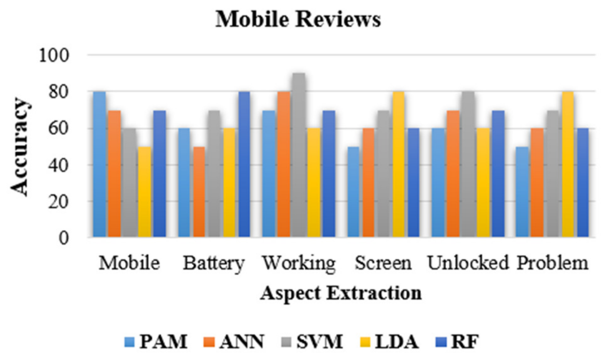

Fig. 3. Aspect sentiment scores for smartphone reviews.

Fig. 4. Aspect sentiment scores for hotel reviews.

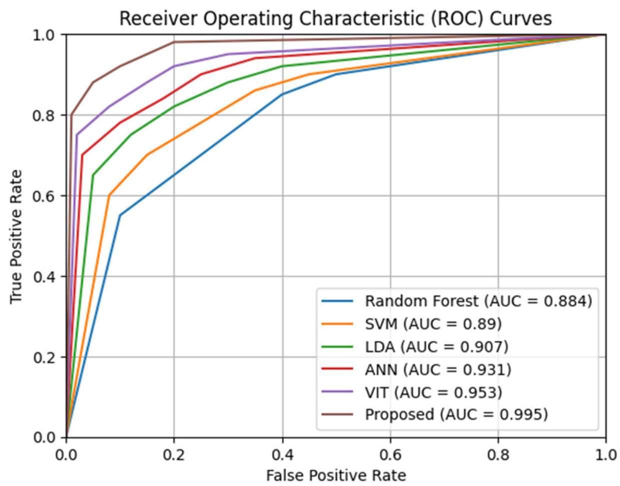

Fig. 5. Analysis of ROC and AUC.

Figure 6 shows the execution time for the proposed and other models in the smartphone and hotel review datasets. Figure 7 compares the accuracy of the proposed method using two other popular sentiment analysis lexicons, Valence Aware Dictionary and sEntiment Reasoner (VADER) [23] and AFINN [24]. SentiWordNet offered a better sentiment classification approach in these datasets in terms of accuracy performance. Figure 8 shows the confusion matrix of the results for the same dataset.

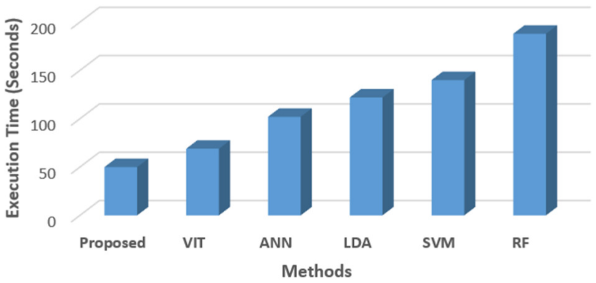

Fig. 6. Execution time for different methods.

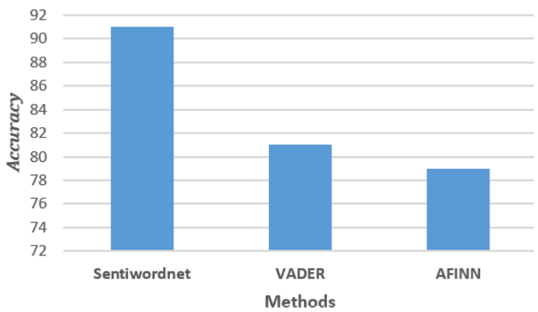

Fig. 7. Accuracy using different sentiment lexicons.

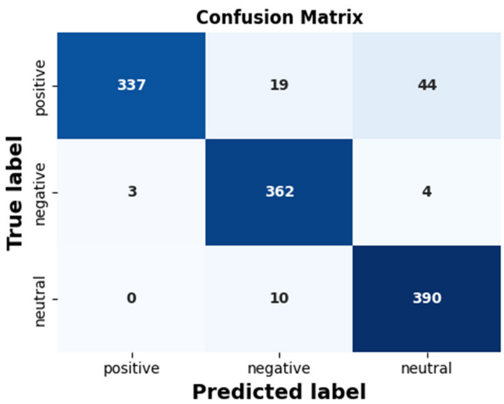

Fig. 8. Confusion Matrix for hotel reviews.

V. CONCLUSION

This study presented a method for ABSA using a PAM model and SentiWordNet. This approach involved preparing the datasets using unstructured review data. This method can be used for several types of review data by making minor adjustments. PAM is a probability-based approach for extracting different sentences and similar product-based comments. The novelty of the work is the use of PAM and SentiWordNet to coextract aspects and opinions. Since deep learning and word embedding techniques have become popular in recent years, several studies have shown that PAM gives better results in aspect-based analysis. Future research directions involve improving sentence segmentation or using different algorithms. Experimental analysis shows that the combination of PAM and WordNet outperforms previous state-of-the-art models for aspect-based sentiment classification tasks.

DATASET AVAILABILITY STATEMENT

The dataset is open source and publicly available.

REFERENCES

[1] J. Singh, G. Singh, and R. Singh, "Optimization of sentiment analysis using machine learning classifiers," *Human-centric Computing and Information Sciences*, vol. 7, no. 1, Dec. 2017, Art. no. 32, https://doi.org/10.1186/s13673-017-0116-3.

[2] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *Journal of Computational Science*, vol. 27, pp. 386–393, Jul. 2018, https://doi.org/10.1016/j.jocs.2017.11.006.

[3] P. Kalarani and S. Selva Brunda, "Sentiment analysis by POS and joint sentiment topic features using SVM and ANN," *Soft Computing*, vol. 23, no. 16, pp. 7067–7079, Aug. 2019, https://doi.org/10.1007/s00500-018-3349-9.

[4] N. C. Hellwig, J. Fehle, and C. Wolff, "Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings," *Expert Systems with Applications*, vol. 261, Feb. 2025, Art. no. 125514, https://doi.org/10.1016/j.eswa.2024.125514.

[5] W. Maharani, D. H. Widyantoro, and M. L. Khodra, "Aspect Extraction in Customer Reviews Using Syntactic Pattern," *Procedia Computer Science*, vol. 59, pp. 244–253, 2015, https://doi.org/10.1016/j.procs.2015.07.545.

[6] C. Liao, C. Feng, S. Yang, and H. Huang, "Topic-related Chinese message sentiment analysis," *Neurocomputing*, vol. 210, pp. 237–246, Oct. 2016, https://doi.org/10.1016/j.neucom.2016.01.110.

[7] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, pp. 139–147, Mar. 2019, https://doi.org/10.1016/j.eswa.2018.08.044.

[8] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," *Expert Systems with Applications*, vol. 148, Jun. 2020, Art. no. 113234, https://doi.org/10.1016/j.eswa.2020.113234.

[9] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, May 2004, pp. 168–177, https://doi.org/10.1145/1014052.1014073.

[10] T. Sokhin and N. Butakov, "Semi-automatic sentiment analysis based on topic modeling," *Procedia Computer Science*, vol. 136, pp. 284–292, 2018, https://doi.org/10.1016/j.procs.2018.08.286.

[11] M. Zaheer, A. Ahmed, and A. J. Smola, "Latent LSTM Allocation: Joint Clustering and Non-Linear Dynamic Modeling of Sequence Data," in *Proceedings of the 34th International Conference on Machine Learning*, Jul. 2017, pp. 3967–3976.

[12] Y. Jo, L. Lee, and S. Palaskar, "Combining LSTM and Latent Topic Modeling for Mortality Prediction." arXiv, 2017, https://doi.org/10.48550/ARXIV.1709.02842.

[13] P. Jansson and S. Liu, "Topic modelling enriched LSTM models for the detection of novel and emerging named entities from social media," in *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, Dec. 2017, pp. 4329–4336, https://doi.org/10.1109/BigData.2017.8258462.

[14] G. Pergola, L. Gui, and Y. He, "TDAM: A topic-dependent attention model for sentiment analysis," *Information Processing & Management*,

vol. 56, no. 6, Nov. 2019, Art. no. 102084, https://doi.org/10.1016/j.ipm.2019.102084.

[15] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10849–10855, Jun. 2023, https://doi.org/10.48084/etasr.5854.

[16] W. Li and A. McCallum, "Pachinko Allocation: Scalable Mixture Models of Topic Correlations," *Journal of Machine Learning Research*, 2008.

[17] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, Feb. 2010.

[18] "515K Hotel Reviews Data in Europe - Skyhighes Technologies," Jan. 16, 2025. https://skyhighes.in/courses/data-visualization-projects/lesson/515k-hotel-reviews-data-in-europe/.

[19] "Download Amazon Review Dataset Of Mobile Phones," *DataStock*. https://datastock.shop/download-amazon-mobile-phone-reviews-dataset/.

[20] T. Liesting, F. Frasincar, and M. M. Trușcă, "Data augmentation in a hybrid approach for aspect-based sentiment analysis," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, Dec. 2021, pp. 828–835, https://doi.org/10.1145/3412841.3441958.

[21] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–30, Mar. 2021, https://doi.org/10.1155/2021/6619088.

[22] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. 66, pp. 245–260, Dec. 2016, https://doi.org/10.1016/j.eswa.2016.09.009.

[23] P. Asthana, M. Barnwal, A. Yadav, M. Aggrawal, and M. Goel, "VADER: A Lightweight and Effective Approach for Sentiment Analysis," in *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, Nov. 2024, pp. 687–692, https://doi.org/10.1109/icaiccit64383.2024.10912371.

[24] K. Tayal, A. Mehta, J. Kamboj, and S. Susan, "Fuzzy Aggregation of Polarity Scores for Unsupervised Sentiment Analysis Using AFINN, VADER and TextBlob," in *Bio-Inspired Computing*, 2025, pp. 47–54, https://doi.org/10.1007/978-3-031-78943-4_6.