Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Deep Learning and Imbalance Handling on Movie Review Sentiment Analysis

Sri Utami[1)], Kemas Muslim Lhaksmana[2)*], Yuliant Sibaroni[3)]
[1,2,3)] School of Computing, Study Program of Informatics, Telkom University, Bandung, Indonesia
[1)]sriutami@students.telkomuniversity.ac.id, [2)]kemasmuslim@telkomuniversity.ac.id,
[3)]yuliant@telkomuniversity.ac.id

Submitted : Jul 24, 2023 | **Accepted** : Jul 26, 2023 | **Published** : Jul 31, 2023

Abstract: Before watching a movie, people usually read reviews written by movie critics or regular audiences to gain insights about the movie's quality and discover recommended films. However, analyzing movie reviews can be challenging due to several reasons. Firstly, popular movies can receive hundreds of reviews, each comprising several paragraphs, making it time-consuming and effort-intensive to read them all. Secondly, different reviews may express varying opinions about the movie, making it difficult to draw definitive conclusions. To address these challenges, sentiment analysis using CNN and LSTM models, known for their effectiveness in classifying text in various datasets, was performed on the movie reviews. Additionally, techniques such as TF-IDF, Word2Vec, and data balancing with SMOTEN were applied to enhance the analysis. The CNN achieved an impressive sentiment analysis accuracy of 98.56%, while the LSTM achieved a close 98.53%. Moreover, both classifiers performed well in terms of the F1-score, with CNN obtaining 77.87% and LSTM achieving 78.92%. These results demonstrate the effectiveness of the sentiment analysis approach in extracting valuable insights from movie reviews and helping people make informed decisions about which movies to watch.

Keywords: CNN; LSTM; movie review; sentiment analysis; SMOTEN

INTRODUCTION

In today's digital era, everyone can access information easily using the internet. The Internet is unique because it combines two diverse communication methods and several types of content into one medium (DiMaggio, Hargittai, Russell Neuman, & Robinson, 2001). Before watching a movie, someone will look for reviews about the movie to watch. The internet is used to share these reviews, one of the platforms that provides a place to share movie reviews is Rotten Tomatoes. Information can be processed using sentiment analysis on the basis of these evaluations.

Sentiment Analysis is the analysis of a person's opinions, attitudes, and emotions using a computer. The objective of sentiment analysis is to discover opinions, identify the emotions conveyed by a person, and categorize the class of opinions (Medhat, Hassan, & Korashy, 2014). The reviews on movie review websites can be used as a resource for movie enthusiasts to discover movie recommendations as a tool for film producers to determine the audience's reaction to recently released film (Nurdiansyah, Bukhori, & Hidayat, 2018).

In recent years, a number of studies have proposed deep learning-based sentiment analysis with varying characteristics and efficacy (Dang, Moreno-García, & De la Prieta, 2020). In pattern recognition and computer vision, deep-learning neural networks have made remarkable progress. Several sophisticated deep learning algorithms, including sentiment analysis, were introduced over time to accomplish complex NLP-related tasks (Ombabi, Ouarda, & Alimi, 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.12770 p-ISSN : 2541-044X

Based on the research of (Zhang, Wang, & Liu, 2018) on the basis of the development of deep learning research and applications, it is anticipated that deep learning research on sentiment analysis will shortly emerge. Before the classification process is performed, the data that has been obtained will be processed first at the preprocessing stage and continued with the use of feature extraction. TF-IDF is the feature extraction method used. In addition, Word2Vec supports expansion and data balancing via SMOTEN. This classification procedure's ultimate objective is to produce an accuracy value. In this study, accuracy and F1-score were used to evaluate the performance of the two classification models.

The objective of this study is to find out the best classification method, between CNN and LSTM, in sentiment classification on an English dataset obtained from the Rotten Tomatoes website. The classification of sentiment is divided into two classes: positive classes and negative classes. The larger the dataset used, combined with the proposed deep CNN and its training strategy, will result in the better ability of the model to generalize, and increase confidence in the generalization (Ouyang, Zhou, Li, & Liu, 2015). LSTMs has proven to be an invaluable tool in learning sequence modeling tasks involving unknown lengths. Its main advantage lies in its ability to store long-term memory (Chen, Lee, & Chen, 2020). TF-IDF represents term frequency-inverse document frequency. This technique is among the most widely used in the fields of information retrieval and text analysis. TF-IDF is a metric for determining the significance of a document's words (Avinash & Sivasankar, 2019). Word2Vec can provide an accurate estimation of the meaning of a word given sufficient data, usage, and content. One important advantage of Word2Vec is its high speed even when used with large datasets. In the context of deep learning, word meaning becomes a simple signal that is useful in classifying more complex entities (Ouyang et al., 2015). SMOTEN is a variant of the SMOTE algorithm that combines oversampling with data deletion. In addition to oversampling the minority class, this method also eliminates instances or data adjacent to the minority class. The primary objective of this method is to identify and eliminate the majority class's adjacent neighbors before oversampling the minority class, thereby removing the most frequent data before balancing the underrepresented class (Alabrah, 2023).

LITERATURE REVIEW

Deep learning-based sentiment analysis has been the subject of numerous previous studies. It is anticipated that research on sentiment analysis using deep learning will shortly advance (Zhang et al., 2018). Document-level, Sentence-level, Aspect-based, Comparative, and Sentiment Lexicon Acquisition are the five categories of sentiment analysis (Feldman, 2013). When there are two basics classes, positive and negative, at the Document-level where documents must be classified and informed, there is training for each class (N Murthy, Rao Allu, Andhavarapu, Bagadi, & Belusonti, n.d.). CNN, LSTM, RNN, and GRU are a few examples of models that employ deep learning.

LSTM is also widely employed in natural language processing (Rhanoui, Mikram, Yousfi, & Barzali, 2019). In research (Hidayatullah, Abida, & Nayoan, n.d.) obtained the highest accuracy value of 90.85% by using the CNN method. Kim (Moschitti, 2014), using an unsupervised model, a simple CNN with a single convolutional layer was deployed on multiple databases, including movie reviews, STS, subjectivity datasets, TREC surveys, and customer reviews. By utilizing modest hyperparameters that yield robust outcomes.

Whereas in (Widayat, 2021), the obtained accuracy findings are at least 85.86%, indicating that the sentiment classification is quite accurate. In research (Sosa, 2017), Pedro M. Sosa combined CNN and LSTM to achieve better performance on sentiment analysis by a significant margin.

LSTM is a type of neural network for processing and predicting significant events with comparatively long intervals and time series delays (Jin, Yang, & Liu, 2020). Then in research (N Murthy et al., n.d.), utilizing an IMDB dataset containing a total of 50,000 evaluations, with 25,000 positive and 25,000 negative classes. With a larger quantity of training data, deep learning techniques such as LSTM can classify sentiments with an accuracy of 85%.

Comparison on CNN, LSTM, and BERT architectures have been carried out in (Colón-Ruiz & Segura-Bedmar, 2020). This study divides subjects into three classes: positive, negative, and neutral. But for a more complex procedure involving 10 categories, which is the aggregate level user satisfaction as determined by drug review's users, see below. As anticipated, the results of the 3-class data are

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.12770 p-ISSN : 2541-044X

significantly higher than those 10-category data. On 10-category data, the hybrid model composed of bidirectional LSTM and CNN yields the best results.

In (Kabra & Nagar, 2023) using CNN and TF-IDF, the accuracy obtained reached 87%. The overall accuracy of the proposed technique is higher by 12% compared to the latest existing methods. Furthermore, in research (Hermanto, Setyanto, & Luthfi, 2021), where LSTM, LSTM-CNN, and CNN-LSTM are added with the use of Word2Vec 51%, 53%, and 62%. The data used is 1200 data on the sentiment of online media classification. By using four split data ratios of 80:20, 70:30, 60:40, and 50:50 in research (Alabrah, 2023). By adding SMOTEN as a data imbalance treatment, the accuracy results reached 99.97% by using one of the split data with ratios of 80:20.

METHOD

The system begins by gathering data from the Rotten Tomatoes website. The data has been classified with positive and negative labels. Figure 1 shows the implementation flow for the built system, where the data undergoes a preprocessing phase in which SMOTEN is used to balance asymmetrical data. Next, the TF-IDF method is used to extract the features. Word2Vec is used to amplify the capabilities. CNN and LSTM models are utilized for data processing. Finally, the system's efficacy is assessed using a confusion matrix.

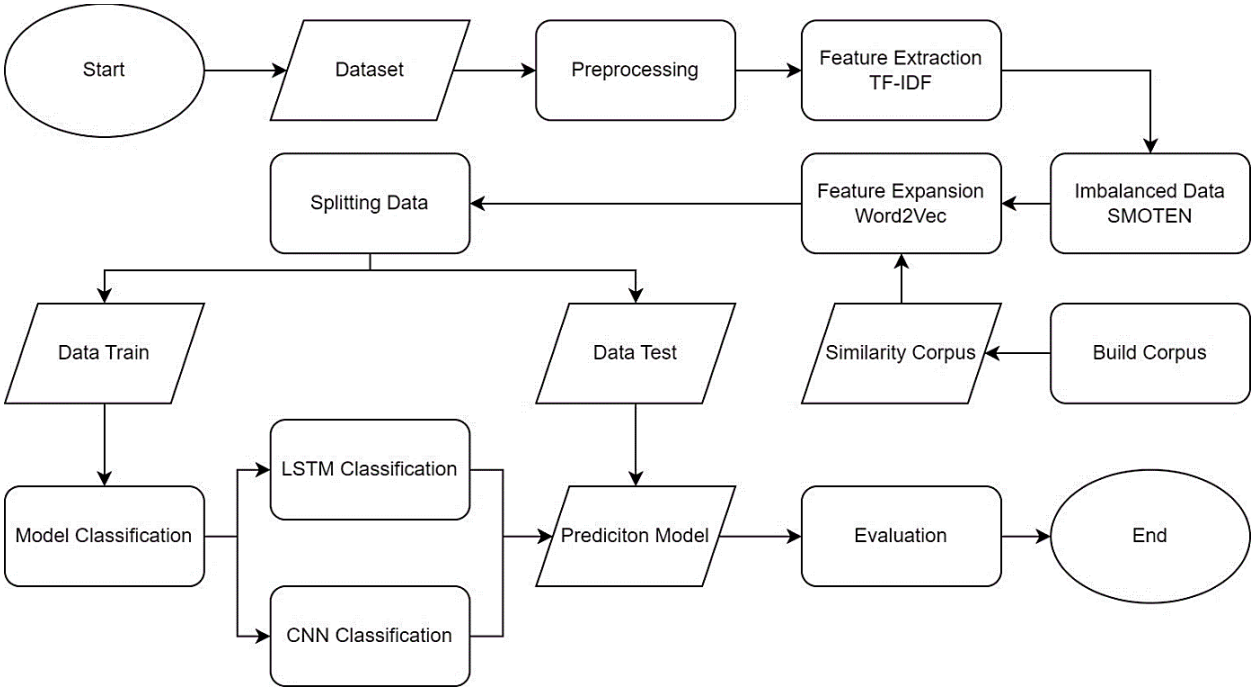

Figure 1. Flowchart of The Implementation of Classification and Evaluation

Dataset

The data used in this study comes from the Rotten Tomatoes website, which is available as an Excel file on Kaggle website (Stefano Leone, 2020). The overall quantity of data was 1,130,117 rows. Because running the entire dataset beginning in 2010 would require a very long time, the dataset used begins in 2020 and contains 45,739 rows of data.

There are two sentiment classes from the audience reviews on the Review_content attribute that related to the Review_type attribute in the dataset. This can be seen in Table 1, where the Rotten sentiment class corresponds to negative classes, whereas the Fresh sentiment class corresponds to positive classes. Where there are 29,011 data points for positive classes and 11,755 data points for negative classes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Sample Data

Review_content	Review_type
The film is stronger in its more sincere moments and weaker where Baumbach tries to push supposed truisms through gurning conceits (Stefano Leone, 2020).	Rotten
What's most frustrating about Knives and Skin is despite all of these wild and weird ingredients, there is simply no connective tissue to keep everything together. Reeder's vision is frustratingly opaque (Stefano Leone, 2020).	Fresh

Preprocessing

Preprocessing is the procedure of removing undesirable data elements. It increases the precision of the results by reducing data errors. Because without preprocessing, such as grammar correction, the system may disregard essential words (Altrabsheh, Cocea, & Fallahkhair, 2014).

Data Cleaning

The first step in the preprocessing phase is data cleansing, the removal of unnecessary characters during the classification process will take place.

Case Folding

Case folding is the process of converting all uppercase characters in a text to lowercase letters (Sari & Ruldeviyani, 2020). The text no longer contains any capital letters. Table 2 depicts an example of the case folding procedure.

Table 2. Example of Case Folding Process

Before	After
The film is stronger in it is more sincere moments and weaker where Baumbach tries to push supposed truisms through gurning conceits (Stefano Leone, 2020)	the film is stronger in it is more sincere moments and weaker where baumbach tries to push supposed truisms through gurning conceits

Tokenization

Tokenization means identifying the base word by segmenting the text into sentences and words (Yasen & Tedmori, 2019). Table 3 is an example of the tokenization process.

Table 3. Example of Tokenization Process

Before	After
the film is stronger in it is more sincere moments and weaker where baumbach tries to push supposed truisms through gurning conceits (Stefano Leone, 2020)	[‘the’, ‘film’, ‘is’, ‘stronger’, ‘in’, ‘it’, ‘is’, ‘more’, ‘sincere’, ‘moments’, ‘and’, ‘weaker’, ‘where’, ‘baumbach’, ‘tries’, ‘to’, ‘push’, ‘supposed’, ‘truisms’, ‘through’, ‘gurning’, ‘conceits’]

Stop Word Removal

Stop word removal is the procedure of removing frequent but unimportant words (Buttar, Kaur, & Kaur Buttar, 2018). Removing stop words is useful for reducing vector space, and improving execution performance speed, computation, and accuracy. Example of use of deleted words such as ‘the’, ‘is’, ‘in’, and others can be seen in the Table 4 (‘NLTK’s list of english stopwords’, 2010).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. Example of Stop Word Removal

Before	After
[‘the’, ‘film’, ‘is’, ‘stronger’, ‘in’, ‘it’, ‘is’, ‘more’, ‘sincere’, ‘moments’, ‘and’, ‘weaker’, ‘where’, ‘baumbach’, ‘tries’, ‘to’, ‘push’, ‘supposed’, ‘truisms’, ‘through’, ‘gurning’, ‘conceits’] (Stefano Leone, 2020)	['film', 'stronger', 'moments', 'weaker', 'baumbach', 'push', 'supposed', 'truisms', 'gurning', 'conceits']

Stemming

The stemming process converts a word into a root word. The famous English stemmer is Porter. Porter Stemmer is a rule-based stemmer. To convert tokens into base words, it needs to remove ‘ing’, ‘ed’, ‘s’, ‘ly’, and others (Gharatkar, Ingle, Naik, & Save, 2018). Examples are in Table 5.

Table 5. Example of Stemming Process

Before	After
['film', 'stronger', 'moments', 'weaker', 'baumbach', 'push', 'supposed', 'truisms', 'gurning', 'conceits']	['film', 'strong', 'moment', 'weak', 'baumbach', 'push', 'suppose', 'truism', 'gurn', 'conceit']

Data Balancing

SMOTE is a synthetic technique for systematically oversampling a given dataset. By incorporating the minority class data into a large number of new optimal examples. SMOTEN is a hybrid form of the SMOTE technique. In addition to oversampling, cases or data that are close to the majority category are eliminated. Before oversampling the minority class, thus excluding the most extensive data (Alabrah, 2023). In this study, the parameter for SMOTEN is random state none, and the random number generator is an instance of RandomState used by np.random (‘SMOTEN — Version 0.11.0’, n.d.).

Feature Extraction

The objective of feature extraction is to present words in a vector format. This feature extraction employs TF-IDF. TF-IDF is an effective method for identifying significant terms in a vector. The TF (Term Frequency) of a given term (t) is calculated as the number of occurrences of the term relative to the total number of words in the document (Ahuja, Chug, Kohli, Gupta, & Ahuja, 2019). In addition, IDF (Inverse Document Frequency), where N is the total number of documents and DF is the total number of documents containing the term (t).

$$IDF(t) = \log\frac{D}{df} \quad (1)$$

Therefore, the TF-IDF formula is as follows:

$$TF - IDF = TF(t) \times IDF(t) \quad (2)$$

Feature Expansion

Word2Vec is a machine learning-based tool for calculating word vector similarity. It converts words into word vectors and calculates the cosine similarity of word vectors (Pan et al., 2019). Figure 2 illustrates the two training model types available in Word2Vec: CBOW (Continuous Bag of Words) and Skip-Gram. Can be seen that the input layer, projection layer, and output layer are present in both models. The CBOW model predicts the current words based on the known context, while the Skip-gram model predicts the context based on the current words. CBOW and Skip-Gram are merged in this study.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

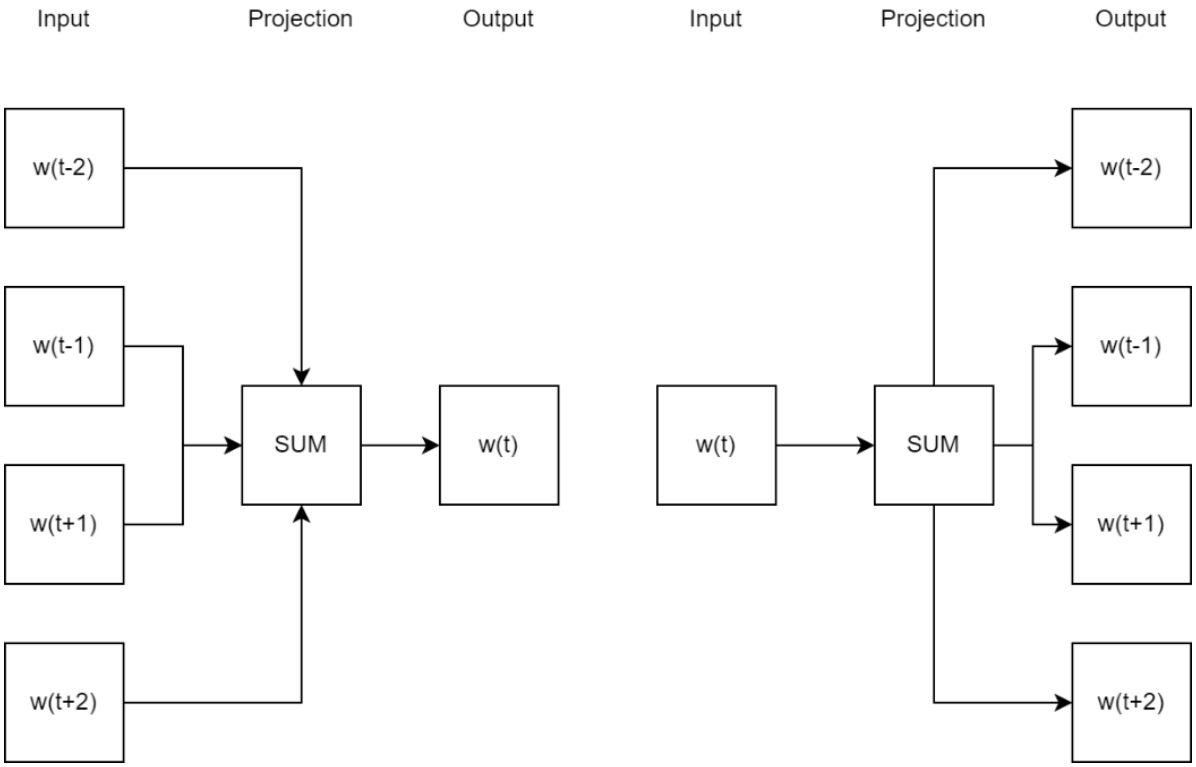

Figure 2. CBOW Model and Skip-Gram Architecture (Pan et al., 2019)

The process of using Word2Vec feature expansion requires a corpus. Initially, a corpus was created using dataset of movie reviews from the website Rotten Tomatoes. Table 6 displays the collected corpus information.

Table 6. Total Corpus Word2Vec

Corpus	Total
Film Reviews	15356

CNN Classification

CNN is a deep learning technique. CNN operates by transferring sentences into matrices, where each column represents a word vector. If the sentences length is s, the dimension of the sentence matrix is s x d (Liao, Wang, Yu, Sato, & Cheng, 2017). It can be seen in Figure 3, the application of the pooling function to each feature map results in a vector length that is fixed. Utilizing the MaxPooling1D method, information is extracted from the feature map. In the final, sigmoid, layer, dropout is utilized to attempt regularization. This study's CNN model parameters include the number of unique terms in dataset's vocabulary and the dimension of the embedding vector used to represent each word in movie reviews.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

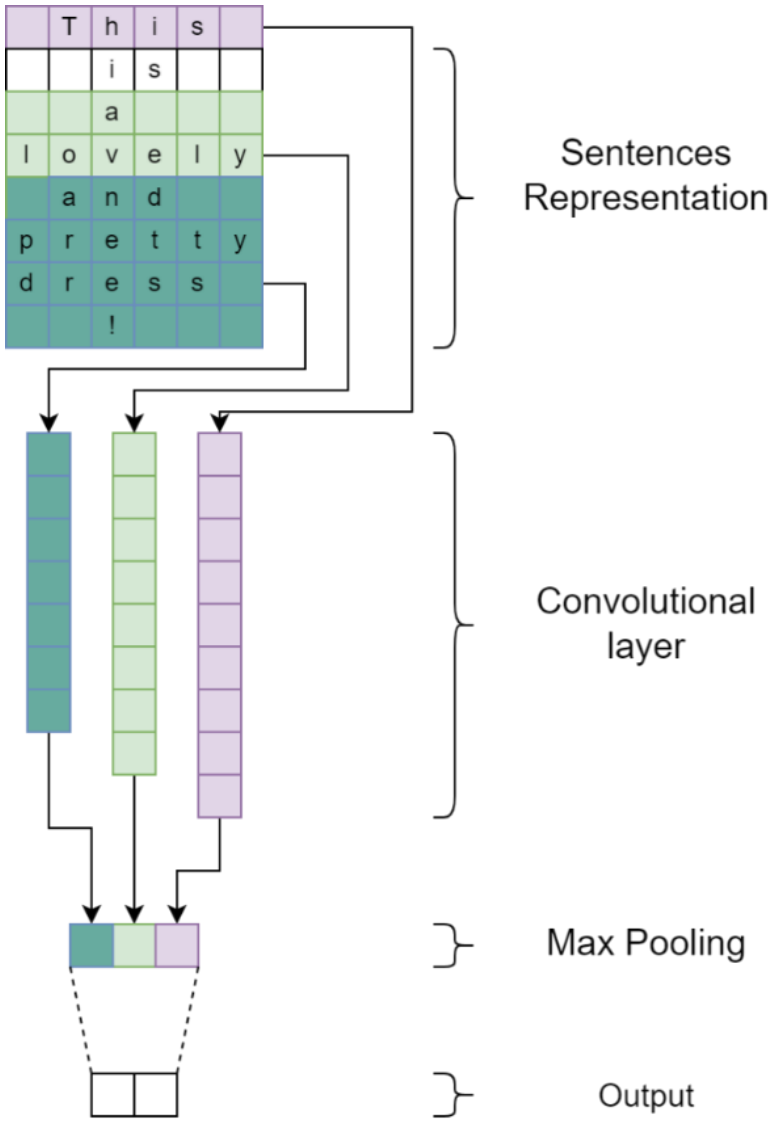

Figure 3. CNN Architecture (Liao et al., 2017)

LSTM Classification

Long-Shor Term Memory (LSTM) is an additional method of deep learning utilized in this study. There are several crucial stages in LSTM. The initial layer is a "forget gate" the determines which information will be removed from the cell state. The next layer is an "input gate" that determines what new information to store in the cell state. The previous cell state is then altered by combining the retained information with the newly generated information. The output of the LSTM is then filtered to generate relevant data (Olah, 2015).

In language models, LSTM can be used to anticipate the next word based on the previous words. For example, in Figure 4. LSTM can retain information about the current subject and remove new information about verbs that may appear next. In this way, LSTM can produce relevant outputs that are useful in building accurate language models. Similarly, to the CNN model, the LSTM model parameters in this study include the number of unique terms in the dataset's vocabulary and the use of the embedding vector dimension to represent each word in movie reviews.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

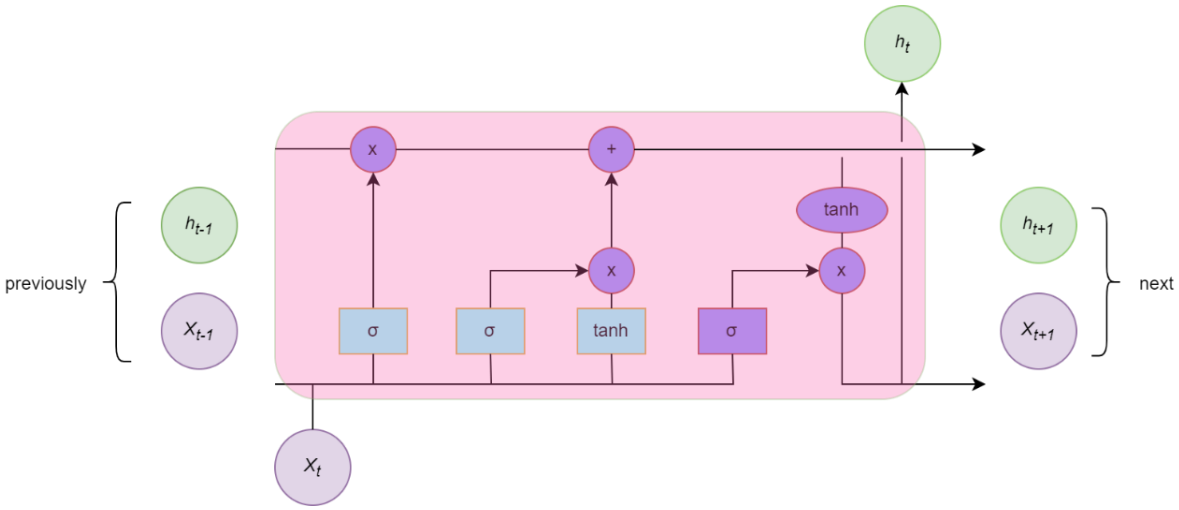

Figure 4. LSTM Architecture (Olah, 2015)

Evaluation

A performance evaluation that compares two methodologies using a confusion matrix. If both methods yield a high score degree of accuracy, then the classification performed was successful. To determine the accuracy of each class (positive and negative), the total number of opinions is divided by the total quantity of data. In addition to the accuracy rate, precision, recall, and F1-score must be calculated (Suhariyanto, Firmanto, & Sarno, 2018). Table 7 shows an example of a confusion matrix.

Table 7. Confusion Matrix

Confusion Matrix		**Actual Value**	
		Positive	**Negative**
Prediction Value	Positive	TP	FP
	Negative	FN	TN

Table 7's terminology is as follows:
True Positive (TP) : The prediction is accurate and corresponds to the actual situation.
True Negative (TN) : Prediction and reality of a negative outcome.
False Positive (FP) : A positive prediction and a fictitious actuality.
False Negative (FN) : A negative prognosis and the incorrect actual circumstance.

The performance value can then be computed using the following formulas:

Accuracy

Accuracy is a metric that describes the proportion of true predicted data relative to the total data.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

Precision

Precision is the degree to which the system's response precisely matches the two requested items of information.

$$precision = \frac{TP}{TP+FP} \quad (4)$$

Recall

Recall is a measure that shows how good the system is at finding relevant answers.

$$recall = \frac{TP}{TP+FN} \quad (5)$$

F1-score

F1-score is value that describes the average of the comparison between precision and recall.

$$F1\ score = \frac{2 \times (recall \times precision)}{(recall+precision)} \quad (6)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

In this research evaluation, a comparison of data was conducted between the two best-performing models in a number of test scenarios. Four test situations were conducted. The initial step involves comparing the data split ratio in the baseline model, with the goal of achieving optimal outcomes for each classification model, which will be carried forward to the subsequent scenario. In the second scenario, the focus is on addressing data imbalance using SMOTEN, aiming to balance the positive and negative classes and enhance the overall performance of both classification models. Moving on to scenario three, TF-IDF is employed not only to transform words into vectors but also to optimize the model for the forthcoming scenarios. Lastly, Word2Vec is utilized for testing, enabling the measurement of word vector similarities.

Scenario 1

In this research evaluation, a comparison of data between the two top-performing models in a variety of test scenarios was conducted. The first involves establishing the split data ratios for CNN and LSTM models by comparing the values of the split data ratio.

The best results from scenario 1 are in Table 8 testing are a 90:10 ratio, with CNN producing accuracy score of 76.46% and F1-score score of 71.14% and LSTM producing an accuracy value of76.21% and an F1-score value of 71.00%. This optimal ratio value will serve as benchmark for future testing.

Table 8. Results of Comparing Split Data Ratios

Test Sizes	CNN		LSTM	
	Accuracy	F1-score	Accuracy	F1-score
0.1	**76.46%**	**71.14%**	**76.21%**	**71.10%**
0.2	76.38%	70.65%	76.00%	71.06%
0.3	76.10%	69.12%	76.24%	70.12%

Scenario 2

Scenario 2 evaluates how SMOTEN handles unbalanced data. As seen in Table 9, the accuracy and F1-score values have significantly increased.

Table 9. Results of Using SMOTEN

Model	CNN		LSTM	
	Accuracy	F1-score	Accuracy	F1-score
Without SMOTEN	76.46%	71.14%	76.21%	71.10%
With SMOTEN	**88.39%**	**81.93%**	**87.21%**	**81.53%**

Table 10. Confusion Matrix of CNN Using SMOTEN

Confusion Matrix of CNN		Actual Value	
		Positive	Negative
Prediction Value	Positive	407	443
	Negative	316	1484

CNNs at Table 10 model has produced positive results in identifying positive and negative film reviews, with relatively high True Positive and True Negative rates. However, it is crucial to be aware of the model's errors in predicting positive reviews as negative (False Positive) and negative reviews as positive (False Negative). Further evaluation and model enhancement may be required to reduce these errors and improve the model's overall performance in film review sentiment analysis.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 11. Confusion Matrix of LSTM Using SMOTEN

Confusion Matrix Of LSTM		Actual Value	
		Positive	Negative
Prediction Value	Positive	434	416
	Negative	425	1375

The LSTM model has produced reasonably accurate results at Table 11 in identifying positive and negative film reviews, with relatively high True Positive and True Negative rates. However, it is essential to note the model's errors in predicting reviews that are genuinely negative as positive (False Positive) and positive reviews as negative (False Negative). Further evaluation and model enhancement may be required to reduce these errors and improve the model's overall performance in film review sentiment analysis.

Scenario 3

In scenario 3, TF-IDF feature extraction is also implemented. Tests are conducted by comparing the maximum feature value. The maximum feature value comparison is divided into 1000 and 5000.

Table 12. Results of Comparing Max Features TF-IDF

Max Features	CNN		LSTM	
	Accuracy	F1-score	Accuracy	F1-score
1000	94.42%	85.56%	94.46%	85.80%
5000	**99.15%**	**87.29%**	**99.13%**	**87.09%**

Table 12 demonstrates that the optimal maximum feature value is 5000. The CNN method has an accuracy value of 99.15% and an F1-Score value of 87.29%, while the LSTM method has an accuracy value of 99.13% and an F1-Score value of 87.09%. The optimal value of maximum features will then be utilized for subsequent testing.

Scenario 4

In scenario 4, the Word2Vec feature augmentation was applied to the previously generated corpus of movie review text. There is a collection of terms with similarity values in the corpus. This test compares the Top 5 and Top 10 values and the result seen at Table 13.

Table 13. Results of Comparing Word2Vec Top Values

Top	CNN		LSTM	
	Accuracy	F1-score	Accuracy	F1-score
5	**98.56%**	**77.87%**	**98.53%**	**78.92%**
10	98.21%	70.39%	98.41%	74.92%

Using the Top 5 value, the accuracy value for the CNN method is 98.56% and the F1-Score value is 77.87%, while the accuracy value for the LSTM method is 98.5% and the F1-Score value is 78.92%.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12770
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

Four different scenarios were conducted by comparing two different classification models namely CNN and LSTM with the application of TF-IDF, Word2Vec, and SMOTEN. Accuracy and F1-score values for each scenario were determined using the average of the five evaluations.

Table 14. Results of CNN and LSTM Comparison

Scenario Test	CNN		LSTM	
	Accuracy	F1-score	Accuracy	F1-score
Baseline	**76.46%**	**71.14%**	76.21%	71.10%
SMOTEN	**88.39%**	**81.93%**	87.21%	81.53%
TF-IDF	**99.15%**	**87.29%**	99.13%	87.09%
Word2Vec	**98.56%**	77.87%	98.53%	**78.92%**

Can be seen in table 14, multiple strategies were employed to achieve the finest CNN model performance results. First, a baseline ratio of 90:10 divides the train data into 90% train data and 10% test data. Second, feature extraction is carried out using the TF-IDF algorithm and only the maximum feature with a value of 5000 is utilized. Thirdly, the Word2Vec model and the top five values from the Movie Review corpus are used to conduct feature expansion. Lastly, the SMOTEN technique is utilized to address data imbalance. In the end, a very excellent performance value was obtained, with an accuracy of 98.56% and an F1-score of 77.87% respectively.

Multiple strategies are utilized to achieve the finest LSTM model performance results. First, similar to the LSTM model, the baseline implements a 90:10 ratio, which divides train data into 90% and test data into 10%. As with the CNN model test, feature extraction is undertaken using the TF-IDF method with a maximum of 5000 features. Thirdly, the Word2Vec model is used to perform feature expansion using only the top five values from the Movie Review corpus. Lastly, the SMOTEN technique is utilized to address data imbalance. In the end, outstanding performance results were obtained, including an accuracy of 98.53% and an F1-score of 78.92%.

Furthermore, a comparison was made with research (Ardhian Fahmi Sabani, Adiwijaya, & Widi Astuti, 2022) where the research used the same dataset (Stefano Leone, 2020) as the current research. In research (Ardhian Fahmi Sabani et al., 2022) SVM and Word2Vec techniques were utilized for sentiment analysis. Using the RBF kernel and K-Fold, with an accuracy value of 79.0% and an F1-Score value of 70.2%, the most accurate model obtained by conducting several experiments can be seen in Table 15.

Table 15. Results of Comparison Between CNN, LSTM, and SVM

Dataset	CNN		LSTM		SVM	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Review Film from Rotten Tomatoes	**98.56%**	**77.87 %**	**98.53%**	**78.92 %**	79.00%	72.00 %

This study's accuracy and F1-Score values are affected by the use of TF-IDF for feature extraction, the management of data imbalance using SMOTEN and deep learning classification models including CNN and LSTM.

*name of corresponding author

CONCLUSION

This study compares the accuracy of Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) classification methods on film review sentiment analysis. The CNN method has the highest accuracy at 98.56%, while the LSTM method has the highest accuracy at 98.53%. The LSTM method has the highest F1-Score at 78.92%, while CNN's F1-Score is 77.87%. The objective of this study is to determine which of the two classification methods, CNN and LSTM, produces the best results when applied to English datasets obtained from the Rotten Tomatoes website. The larger the dataset used, combined with the proposed deep CNN and its training strategy, will result in the better ability of the model to generalize and increase confidence in the generalization. TF-IDF stands for term frequency-inverse document frequency, a commonly employed technique in information retrieval and text analysis. Word2Vec provides an accurate estimation of the meaning of a word given sufficient data, usage, and context. SMOTEN is a variant of the SMOTE algorithm that combines oversampling with data deletion, aiming to identify and eliminate the majority class's adjacent neighbors before oversampling the minority class. Among the two classification models, the CNN model achieves the highest accuracy when using a baseline ratio of 90:10, TF-IDF feature extraction with a maximum of 5000 features, Word2Vec feature expansion with the Top 5 words on the Movie Review corpus, and SMOTEN for addressing imbalanced data. The LSTM model has the highest F1-Score value. Utilizing additional models for feature expansion and feature extraction is a suggestion for future research. Additionally, it can perform variations by comparing other models.

REFERENCES

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). *The impact of features extraction on the sentiment analysis*. In *Procedia Computer Science* (Vol. 152). doi:10.1016/j.procs.2019.05.008.

Alabrah, A. (2023). An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method. *Sensors (Basel, Switzerland)*, *23*(9). doi:10.3390/s23094406.

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). *Sentiment Analysis: Towards a Tool for Analysing Real-Time Students Feedback*. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* (Vol. 2014-December). doi:10.1109/ICTAI.2014.70.

Ardhian Fahmi Sabani, Adiwijaya, & Widi Astuti. (2022). *Analisis Sentimen Review Film pada Website Rotten Tomatoes Menggunakan Metode SVM Dengan Mengimplementasikan Fitur Extraction Word2Vec*. Retrieved from https://openlibrary.telkomuniversity.ac.id/home/catalog/id/178462/slug/analisis-sentimen-review-film-pada-website-rotten-tomatoes-menggunakan-metode-svm-dengan-mengimplementasikan-fitur-extraction-word2vec.html.

Avinash, M., & Sivasankar, E. (2019). *A study of feature extraction techniques for sentiment analysis*. In *Advances in Intelligent Systems and Computing* (Vol. 814). doi:10.1007/978-981-13-1501-5_41.

Buttar, P., Kaur, J., & Kaur Buttar, P. (2018). A Systematic Review on Stopword Removal Algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *4*(4).

Chen, L. C., Lee, C. M., & Chen, M. Y. (2020). Exploration of social media for sentiment analysis using deep learning. *Soft Computing*, *24*(11). doi:10.1007/s00500-019-04402-8.

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*. doi:10.1016/j.jbi.2020.103539.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics (Switzerland)*, *9*(3). doi:10.3390/electronics9030483.

DiMaggio, P., Hargittai, E., Russell Neuman, W., & Robinson, J. P. (2001). Social implications of the internet. *Annual Review of Sociology*, *27*. doi:10.1146/annurev.soc.27.1.307.

EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. (2014). *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4). doi:10.1145/2436256.2436274.

*name of corresponding author

Gharatkar, S., Ingle, A., Naik, T., & Save, A. (2018). *Review preprocessing using data cleaning and stemming technique*. In *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017* (Vol. 2018-January). doi:10.1109/ICIIECS.2017.8276011.

Hermanto, D. T., Setyanto, A., & Luthfi, E. T. (2021). Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online. *Creative Information Technology Journal*, *8*(1). doi:10.24076/citec.2021v8i1.264.

Hidayatullah, A. F., Abida, R., & Nayoan, N. (n.d.). *Analisis Sentimen Berbasis Fitur pada Ulasan Tempat Wisata Menggunakan Metode Convolutional Neural Network(CNN)*. Retrieved from www.cnet.com.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, *32*(13). doi:10.1007/s00521-019-04504-2.

Kabra, B., & Nagar, C. (2023). Convolutional Neural Network based sentiment analysis with TF-IDF based vectorization. *Journal of Integrated Science and Technology*, *11*(3).

Liao, S., Wang, J., Yu, R., Sato, K., & Cheng, Z. (2017). *CNN for situations understanding based on sentiment analysis of twitter data*. In *Procedia Computer Science* (Vol. 111). doi:10.1016/j.procs.2017.06.037.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4). doi:10.1016/j.asej.2014.04.011.

N Murthy, G. S., Rao Allu, S., Andhavarapu, B., Bagadi, M., & Belusonti, M. (n.d.). *Text based Sentiment Analysis using LSTM; Text based Sentiment Analysis using LSTM*. Retrieved from www.ijert.org.

NLTK's list of english stopwords. (2010, August).

Nurdiansyah, Y., Bukhori, S., & Hidayat, R. (2018). *Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method*. In *Journal of Physics: Conference Series* (Vol. 1008). doi:10.1088/1742-6596/1008/1/012011.

Olah, C. (2015). Understanding LSTM Networks [Blog]. *Web Page*.

Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, *10*(1). doi:10.1007/s13278-020-00668-1.

Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015). *Sentiment Analysis Using Convolutional Neural Network*. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 2359–2364). IEEE. doi:10.1109/CIT/IUCC/DASC/PICOM.2015.349.

Pan, Q., Dong, H., Wang, Y., Cai, Z., Zhang, L., & Nogueira, M. (2019). Recommendation of crowdsourcing tasks based on word2vec semantic tags. *Wireless Communications and Mobile Computing*, *2019*. doi:10.1155/2019/2121850.

Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A CNN-BiLSTM Model for Document-Level Sentiment Analysis. *Machine Learning and Knowledge Extraction*, *1*(3). doi:10.3390/make1030048.

Sari, I. C., & Ruldeviyani, Y. (2020). *Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers*. In *2020 International Workshop on Big Data and Information Security, IWBIS 2020*. doi:10.1109/IWBIS50925.2020.9255531

SMOTEN — Version 0.11.0. (n.d.).

Sosa, P. M. (2017). Twitter Sentiment Analysis using combined LSTM-CNN Models. *Eprint Arxiv*.

Stefano Leone. (2020). Rotten tomatoes movies and critic Reviews Dataset.

Suhariyanto, Firmanto, A., & Sarno, R. (2018). *Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet*. In *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*. doi:10.1109/ISEMANTIC.2018.8549704.

*name of corresponding author

Widayat, W. (2021). Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(3). doi:10.30865/mib.v5i3.3111.

Yasen, M., & Tedmori, S. (2019). *Movies reviews sentiment analysis and classification*. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*. doi:10.1109/JEEIT.2019.8717422.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). doi:10.1002/widm.1253.

*name of corresponding author