



Cross Domain Sentiment Analysis On Amazon Reviews

Nehal Parakhi¹ Arti Khaparde²

¹School Of Electronics and Communication, MIT World Peace University, Pune, Maharashtra

²School Of Electronics and Communication, MIT World Peace University, Pune, Maharashtra

Abstract— Cross domain sentiment analysis refers to divide sentiments of people into different categories. Such algorithms are used by different large scale corporations, hospitality departments, etc to understand feelings of the people towards their product or company as a whole.

Previously, hand written reviews were considered. But as the companies and online purchases are increasing, this method proved to be insufficient and tedious. Also the reviews had to be checked by a person, thus decreasing the efficiency and the accuracy of this method.

Our study aims at using different machine learning, deep learning algorithms to automate the process of product analysis.

The results prove that the study successfully provides an algorithm to understand and divide the sentiments of people into positive, negative and neutral.

Keywords – Sentiment Analysis, machine-learning, positive, negative, neutral.

I. INTRODUCTION

As mentioned in the previous sections sentiment analysis is a tool in machine learning and natural language processing used to identify, extract, quantify and study opinion of people. With increase in sell of online products and online services, this method is widely used to analyze and improve the products or services of an organization.

Applications of Sentiment Analysis are:

1. Social media monitoring.
2. Customer support
3. Customer feedback
4. Brand monitoring and reputation management.
5. Voice of customer (VoC)
6. Voice of employee
7. Product analysis
8. Market research and competitive research.

Two main approaches are used to detect the sentiment i.e Lexicon based and Machine learning based. Lexicon directly uses a dictionary to identify the words that describe the words related to sentiment of people. The review sentences for example are scanned and the words are searched through the dictionary to detect the polarity. This method is easy to apply but is very time consuming and takes lot of space. The words in the dictionary have to be of sufficient length to actually obtain the accuracy as desired. Another method used is of machine learning where the data is split into training and testing data. Accordingly the

machine is trained such that any new data provided will be automatically categorized and classified. Now a days many algorithms are used which combine both these methods taking the best of the both methods. The main drawback can be seen when the reviews are of different domains. Different words used to categorize the sentiment can have different meaning in different domains. The words “long running” can be used as positive in electronics domain, but the same can be a negative word in other domain. This reduces the accuracy of the algorithm to a high extent.

Figure 1 provides basic steps involved in sentiment classification. Steps involve data preprocessing, Vectorization, CNN algorithm and classification of data into positive, negative and neutral sentiments.

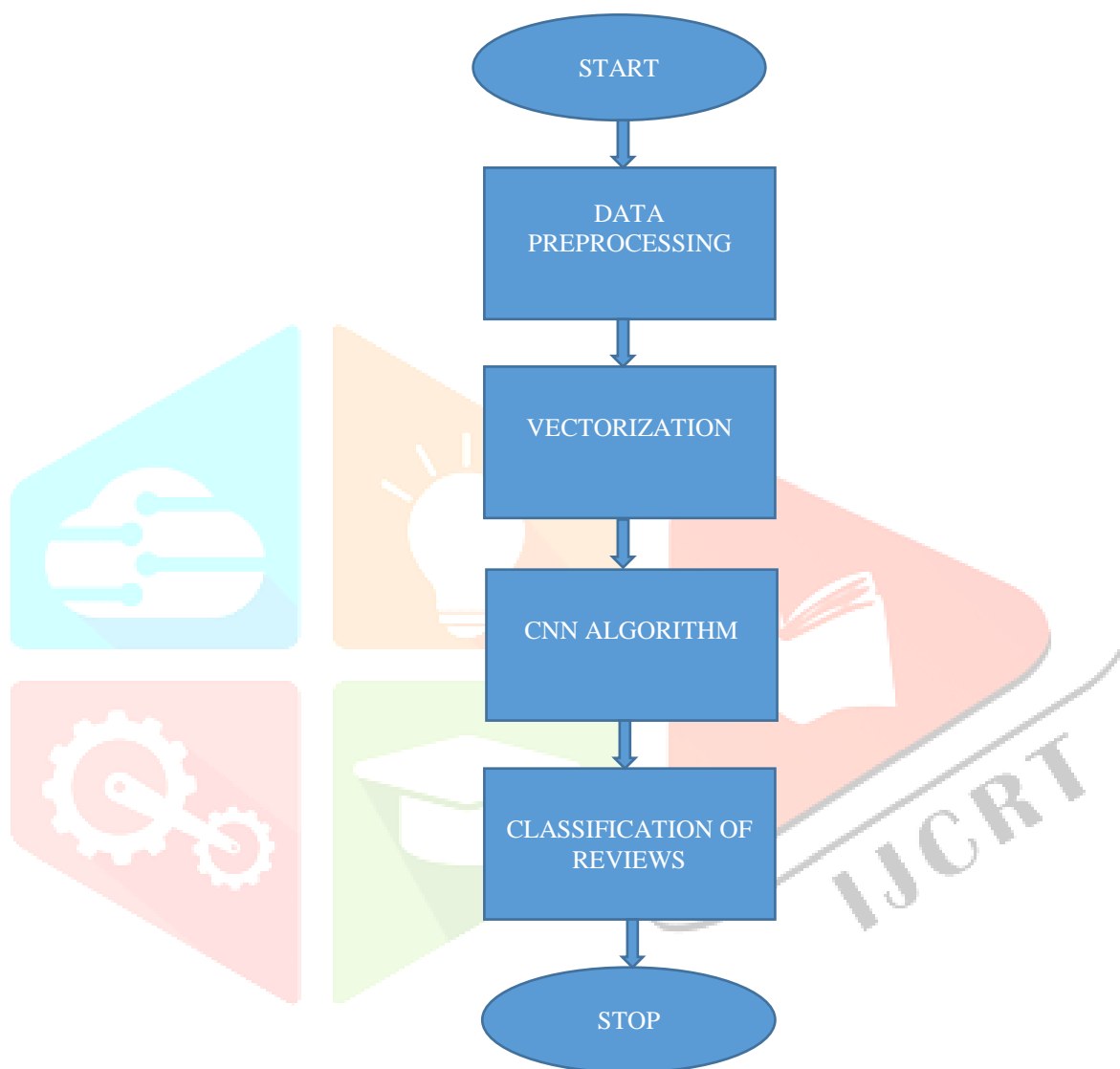


Figure 1: Basic Steps in Sentiment Analysis

The paper is divided into following sections. Next section includes basic literature survey. This section described and defines the algorithms that are being used in the study. These are obtained by reading and searching previous papers that are mentioned in the references in the last section. Next section allows detailed steps of the algorithm. It will also describe the convolutional neural network layers in detail. Further section will provide details of the experiment conducted with the results obtained in the study. Next sessions will include conclusion and future scope for the study.

II. LITERATURE SURVEY

Different authors have already provided various algorithms for sentiment analysis in cross domain. But there are certain disadvantages in the algorithm such as decrease in accuracy, execution time, etc. Considering all these, our aim is to provide algorithm for cross domain sentiment analysis using amazon reviews.

Certain parameters are used in the study that are defined and used in the previous papers. These parameters are described in this session below.

A. Sentiment Score

Let AR be the set of amazon reviews.

Each review ar consists of WO set of words. Each word 'wo' will contain synset SY. Let sy be a single synset

For $ar \in AR$: let WO be the words in ar.

For $sy \in SY$: let SY be set of synset of word wo.

For a random synset syn;

For $syn \in SY$:

let PO be POS tags of words in WO,

Let us denote "Positive" to positive and "Negative" to negative score assigned by SentiWordNet. It is a dictionary available to users, which contain scores for words.

The value of Positive and Negative will always be between the range of 0.0 and 1.0. where 0.0 denotes negative score and 1.0 denotes positive score.

let $Xp(wo,po)$ be positive score of word.

let $Xn(wo,po)$ be negative score of word.

The scores are defined in terms of word wo and the POS tag of synset po.

The positive and negative scores can be defined as follows:

$$Xp(wo, po) = 1/n \sum_{syn \in SY} Positive(syn) \quad (1)$$

$$Xn(wo, po) = 1/n \sum_{syn \in SY} Negative(syn) \quad (2)$$

Where n defines the total synsets of word wd.

B. Formal Model for N-Gram Vectorization

A set of AR amazon reviews are given as input to the N-Gram Vectorization model.

For $ar \in AR$ and $wo \in WO$, let N_v be the N-Gram vector.

The N-Gram vector for the wo is denoted by N_v . It is a k dimension vector of TF-IDF values for 1 gram, 2 gram and 3 gram of word wo respectively. Let l_1, l_2 and l_3 be vectors containing TF-IDF values of word w for 1 gram, 2 gram and 3 gram respectively. The N Gram vector is given by,

$$\begin{aligned} N_v(w) &= \{(l_1+l_2+l_3) | l_1 \in T^1 \wedge \\ U(w) &= l_1, l_2 \in T^h \wedge \\ B(w) &= l_2, l_3 \in T^q \wedge \\ R(w) &= l_3 \} \end{aligned} \quad (3)$$

Where total of 1 to q gives the length k and U, B and R are mapping functions.

III. RESEARCH STRUCTURE

The basic steps involved in sentiment classification are shown in Fig.1. This section further provides details of the process with specific algorithm used in the process.

1. Data preprocessing involves the method to clean data to make it ready for algorithm.

2. N-Gram vectorization is used to make the data ready for machine learning algorithms. This layer converts words from the reviews to vectors.

3. A CNN model is used to finally classify the reviews into three divisions; positive, neutral and negative.

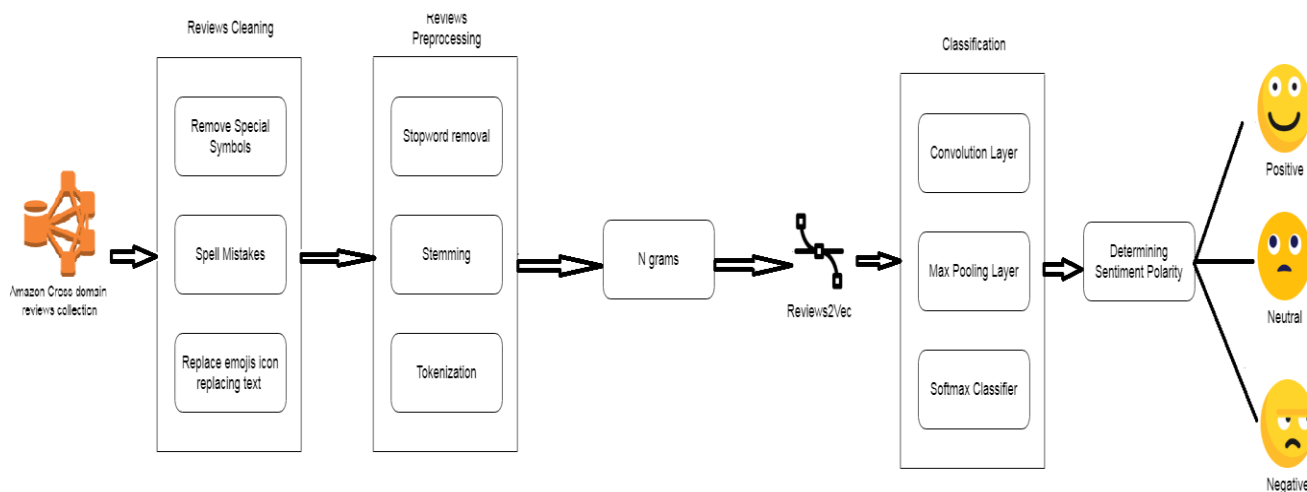


Figure 2: Architecture of the proposed system

A. N-Gram Vectorization

After the text is cleaned and ready, the next step is to convert the words into vectors. The study uses a process called N-Gram vectorization to convert the words into vectors. The method is very simple to use and provides good result and hence can be used for the study.

The process converts words into a matrix. For example consider a sentence such as “Vectorization is used for sentiment classification”.

Depending on the value of n, the sentence gets divided into matrix.

When n=1 “vectorization”, “is”, “used”, “for”

When n=2 “vectorization is”, “is used”, “used for”, “for sentiment”

When n=3 “Vectorization is used”, “is used for”, “used for sentiment”.

The value of n depends on the performance of the algorithm. The value cannot be low as the accuracy is decreased, but large value of n increases the matrix space. The number of features also depends on the value of n. Small n value mean less number of features which won’t determine the sentiments, but large number of features is not preferable as they increase the run time.

B Convolutional Neural Network (CNN):

The convolutional neural network is a type of neural network that finds it’s applications where the input is in the form of grid like structure for example image. The inputs are arranged in such a way that that each pixel is arranged in the form of grids. When we see an image, each neuron in our brain extracts a certain feature and then these features are connected to form a complete picture. The layers of CNN are as shown in Fig.3.

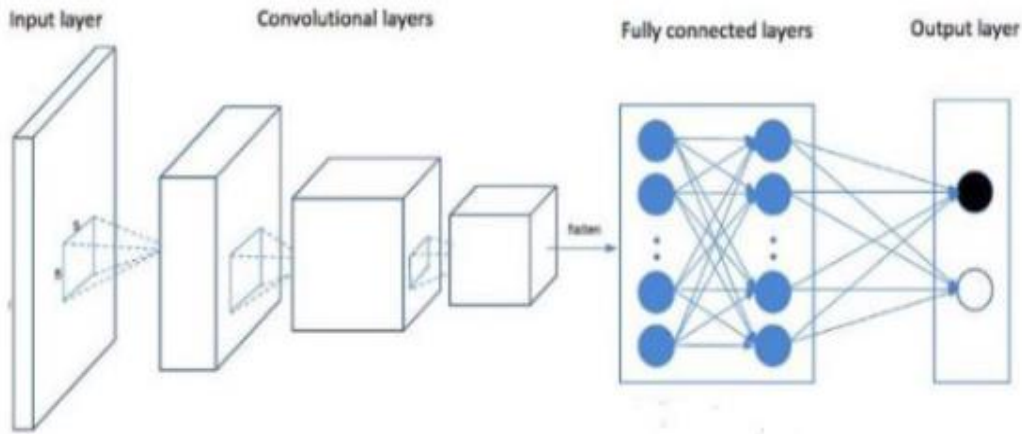


Figure 3: CNN Algorithm

Convolution Layer

Convolution layer is the most important and basic layer in the CNN algorithm. Convolution is performed in this layer to extract features from the image. The convolution is performed on the input with a smaller matrix called kernel. The kernel is small in size but not in dept. If an image has 3 channels i.e. RGB, the kernel will have the same number of channels. The kernel moves from the top left of the input and slides through the entire image. This produces a two dimension representation of the image called the activation map.

The final output of this stage will be the activation map which combines all the features of the words in the review. It can be shown as follows;

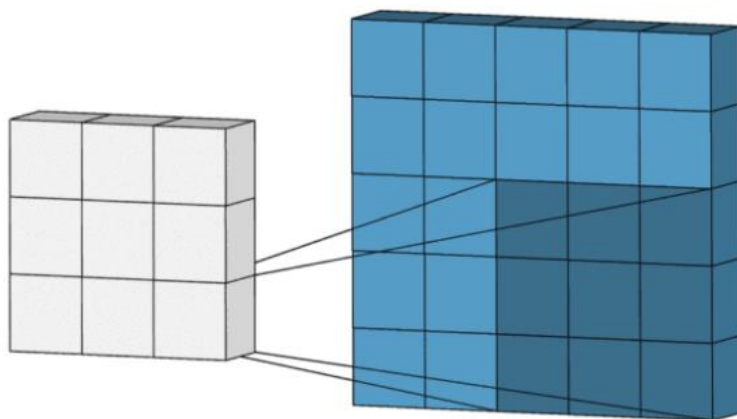


Figure 4: Convolution Process on Image(5x5) using kernel(3x3)

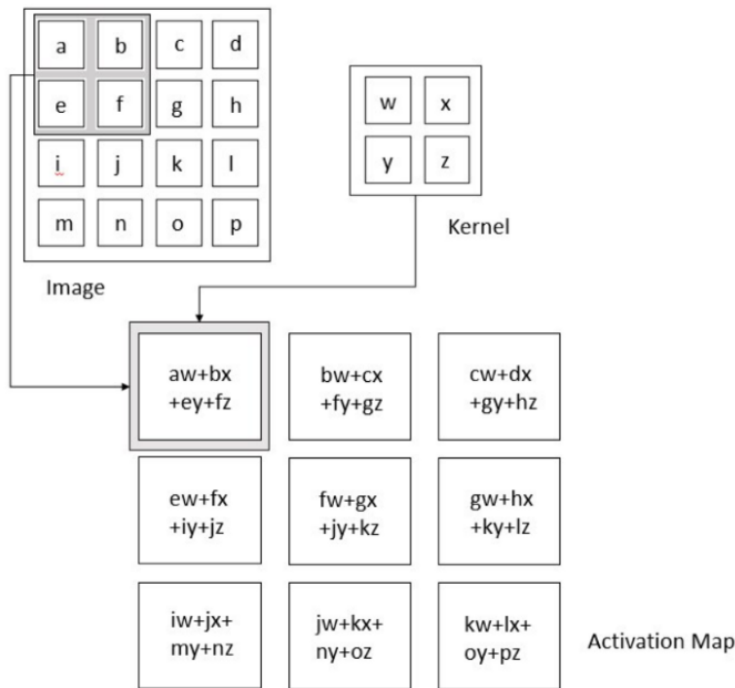


Figure 5: Activation Map Generated After Convolution Layer

Pooling Layer

The output of the convolution layer is taken as input. The pooling layer performs operation on the activation map. It replaces parts of the map with a summary of the nearby outputs thus reducing the size of the map when the dataset is large. Spatial pooling also known as subsampling or down examining, diminishes the dimensionality of each outline but holds the information of the data.

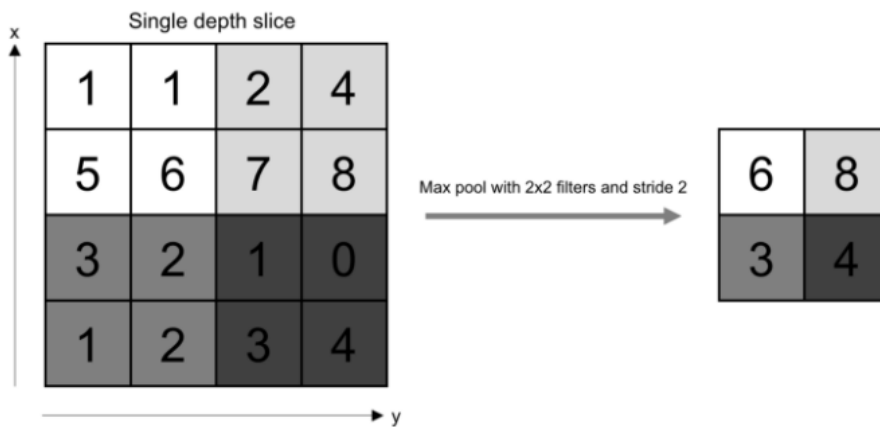


Figure 6: Example of Pooling Layer

Fully Connected Layer

Fully connected layer helps in connecting all the neurons in the input and output layers, thus providing the output to the softmax layer for classification..

Softmax Layer

A additional layer called softmax layer is added to the CNN algorithm. This layer classifies the text into three categories i.e positive, negative and neutral based on the features extracted form the previous layers

IV. EXPERIMENT AND RESULT

System used for the experimentation: Intel (R) Core (TM) i5-8250 CPU @ 1.60GHz, 1800MHz, 8GB memory, Windows 10, MySQL 5.1 backend database and Jdk 1.8. The architecture uses web application tool in Eclipse and executed on Tomcat server.

A total 2000 reviews are taken from amazon. Different domains are considered such as electronics, fashion, jewelry, etc.

The sentiment score is calculated and then the sentiment results are obtained. i.e 0.0 will give a negative sentiment while 1.0 will indicate a positive sentiment and 0.7 describes a neutral review. The results obtained are shown below in this session.

Review No.	Reviews	Sentiment Score	Sentiment Result
1	Bad product. I wanted it to return back but it is not returnable. Go for some other brands. I have been using spiriluna since long time but surely this is not spiriluna . Wastage of money. Some green leaf powder in the name of spirulina. Don't buy this product.	0.0	Negative
2	Good product. Price slightly on the higher side.	0.7	Neutral
3	Looks good for me. Have been using for the past 90 days. My immunity and sleeping time have improved 🍊	1.0	Positive
4	It is nice but box could be more strong.	0.7	Neutral
5	Good for health And value for money product	1.0	Positive

Table 1: Sentiment Classification of Amazon Reviews. [9]

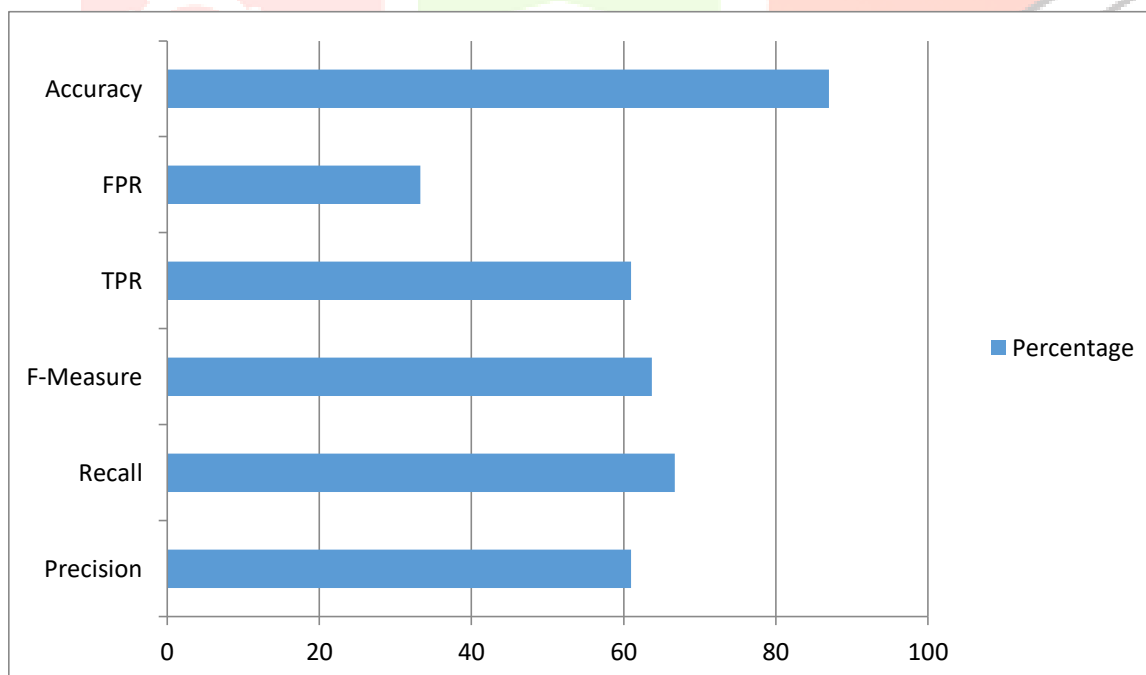


Chart 1: Performance Of The System

<i>Parameter</i>	<i>Value</i>
Precision	61
Recall	66.7
F-Measure	63.7
TPR	61
FPR	33.3
Accuracy	87

Table 2: Performance Parameter Of The Algorithm

V. CONCLUSION

The paper proposes an algorithm for cross domain sentiment analysis of amazon data. The algorithm is successfully implemented. The performance of the algorithm is checked. Different parameters are considered to test the algorithm. The accuracy of the system obtained is about 87%.

VI. FUTURE WORK

The application for the algorithm tested is limited to amazon reviews. Further increase in the performance is necessary. Further tests need to be conducted to analyze the effects of cross domain. Other methods based on lexicon and machine learning can be incorporated to get better results.

REFERENCES

- [1] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," in *IEEE Access*, vol. 8, pp. 17722-17733, 2020, doi: 10.1109/ACCESS.2019.2958702.
- [2] H. Yin, P. Liu, Z. Zhu, W. Li and Q. Wang, "Capsule Network With Identifying Transferable Knowledge for Cross-Domain Sentiment Classification," in *IEEE Access*, vol. 7, pp. 153171-153182, 2019, doi: 10.1109/ACCESS.2019.2948628.
- [3] B. Myagmar, J. Li and S. Kimura, "Cross-Domain Sentiment Classification With Bidirectional Contextualized Transformer Language Models," in *IEEE Access*, vol. 7, pp. 163219-163230, 2019, doi: 10.1109/ACCESS.2019.2952360.
- [4] H. T. Phan, V. C. Tran, N. T. Nguyen and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," in *IEEE Access*, vol. 8, pp. 14630-14641, 2020, doi: 10.1109/ACCESS.2019.2963702.
- [5] H. Wang, M. Hou, F. Li and Y. Zhang, "Chinese Implicit Sentiment Analysis Based on Hierarchical Knowledge Enhancement and Multi-Pooling," in *IEEE Access*, vol. 8, pp. 126051-126065, 2020, doi: 10.1109/ACCESS.2020.3008874.
- [6] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu, "Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary," in *IEEE Access*, vol. 7, pp. 43749-43762, 2019, doi: 10.1109/ACCESS.2019.2907772 .
- [7] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in *IEEE Access*, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
- [8] T. Al-Moslmi, N. Omar, S. Abdullah and M. Albared, "Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review," in *IEEE Access*, vol. 5, pp. 16173-16192, 2017, doi: 10.1109/ACCESS.2017.2690342.
- [9] <http://www.amazon.in/product-reviews/>