INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 17, Issue: 2, Month: April, Year: 2022
Article Number: 4351, https://doi.org/10.15837/ijccc.2022.2.4351

CCC Publications

AGORA UNIVERSITY PRESS

Sentiment Analysis using Improved Novel Convolutional Neural Network (SNCNN)

M. Kalaiarasu, C. Ranjeeth Kumar

M. Kalaiarasu
Associate Professor
Department of Information Technology
Sri Ramakrishna Engineering College, Coimbatore
Corresponding author:kalai.muthuswamy@srec.ac.in.

C. Ranjeeth Kumar
Assistant Professor (Sr. Gr)
Department of Information Technology
Sri Ramakrishna Engineering College, Coimbatore
ranjeeth.chandran@srec.ac.in

Abstract

Sentiment Analysis is an important method in which many researchers are working on the automated approach for extraction and analysis of huge volumes of user achieved data, which are accessible on social networking websites. This approach helps in analyzing the direct falls under the domain of SA. SA comprises the vast field of effective classification of user-initiated text under defined polarities. The proposed work includes four major steps for solving these issues: the first step is preprocessing which holds tokenization, stop word removal, stemming, cleaning up of unwanted text information like removing of Ads from Web pages, Text normalization for converting binary format. Secondly, the Feature extraction is based on the Bag words, Word2Vec and TF-ID which is a Term Frequency-Inverse Document Frequency. Thirdly, this feature selection includes the procedure for examining semantic gaps along with source features using teaching models and this involves target task characteristic application for Improved Novel Convolutional Neural Network (INCNN). The Feature Selection accompanies the procedure of Information Gain (IG) and PCC which is a Pearson Correlation Coefficient. Finally, the classification step INCNN gives out sentiment posts and responses for the user-based post aspects which helps in enhancing the system performance. The experimental outcome proposes the INCNN algorithm and provides higher performance rather than the existing approach. The proposed INCNN classifier results in highest accuracy.

Keywords: Sentiment Analysis (SA), Improved Novel Convolutional Neural Network (INCNN), TF-IDF is Term Frequency-Inverse Document Frequency, SVM is Support Vector Machine, Information Gain (IG) and Pearson's Correlation Coefficient (PCC).

1 Introduction

Community's view and various feedbacks are constantly enhanced with most essential and valuable resources for organizations or companies. Social media helps in raising new trends among each and every one; it paves way for unprecedented analysis and determination of numerous aspects in which organizations had to depend on unconventional, time consuming and error prone approaches earlier. The advancement of social media web is an important prospect to gain the valuable opinion of different people over numerous business works, political party areas, health-care causes and social media issues. These certain things have enhanced the growth of Sentiment Analysis in a dynamic principle of research fields [1]. Social media sites like FB, Twitter, Blogs include specific prime platforms where users can share their important opinion on certain topics with this kind of platform diverse opportunities and challenges arise to actively use mixed techniques to extract and understand the opinion of others [2]. Here, Sentiment Analysis holds a research area which helps in investigating the knowledge of people towards numerous matters such as various organizations, events and products. The part of sentiment analysis helps to improve the fast spread of micro blogging applications, social area networks and specific forums. Presently, every single web page has its own space for responding over user's feedback on products, services, and shares with friends on twitter, Facebook or Pinterest etc. It didn't happen a few years before but for each mining the amount of opinion gives out with large information and more understanding over human actions upon precious commercial interests [3].

The Sentiment Analysis (SA) considers Natural Language Processing (NLP) that consists of investigating and determining opinions with respect to products or movies. Further, this is mentioned as opinions of mining. These views and analysis are posted over an online network by numerous customers of the web resources for accessing blogs and other social media related platforms. This review helps the person to obtain a sufficient study on overall people's opinions regarding the specific tool on positive, negative and neutral [4-5]. At the same time, it is difficult to handle tasks based on time consumption. Here, a single person has to undergo a task for consumption of time which basically focuses on a certain restaurant or a movie in the earlier stage of settlement for selecting and building a finite number of decisions. The important classification required for sentiment analysis is based on the three execution levels: (i) document level (ii) sentence level (iii) sub sentence level. This level helps in determining the exact sentiment analysis for documenting, processing of reviews in individual sentences and respectively obtaining a sub-expression of opinions in the required sentences [6-7]. CNN with air pollution index prediction model is studied in [24] and document classification model-based Pearson correlation-based feature analysis is reviewed in [25]. Single dimension CNN with feature extraction process is modeled for signal status recognition model [26] and the intrusion detection and normalization process is done with CNN [27]. The CNN - LSTM layer on forecasting particulate matter is reviewed in [28].

The sentiment analysis describes the perspective based on several events and situations. This repeatedly consists of infinite self-considerate feelings like happiness, sadness, kindness and regretless. This sentiment analysis label is splitted into polarity and valence of positive, negative and neutral which shows numerous feelings such as angry, sad, happy and proud. This sentiment label clarifies and influences the outcome of the analysis therefore, we must express the sentiment analysis carefully. This study defines more than three sentiment labels such as emotional feelings, score rating and acquires labels of two dimensions such as positive or negative [8-10]. Together sentiment analysis performs better work for field improvement, and classifies all kinds of binary classification challenges such as performance of accuracy varied from 70 to 90 percent in characteristics with data terms. This binary classification has both positive and negative terms while the ternary classification includes all three types of terms so called positive, negative and neutral. In General, the methodology of sentiment analysis is shown in Figure 1.

In numerous studies, the classification of sentiments is done based on machine learning models like Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and ensemble. The important features used for generating ML i.e., machine learning trained models are n-grams. The feature of unigram is based on Sentiment Analysis of binary classification with an accuracy of 88.94% by assisting the SVM. Further, it adopts the features of Unigram and Bigram with analysis of binary sentiment classification to obtain an accuracy of 84.4% using SVM

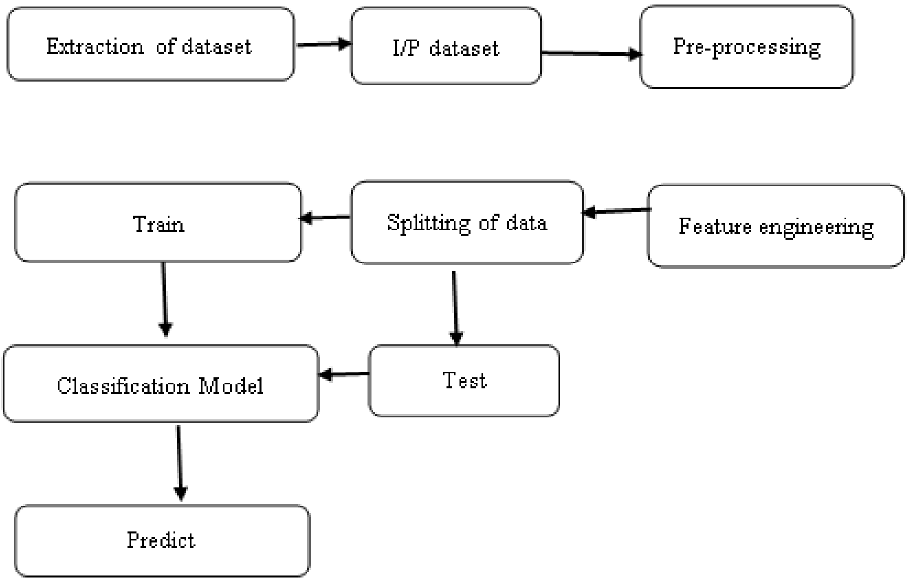


Figure 1: GENERAL BLOCK OF SENTIMENT ANALYSIS WORK PROCESS

movie review data. This tackles all classification of binary data to process the features of Unigram and Bigram in order to attain an accuracy of 86.1% from the translated product review information of Amazon.

The proposed sentiment analysis of INCNN framework holds four important key steps: (i) Preprocessing (ii) Feature Extraction (iii) Feature Selection (iv) Classification.

1. In the first step, Preprocessing holds the use of tokenization, stop word removal, stemming and text cleaning features approaches. The Preprocessing steps include message cleaning to get rid of hyperlink frequent posts. In the initial stage, the posts are clustered based on numerous features, Tokenization is a function which segments text into words, clauses or sentences. Stemming is an action of progressing the conflict word to a common system of representation. Text cleanup focuses on removal of unwanted and unnecessary information like removing ads from web pages, normalizing the text converted from binary formats.

2. In step two, Feature Extraction is done based on the Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec. Text feature extraction belongs to a task function which lists out text data words and transforms it to feature set classifiers. In general, it recognizes correctly the importance of features related text. Word2vec for generating words as vectors. In a textual corpus, this extracts the relationship between the words and it is termed as word combinations.

3. In the third step, feature selection procedure is evaluated for semantic gaps along with source features produced using teaching model and target task characteristics before the application of Improved Novel Convolutional Neural Network (INCNN). This Feature Selection accompanies the procedure of IG which is an Information Gain and PCC which is a Pearson's Correlation Coefficient. The term IG is used for measuring the obtained number of information bits for predicting the category of presence or absence of a document.

4. Finally, Classification of IN CNN provides sentiments towards the user post and responses to build up with a better perform ance system.

The whole paper is organized as following things,

1. Literature review on sentiment analysis on Facebook existing schemes are briefly discussed.
2. The proposed methodology uses sentiment analysis over Deep learning algorithms on the Facebook system.
3. Discussion and outcome of the study is presented.
4. Conclusion and future enhancement of the presented work is expressed.

2 Literature survey

Alnawas and Arici [11] sketched a work on Logistic Regression (LR), Naive Bayes (NB), Decision Trees (DT) and Support Vector Machine (SVM)) in Sentiment Analysis word embedding from Iraqi Arabic. Firstly, word representation helps to learn a huge corpus of work and secondly, it assists the embedded word to generate the model for corpus training in Doc2Vec paragraphs and constructs a architecture on DM-PV which focuses on Distributed Memory Model of Paragraph Vector. At the end, four binary classifiers are used for representation of feature training such as LR, SVM, DT and NB which detects sentiment analysis. Here. different values of parameters are analyzed like window size, dimension and negative samples. The brighter side of the experiment concludes with a proposed approach to achieve a good performance of work based on LR and SVM classifiers.

Salloum et al [12] presented a paper on Natural Language Processing (NLP) which determines the classification of various comments and posts in Facebook pages of Arabic newspapers. Here, entirely 24 numbers of Arab gulf newspapers are explored and studied based on the Facebook page with 6237 posts and 9372 comments. The data extraction includes various kinds of text mining methods to analyze the process of operations. In concern with Arab Gulf region, UAE is the only country which indicates more shared posts over Facebook. Further, KSA and Oman follow the same position. Particularly, finding helped in enhancing the most attractive video post over the Facebook page of Arabic newspapers.

Nahar et al [13] designed a work on lexicon-based method, Machine Learning (ML) algorithms with Naive Bayes (NB), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) for identifying polarity of the provided Facebook comments in classification and sentiment analysis. The data samples are from local Jordanian people who leave their comments on a public issue related to the services provided by the main telecommunication companies in Jordan (Orange, Zain and Umniah). The produced results regarding the evaluated Arabic sentiment lexicon were notably promising. By trying the user-defined lexicon-based Facebook posts on the common and comments used by Jordanians, it scored (40%) negative and (60%) positive. The lexicon was used to label a set of unlabeled Facebook comments to formulate a big dataset using supervised Machine Learning (ML) algorithms that are usually used in polarity classification.

Tran and Shcherbakov [14] outlined an unsupervised clustering pattern on sentiment analysis in text identification and predication with positive or negative connection of Facebook comments.
(*i*) Discovery of patterns with real-time sentiment text analysis.
(*ii*) Creation of forecasting algorithm choices based on data processing batch system.
Proposed methods are of followings:
(1) Batch analysis is performed based on Facebook data collection for forecasting the model pattern.
(2) Detection of real-time patterns using text processing.
(3) Pattern prediction is built based on the grasping situations.
(4) Change in action with pattern analysis.
This includes achievement of the proposed forecast with two step algorithms as follows:
(*i*) Unsupervised technique is used for pattern clustering.
(*ii*) The nearest pattern is identified based on cluster prediction.
Also, this described and revealed about the three major kinds of user patterns.

Soliman et al [15] introduced SSWIL which is a Slang Sentimental Words and Idioms Lexicon and enhanced SVM which is a Support Vector Machine. This is used in sentiment analysis over Facebook comments in Arabic slang. Further, this proposes a Support Vector Machine on the Gaussian kernel to classify all the comments in Arabic newspapers. The growth of the outcome has 1355 random comments and major three types of classifications. Here, web users' comments are written in the form of new brand syntax where the outcome of classification may be affected. Moreover, in the initial stage of classification the extraction method fails to take out the words opinion but it gives a better performance by adding SSWIL. The proposed mining application applies all the comments and classifies the comments based on the first classification type known as classic lexicon which generates a lower accuracy level rate. The second classification type is expressed as SSWIL with classic lexicon which generates a higher accuracy level rate. Finally, the testing is performed on proposed classifiers with the help of various Facebook comments that are obtained on newspapers where the rate of accuracy is high based on Precision(P) and Recall(R).

Meire et al [16] sketched an outline on SVM - Support Vector Machine and RF - Random Forest over Facebook posts using sentiment analysis. The main motive of the review is (i) to evaluate the value attached information based on leading the existing time and to know the lagging of proposed time in creating focal posts for Facebook by sentiment analysis (ii) to determine the prediction analysis (iii) to investigate the relation between sentiment and predictors. Construct a sentiment prediction model that consists of information based on prime, lagging and traditional posts for the process of cross validation in RF and SVM. The outcome mentions both are lagging and leading information can increase the model's predictive performance. The key feature of a predictor consists of the integer in uppercase letters with a number of negative comments. A higher number likes increases the uppercase of likelihood positive posts. While, a higher number of comments reaches the likelihood of a negative post. The important contribution of this work phase is to context the sentiment analysis and access the added value of lagging information's and leading information's.

Ortigosa et al [17] developed a Sent Buk following hybrid approach that combines lexical-based SVM-Support Vector Machine and enhanced NB-Naive Bayes for the process of sentiment analysis over comments of the Facebook. Users support certain messages: (i) to clear user sentiment polarity information with positive, negative and neutral message transmission. (ii) to determine the sentiment analysis of the user model with certain emotional noted changes. The messages written by Facebook users are easily retrieved from Sent Buk. Accordingly, this classifies the polarity of the messages and results the user through GUI which is a Graphical User Interface. Here, detection of emotional changes is recorded based on the statistics and findings of messages by our friends. This is a hybrid approach where the method of classification is implemented over Sent Buck with lexical machine learning based techniques. This approach enhances the outcome of feasible study and helps to perform sentiment analysis with best accuracy over Facebook. The users of sentiment analysis hold all the context of e-learning and also supports all the customized learning operations with various emotions. Here, all the user's activities are tackled by assisting the recommendation of suitable operators. On other side, Course of the students are sentimental with feedback towards the teaching staff. In particular, during online class there will be less contact over various face connections.

3 Proposed methodology

Sentiment analysis using proposed INCNN framework contains four major steps called preprocessing, feature extraction, feature selection and classification. First things first, preprocessing step, stop word removal, tokenization, stemming. The preprocessing step includes messing up the contents to erase repeated posts and also hyperlinks. Tokenization will process the segmenting of text into words, clauses or sentences. Next part is feature extraction based on Word2Vec, Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words. In a textual corpus, it extracts the relationship between words and it is termed as word embeddings. Feature selection is the third step; its procedure for evaluating semantic gaps among source features produced using a teaching model and target task characteristics before INCNN applications. Feature selection follows the structure of Pearson's Correlation Coefficient (PCC) and Information Gain (IG). The term measured in IG is the number of bytes memory storage information acquired from the category of prediction by the absence or presence in a document. The figure-2 shows the overall idea of the proposed system.

3.1 Preprocessing

The preprocessing method includes erasing of messages to remove the replicate and hyperlinks the posts and also it is initially trying to cluster the posts formulated on different aspects. Those messages are hashtags and trimmed off are gathered with them. Here, crawlers collect the amount of data from Facebook that has several kinds of meanings about punctuations. Hence, the data is altered and punctuated to return back to text. This preprocessed text is crossed to the module for detection posture where the text features are cleaned up for removing stop words, tokenization and stemming.

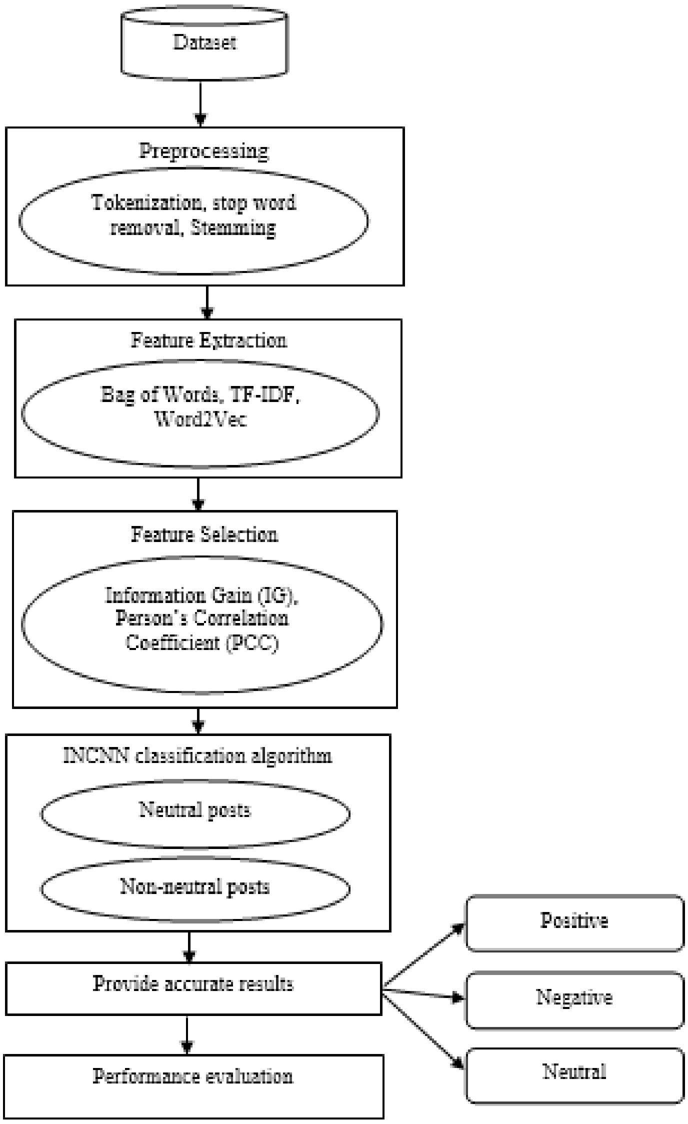

Dataset
Preprocessing
Tokenization, stop word removal, Stemming
Feature Extraction
Bag of Words, TF-IDF, Word2Vec
Feature Selection
Information Gain (IG), Person's Correlation Coefficient (PCC)
INCNN classification algorithm
Neutral posts
Non-neutral posts
Provide accurate results
Performance evaluation
Positive
Negative
Neutral

Figure 2: OVERALL DIAGRAM OF THE PROPOSED SYSTEM

A.Tokenization

The tokenization is the process of cutting down a stream of text into phrases, words, symbols or other kinds of meaningful elements. The main purpose of the tokenization is to explore the words into a particular sentence. The overall list of tokens would become an input for further processing areas such as text mining/parsing. The tokenization is applicable for both in computer science and in linguistics, where it is the form part of the lexical analysis. The textual data is the only block of characters in the initial. Thus, the overall requirement of documents for that parser is tokenized. The tokenization is the operation of segmenting text into words, sentences or clauses and it is the process by the big quantity of text to be divided into smaller parts. It can be very useful in seeking such patterns to be the same as a base level step for lemmatization. Usage of the function word tokenize () is to split a sentence into words and the result of word tokenize () can be transform into pandas Data frame for better text performance and understanding in machine learning use cases [18]. description = nlp. word_tokenize (description) is used for tokenization.

B.Stop Word Removal

Stop word removal is a very important procedure of preprocessing. The procedure ensures that superfluous words also with more or less not a big data content for the task under the considerations are discarded. It makes the purpose of a linked resource to perform to remove the stop word. This is part of Stanford resources in NLP. According to their high-level frequency of occurrence, their presence in text mining presents a barrier in the understanding of the content in the documents. The important method of text data is to decrease the process and also to improve performance of the system where all the text document manages to operate the words which are not needed for text mining analysis. Those words are already captured in this corpus named corpus. A list of stop words is made and the source document is compared with this stop list. If any word matches with this list, that word is removed from the source document. Stop words don't carry any value in matching queries to the document by removing the stop words did not affect the document semantics [18].

C.Stemming

The process of stemming helps in consolidating the various word forms into simpler text descriptions which is often used for text processing IR - Information Retrieval assumptions. This poses all the queries for representing the document terms and executes all identified tokens for Facebook with user generated strings. The following two features of token are: (i) Main token and (ii) stemmed token where the process accepts the porter stemmer for stemming operations [19].

3.2 Feature Extractions

Feature Extraction is an important operation for extracting data in feature subsets in order to improve jobs of data classifications. The process of feature extraction text helps in identifying the text data of words list and transforms all the data into a set of feature classifiers which exactly finds out the key features of the text. The following are the feature extraction techniques used.

Bag of Words
The most common and simplest method of feature extraction is bag of words. This forms an instant feature set in the presence of words and that is described as a "bag" of words. These features help in bringing out the text from angles. Further, this identifies the frequency of document words and it consists of (i) Known words have frequency existence (ii) Words of lexicon which is generally a word. The worst model case determines bags of words which processes the operation for scoring the famous words presence and to come out with familiar vocabulary words design. Word2Vec An open-source tool called Word2Vec developed by Google for representing words as vectors. In a textual corpus, it extracts, relationship between words. It is termed as word embeddings [20]. Unsupervised learning is involved in word2vec for producing word embedding representation in a meaningful manner. Further, the creation of vector space consists of more dimensions such that each of the corpus words is distinctive with allocation and correspondence of the vector space. . The common word context is marked near the proximity of vector space. Word2vec uses more than one architecture for processing CBOW which is a Continuous Bag of Words that holds frequent skip grams. In these current words are used for predicting context words near windows. In the architecture of CBOW, prediction is not at all impacted by context word sequence and the basic model is built on bags of words. Term Frequency - Inverse Document Frequency (TF-IDF) The particular way of scoring is known as Term Frequency - Inverse Document Frequency.

- TF is the term frequency of new documents in word frequency
- IDF is the inverse document frequency of all documents to score the words

The score highlights the important unique word in which a particular word represents necessary document information. Thus; IDF obtains high infrequent terms with low frequent terms. In order to create the glossary, the weight of all words is listed in connection with TF - IDF for recognizing the text mining algorithms. The frequency of word denotes repeated number of terms in the text and expands the IDF as Inverse Document Frequency. In specific, these algorithms are used for determining the word text in inverse probability. Following equation (1) helps in classifying TF - IDF equation and this equation is used for the calculation of weight as follow:

$$W_{ij} = tf_{ij} * loglogN/(df_i)(1) \tag{1}$$

This equation specifies,w_ij is word i weight in document j. Here, N is the document number in the total set of documents. Next, tf_ij is word i frequency in document j and df_i is word i holding documents number.

3.3 Feature Selections

Feature selection technique is used for evaluating semantic gaps with all sources of features produced by teaching models and target task characteristics before the application of INCNN. This assists the procedure of Information Gain - IG and Pearson Correlation Coefficient - PCC. Feature selection methods can enhance the best efficiency.

A.Information Gain (IG) The term Information Gain (IG) holds bit information number to obtain a categorization of predictive value which gives the presence or absence term of document forms [21]. Further, m is denoted as class number and the gained information t term is expressed as

In class C, feature F's is analyzed using an information gain in sentiment analysis area. In a mutual information the value between feature F's sentiment classes is high, then there will be an immense value of relevance between feature F's sentiment.

$$I(C,F) = H(C) - H(C|F) \tag{2}$$

$$H(C) = -\sum cECp(C)logp(C) \tag{3}$$

Here, the class entropy is denoted as $H(C)$, and the class features of conditional entropy is presented as $H(C|F)$,

$$H(F) = -\sum cECp(C|F)logp(C|F) \tag{4}$$

In this dataset the classes are balanced. Moreover, for negative and positive where class C's probability is 0.5 and it generates 1 as a class entropy (C). Information gain is expressed as,

$$I(C,F) = 1 - H(C|F) \tag{5}$$

If $H(C|F) = 1, I(C,F)$'s minimum value is generated and it presents there is no relation between classes C and feature F. Thus, the feature F is selected in such a way that it emerges either in negative class or positive class. If $P(F|C_1)$ is same as $P(F)$, maximum value ofI $(C|F)$ can be achieved and it results in H(C_1 | F) and P(C_1 |F) value of 0.5. The $P(F|C_2)$ value results in $P(C_2|F) = 0$, if $P(F) = P(F|C_1)$ and makes $(C_1|F)$=0. The $I(C,F)$ value is changed between 0 to 0.5.

B.Pearson Correlation Coefficient (PCC) The Pearson Correlation Coefficient (PCC) is the linear computation between two variables associated. The regression line slope is equal to the ratio of standard PCC deviations. The Pearson Correlation tackles all the variables and eliminates the single outlier correlation line to decrease the error in mainstream to follow up with normal distribution variable homoscedasticity where the same regressed variance lines are used.

$$p_pearson(X,Y) = \frac{\sum_n^(i=1)(x_i-(x))(y_i-y))}{\sqrt{(\sum_n^i = 1(x_i(x)^2)\sum_n^i = 1(y_i(y)^2)}} \tag{6}$$

This case focuses on Joint Normal Distribution of Pearson Correlation Coefficient (PCC) that accompanies with t-distribution n^2 which is a non-dependent X and Y degree of freedom. Here, X and Y are linearly dependent with a set pearson (X,Y)=±1 . This case includes a perfect positive that denotes the linear relationship increment where Pearson Correlation Coefficient is an important set of +1 and the case of perfect negative denotes linear relationship decrement with a set of -1.

If X and Y are nonlinear association. However, Pearson(X,Y)=0 and partial linear dependency case shows -1 <P_Pearson (X,Y)<1 . Even Though, no data association scores zero correlations and it does not imply any association. Pearson Correlation is to analyze forms of different words to calculate all correlated event days. These words enhance the classifier feature for training and testing. An accurate prediction of a class in classification by a feature is examined using PCC while riding other features. According to the realized score of correlation p, every feature is ranked sequentially.

Here, the feature of X_i's variance is defined as var(X_i) with a covariance between target class Yand the feature X_i is defined as $cov(X_(i,),Y)$ where the Variable covariance is denoted as cov. Simultaneously, every variable's variance is expressed as var.

3.4 Sentiment Classification using INCNN

A specific classification of Improved Novel Convolutional Neural Network (INCNN) provides a process description for text classifications. Here, CNN has multistage trainable Neural Network architecture is developed for task classifications [22].

1.The Convolutional Layers are important parts of Convolutional Neural Network model. This layer holds the amount of kernel matrix. It produces an output matrix of features and convolution on their inputs where a bias value is joined. To learn the method it aims to train the biases as shared and kernel weights in neuron connection weights.

2.The Pooling Layers are the specific and essential component of CNN. The purpose of the pooling layer is to behave as dimensionality depletion of the input feature. The pooling layers form subsampling output matrices of convolutional layers to connect up with nearest elements. One of the most familiar pooling functions is max-pooling. That takes in the peak value of local neighborhoods.

3.Hidden Layers are the exclusive part of text classification in CNN. To convert the text inputs into an applicable form of CNN is the major usage of the embedded layer. Thus, convert the whole text document into a dense vector of stable size. Softmax Layers are the finest hidden layers which have an exceptional case of convolutional layer. It interpreted with a kernel size of 1×1. That type of layer resides with these are used in the concluding stages of CNN and a class of trainable layer weights.

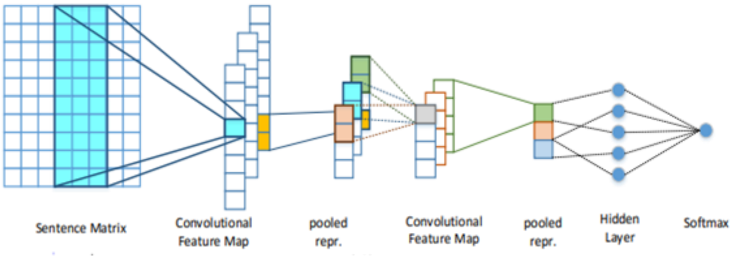


Figure 3: ARCHITECTURE OF INCNN

CNN holds the advantage of Convolutional filters. Those filters automatically learn the features which are apt for the specified task. Following the example, CNN is used for the sentiment classification of Convolutional filters that captures out semantic features of sentimental expressions and inherent syntactic where the threshold varies based on the dataset.

$$p = \frac{cov(X_{iY})}{\sqrt{var(X_i).var(Y)}} \tag{7}$$

n_w k_p denotes the expected frequency of co-occurrence, (freq(w,p)-n_w k_p indicates the difference between observed and expected frequencies.

Algorithm 1: INCNN Input: Analyze a labelled dataset

Output: Based on polarity of extracted reviews whether it is positive, negative and neutral. Step 1: Extract the reviews.

Step 2: Obtain the feature vectors.

Step 3: Joining up training dataset and feature vector list by computing the probability features using

(9)
Step 4: Performance of Training and Testing using INCNN algorithm.
Step 5: Identify similar feature vectors using convolution layers.
Step 6: Give out positive, negative and neutral post

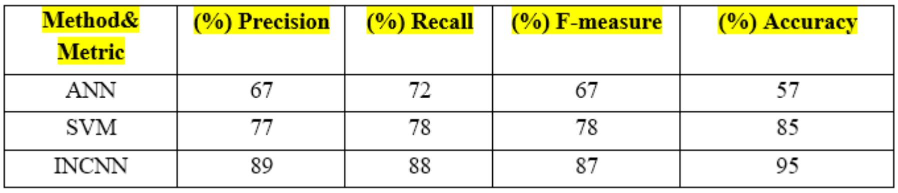
Convolutional layer is used for extracting patterns, i.e., specific discriminative sequence words are obtained inside of the input words that would commonly follow during the instances of training. The two-word types used for encoding the elements are token spanning phrases which are used to encode all others with 0s and with 1s. Here, every word of type is combined with its self-embedding. Thus, when we tackle the phrase level classification of sentiment it forms a matrix in sentence S as following: every token of a comment has a corresponding alternate word embedded in the matrix of word W. It is embedded for each one of the words in two types. Eventually, the whole input of the matrix in sentence is augmented with an additional set of the rows from embedded words. Overall, the architecture of the proposed network remains constant.

4 Results and discussion

Originally, the dataset was downloaded from Kaggle [23]. This experimental setup indicates the online social workers in real-time environments. These datasets are gathered from a group named Cheltenham Facebook and major three groups of open source are selected for enhancing the experimental features. Here, the dataset gives an exact set-up and senses the information of the gained matrix for sparse ratings. Following are two set-ups of conducted experiments are (i) To prove effectiveness of proposed techniques in executing customized post solutions and to give out a free clutter environment group. (ii) To prove capability of the proposed work to identify the community of members in a population. These classification methods are analyzed in the terms of metric like Precision (P), Recall(R), F-measure(F1) and Accuracy(A). In this part, SVM and ANN algorithms are compared for determining the performance metrics against proposed work of INCNN algorithms which is further tested by a dataset of Cheltenham Facebook Groups.

Dataset Description.
In general, the dataset is downloaded from Kaggle [23]. These attributes include gid, pid, Cid, timestamp, id, name, rid, msg. In the first case, for more the required models are trained on a training set (80%) and tuned using validation set with final test set (20%).

Evaluation Metrics
In order to classify each and every model various estimation of metrics are used they are listed as Precision, Recall, F-measure and Accuracy. This needs multiple usages of metrics because they don't specify the same values for all accounts.



Method& Metric	(%) Precision	(%) Recall	(%) F-measure	(%) Accuracy
ANN	67	72	67	57
SVM	77	78	78	85
INCNN	89	88	87	95

Table 1: PERFORMANCE COMPARISON METRICS VS. SENTIMENT ANALYSIS METHODS

1.Precision(P) By considering Posts as positive samples in the binary classification, precision is expressed as,

In the above Figure 4, this recognizes about the analysis of comparison metric in order to determine the usage of existing and approached methodology in the terms of Precision. Further, the methods of x-axis are grasped and the approach of y-axis helps to plot the values in Precision. Here, the existing method holds the SVM and ANN algorithms that provide a lower Precision whereas the proposed INCNN approach gives out the uppermost Precision towards the labeled datasets. Finally, the result concludes with a proposed INCNN that helps improve the sentiment analysis classification method by

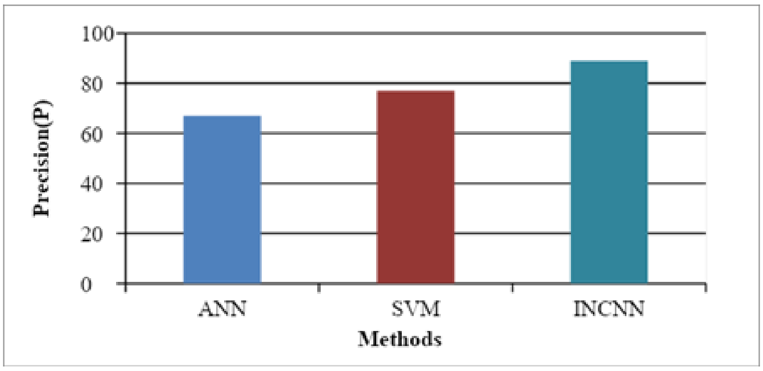


Figure 4: PRECISION OF SENTIMENT ANALYSIS METHODS

classifying the neutral and non-neutral posts effectively for the defined dataset.

2.Recall(R) The calculation of the recall(R) value is defined as:

$$Recall(R) = \frac{tp}{(tp + fn)} \tag{8}$$

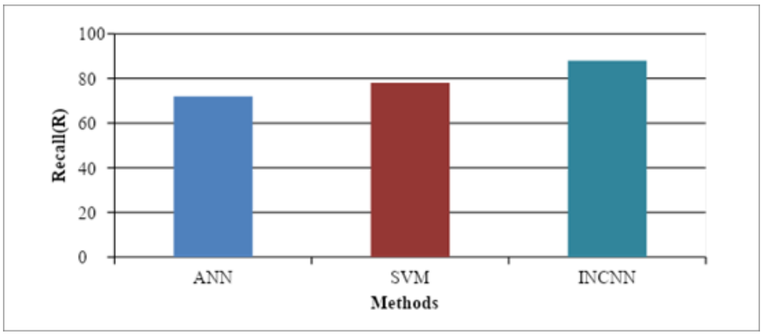


Figure 5: RECALL(R) COMPARISON OF SENTIMENT ANALYSIS METHODS

In the shown figure 5, we observe the compatibility of the required metrics to examine the usage of previous and presented methodology in the term of Recall. This approach grasps a specified process on the x-axis and recall for accurate marked y-axis values. SVM and ANN algorithms group the existing system method that provides a lower Recall and the INCNN proposed technique gives out the higher Recall for a specified dataset. In general, the result concludes that the proposed INCNN enhances the sentiment analysis classification process by analyzing the neutral and non-neutral posts are more efficient for the specified dataset.

3.F-measure(F1) F-measure(F1) is denoted as

$$F1 - score = \frac{(2XprecisionXrecall)}{(precision + recall)} \tag{9}$$

The above Figure 6, discovers about the comparison metric in order to estimate it using existing and proposed methodology in the terms of F-measure. Further, analysis of x-axis methods are noted and simultaneously y-axis F-measure (F1) values are also plotted out. In the previous methods, SVM and ANN algorithms delivered bottommost F-measure in which the presented INCNN technique gives out uppermost F-measure for a stated dataset. Thus, the outcome concludes that the proposed INCNN builds up a sentiment analysis for classification process by associating the neutral and non-neutral posts more efficiently for the given dataset.

4.Accuracy(A) Accuracy(A) is determined as the total accurateness of detection outcome and it is examined as the additional classification of parameters (tp+tn) divided by overall number of

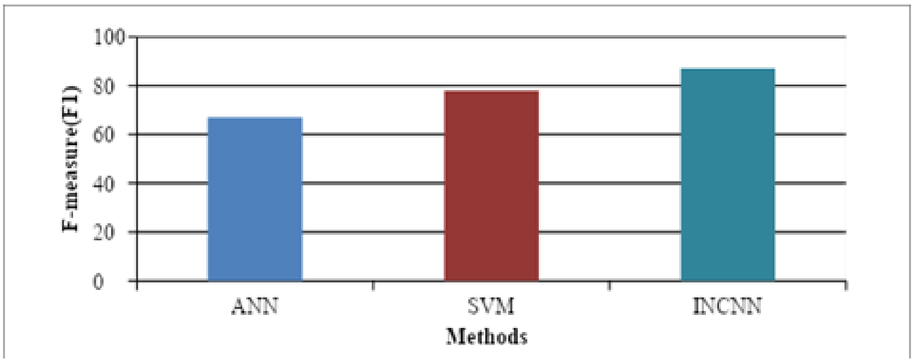


Figure 6: F-MEASURE COMPARISON OF SENTIMENT ANALYSIS METHODS

detections with classification of specific parameters (tp+tn+fp+fn)

$$Accuracy = \frac{(tp + tn)}{(tp + tn + fp + fn)} \tag{10}$$

In respect, the parameters tp,tn,fp and fnconsists of numbers such as true positives, true negatives, false positives and false negatives.

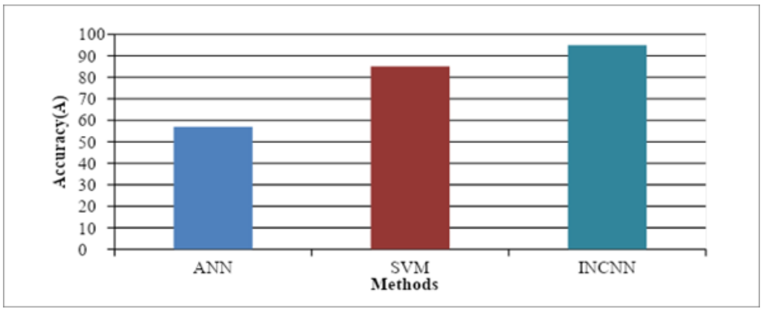


Figure 7: ACCURACY COMPARISON OF SENTIMENT ANALYSIS METHODS

In the above Figure 7, the approach of the existing and proposed methodology is to assess the comparison metric in the terms of Accuracy. Further, in x-axis the methods are grasped and in y-axis the accuracy value is marked. Here, the technique of the existing method includes SVM and ANN algorithms which provide a lower Accuracy whereas the initiated INCNN approach contains huge Accuracy for a specified dataset. Finally, the outcome of the proposed INCNN helps to improve the sentiment analysis classification task by recognizing the neutral and non-neutral posts effectively for the given dataset.

5 Conclusion and future works

The four major steps included in INCNN proposed work are (i) Tokenization of preprocessing, terminates word removal, stemming, cleanup of text this means discarding unwanted. (ii) Feature extraction based upon Bags of Words, Word2Vec, TF-ID and in this an important subset feature is extracted from classification tasks for improving data. (iii) Selection methods for evaluating semantic feature gaps which are generated using teaching models and characteristics of target tasks in previous INCNN applications and the feature selection follows up IG/PCC procedures. (iv) INCNN provides sentiment posts, user replies are based upon post aspect and the system performance is enhanced. Hence, the dataset selects up a prominent keyword which is used for classifying posts more effectively. Here, the INCNN algorithm proposes a sentiment classification algorithm that classifies the posts as neutral or non-neutral. Moreover, the proposed INCNN algorithm finds the greater performance in terms of higher Accuracy, Precision, Recall and F-measure where, the values are compared to other state-of-the art methods known as SVM and ANN classifiers. The Future work holds (i) To improve accuracy classification to tune up parameter weights and biases of INCNN classifier, certain optimization algorithms are introduced for Animal Migration Optimization (AMO), Harmony Search Optimization (HSO), etc., (ii) Hybrid INCNN combined with Deep Neural Network (DNN).

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflict of interest

The authors declare no conflict of interest.

References

[1] R. Feldman, 2013. “Techniques and applications for sentiment analysis”, Communications of the ACM, Vol.56, No.4, Pp.82-89,

[2] M. Ghiassi, J. Skinner and D. Zimbra, 2013. “Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network”, Expert Systems with applications, Vol.40, No.16, Pp.6266-6282,

[3] N. Zainuddin, A. Selamat and R. Ibrahim, 2016. “Improving twitter aspect-based sentiment analysis using hybrid approach”, Asian conference on intelligent information and database systems, Pp.151-160,

[4] D. Antenucci, M. Cafarella, M. Levenstein, C. Ré and M.D. Shapiro, 2014. “Using social media to measure labor market flows”, National Bureau of Economic Research, Pp. 1-50,

[5] N. Zainuddin, A. Selamat and R. Ibrahim, 2018. “Hybrid sentiment classification on twitter aspect-based sentiment analysis”, Applied Intelligence, Vol.48, No.5, Pp.1218-1232,

[6] G. Wang, J. Sun, J. Ma, K. Xu and J. Gu, 2014. “Sentiment classification: The contribution of ensemble learning”, Decision support systems, Vol.57, Pp.77-93, 2014.

[7] O. Kolchyna, T.T. Souza, P. Treleaven and T. Aste, 2015. “Twitter sentiment analysis: Lexicon method, machine learning method and their combination”, Computation and Language (cs.CL), Pp.1-32,

[8] N. Öztürk and S. Ayvaz, 2018. “Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis”, Telematics and Informatics, Vol.35, No.1, Pp.136-147,

[9] K. Philander and Y. Zhong, 2016. “Twitter sentiment analysis: Capturing sentiment from integrated resort tweets”, International Journal of Hospitality Management, Vol.55, Pp.16-24.

[10] A. Hasan, S. Moin, A. Karim and S. Shamshir band, 2018. “Machine learning-based sentiment analysis for twitter accounts”, Mathematical and Computational Applications, Vol.23, No.1, Pp.1-15,

[11] A. Alnawas and N. Arici, 2019. “Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents”, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Vol.18, No.3, Pp.1-17,

[12] S.A. Salloum, C. Mhamdi, M. Al-Emran and K. Shaalan, 2017. “Analysis and classification of Arabic newspapers’ Facebook pages using text mining techniques”, International Journal of Information Technology and Language Studies, Vol.1, No.2, Pp.8-17,

[13] K.M. Nahar, A. Jaradat, M.S. Atoum and F. Ibrahim, 2020. "Sentiment analysis and classification of arabjordanian facebook comments for jordanian telecom companies using lexicon-based approach and machine learning", Jordanian Journal of Computers and Information Technology (JJCIT), Vol.6, No.03, Pp.247-263,

[14] H. Tran and M. Shcherbakov, 2016. "Detection and prediction of users attitude based on realtime and batch sentiment analysis of facebook comments", International conference on computational social networks, Pp.273-284,

[15] T.H. Soliman, M.A. Elmasry, A. Hedar and M.M. Doss, 2014. "Sentiment analysis of Arabic slang comments on facebook", International Journal of Computers & Technology, Vol.12, No.5, Pp.3470-3478,

[16] M. Meire, M. Ballings and D. Van den Poel, 2016. "The added value of auxiliary data in sentiment analysis of Facebook posts", Decision Support Systems, Vol.89, Pp.98-112,

[17] A. Ortigosa, J.M. Martín and R.M. Carro, 2014. "Sentiment analysis in Facebook and its application to e-learning", Computers in human behavior, Vol.31, Pp.527-541,

[18] F. Millstein, 2020."Natural Language Processing With Python: Natural Language Processing Using NLTK", Frank Millstein, Pp.1-116,

[19] A. Jabbar, S. Iqbal, A. Akhunzada and Q. Abbas, 2018. "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach", Journal of Experimental & Theoretical Artificial Intelligence, Vol.30, No.5, Pp.703-723,

[20] M.A. Fauzi, 2018. "Word2Vec model for sentiment analysis of product reviews in Indonesian language", International Journal of Electrical and Computer Engineering, Vol.9, No.1, Pp.525-530,

[21] A.I. Pratiwi, 2018. "On the feature selection and classification based on information gain for document sentiment analysis", Applied Computational Intelligence and Soft Computing, Pp.1-5,

[22] S.V. Georgakopoulos, S.K. Tasoulis, A.G. Vrahatis and V.P. Plagianakos.2018., "Convolutional neural networks for toxic comment classification", Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Pp.1-6,

[23] https://www.kaggle.com/mchirico/cheltenham-s-facebook-group

[24] Ragab et al., (2020). A Novel One-Dimensional CNN with Exponential Adaptive Gradients for Air Pollution Index Prediction, sustainability, 12(23): 10090. https://doi.org/10.3390/su122310090

[25] Nasir et al. (2020). Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training, sensors, 20(23): 6793, https://doi.org/10.3390/s20236793

[26] Huang, et al. (2019). Signal status recognition based on 1DCNN and its feature extraction mechanism analysis. Sensors, 19(9), https://doi.org/10.3390/s19092018

[27] Meliboev et al. (2020). CNN Based Network Intrusion Detection with Normalization on Imbalanced Data, International Conference on Artificial Intelligence in Information and Communication, 19-21 Feb. 2020, Japan. https://doi.org/10.1109/ICAIIC48513.2020.9064976

[28] Li, et al. (2020). A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5), IEEE Access, 8, 26933-26940. https://doi.org/10.1109/ACCESS.2020.2971348

Cite this paper as:
Kalaiarasu M.; Ranjeeth Kumar C. (2022). Sentiment Analysis using Improved Novel Convolutional Neural Network (SNCNN), *International Journal of Computers Communications & Control*, 17(2), 4351, 2022.
https://doi.org/10.15837/ijccc.2022.2.4351