Stance Detection using Extreme Learning Machine with Improved Agglomerative Hierarchical Clustering on Social Network

Saini Jacob Soman, R. Anandan

Abstract--- People can communicate and exchange their views through online social networks and it laid the host for the online social groups. Among numerous groups: privacy violation, groups with no choices of opt-in, disorder were the main problems, which inhibit security of the user and comfort, and we consider user as the member so managing the group principles becomes tedious one. The stylistic, thematic, emotional, sentimental paves the ways for clustering the posts within the group and psycholinguistic rectify these major problems. Stance detection has recently gained significant attention in research and it is sentimental clustering, further it forms a primary segment of the larger research challenge posed from Facebook groups. Recognizing the position of certain Facebook, with regard to the topics given, from user-made text was the main issue addressed in this work. Preprocessing, keyword extraction, stance detection using ELM and clustering using IAHC algorithm were there in this system. Initial one (pre-processing) helps to eliminate the unnecessary data and further assist us to enhance the clustering accuracy in the given dataset. Then for choosing the prominent keywords based on the frequent terms, the keyword extraction was done. Then for classifying neutral and non-neutral posts, Extreme Learning Machine (ELM) paves the way. Further, personal posts are categorized to be positive vs. negative. The members of the group were classified, based on the response to the post of various features. This classification fulfills the performance of huge members present in a strong group by executing the clustering on the basis of linguistic characteristics. For providing sentiments that depend on the posts and the response given by users for the posts, then enhancing the performance of the system, Improved Agglomerative Hierarchical Clustering (IAHC) was established. As observed from the experimental analysis, it is proven that the proposed ELM with IAHC algorithm yields higher performance when compared with the existing methods.

Index Terms--- Stance Detection, Extreme Learning Machine, Social Network, Sentiment Analysis, Improved Agglomerative Hierarchical Clustering.

I. INTRODUCTION

Online social networks paves the way, for providing a link between people using several mutual associations and this online social network it plays an essential role in various applications like email, online shopping, online business networking and instant messaging, in the people's work and life [1].The data addresses the accessibility once after it appears on the online social network. Instead of the traditional search engine, the users generally seek for information from their online social networks. Furthermore, it behave as the popular medium of interaction for sharing, communicating and exchanging enormous details such as text, image, audio, video etc. The users who are all connected in the blog or networks will explicitly view this publicly shared information and it has numerous social impacts in human mind.

From online social networks, blogs, and other kind of media numerous data were created and distributed in the World Wide Web. Very crucial opinion related information helps to be beneficial forth businesses and other features of commercial and scientific organizations. Sentiment analysis is necessary, because there is possibility for manual tracking and extraction.

Sentiment analysis is nothing but an extraction of sentiments or thoughts from reviews stated by users on a specific topic, field or an online product. In order to recognize the subjective information from source data, this analysis is considered to be an application involving natural language processing, computational linguistics, and text analytics. Sentiments were classified as positive or negative one [2]. Hence, it defines the common attitude shown by the speaker or a writer in terms of the topic given.

Excessive messages or sensitive data were there in posting or commenting on specific online social network like Facebook, Twitter, etc. A strong impact was there in these online social networks, which assists us to organize the messages, where it has been expressed publicly through the separation of unnecessary statements.

The methods in [3], utilizes information filtering and it allows the online social network users to have the benefit of directly controlling the posts or comments made in their group. Text pattern matching system helps in accomplishing this work and further it allows the user for filtering out their public space and a license to include new sentences and they were considered to be irrelevant.

Wide range of shared knowledge was the primary issue here, we do not have mechanism to define their validity, and it results in the untrusted information. The experts in the social networks define their validity and examining their knowledge level. Hence, the solution for addressing the issue is from the experts.

More than 90% facts were declared in the Tweets, while false tweets were primarily questioned or denied and this information was given in Mendoza [4]. In natural language, processing, main research hotspot was recognizing the user stance from the massive online text. Furthermore, big corporates such as Facebook look a way for mining the microblog information, for recognizing people's opinion on their products and services

Revised Manuscript Received on November 22, 2019.

Saini Jacob Soman, Research Scholar, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies, Pallavaram, Chennai,India. (e-mail: sainijacobs@gmail.com)

R. Anandan, Professor, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies, Pallavaram, Chennai,India.

provoked by the development in utilizing the micro blogging sites. For instance: at the time of US general election, the debate on people supporting Trump or not would affect the text or posts written online. Identifying the stance detection generally highlighted on debates [5] or news [6].

Stance detection categorizes the attitudes hown by one user in a text on the target provided, i.e. whether a text is in favour of or against the certain target, or neither of them. Which gives huge variation among the stance detection and classical sentiment classification [7]? Initially, the sentiment classification's target always shows up in the text, however as a contradictory, the target of stance detection does not address it openly. Thus, it is significant to categorize the emotion pertaining to the text in the sentiment classification tasks rather than recognizing the stance in the text to the related target. For instance: the text given below suggests a stance made against the target Donald Trump, however the target does not show up anywhere in this scenario.

Example:

The HATE seen within America is stronger NOW rather than I remember since 1969? Cannot go back to such division.

It is necessary to understand the stance pertaining to the text in terms of the unobserved target Donald Trump. With respect to the target towards feature, this instance does NOT provide any view on the target, however it HAS a view regarding something or someone except the target.

An efficient technique was established like filtering method, classification algorithms and clustering approaches, for rectifying the problems in sentiment analysis. Information filtering techniques deals with the massive amounts of dynamically generated data and data or informational sources will be given to the client, which is probable to accomplish his or her information necessity.

Filtering technique, like you tube, Facebook and Twitter, helps to perform the filtering process on social platforms. They generally have huge text like reviews and comments that help usin extracting the view and regulating the adequacy of the contents existing over the internet [8]. In order to filter the document, these techniques work with the similarity value computations, and keyword matching approaches.

We cannot get the accurate result, by separating out the single similarity value. Next, though techniques give accurate result as a domain specific, we make use of the expert's system approach. Therefore, while enforcing the same technique on social platforms [9] we face huge issue. Likewise, the overall filtering approaches use keywords matching between the knowledge base and the filtering contents. When we filter the content with no knowledge on the semantics of the text, incorrect results will be acquired, since the word usage depends on the link or context presented by the information. Classification methods were established for maximum accuracy of given dataset or documents, in order to rectify this issue.

The sentiment is classified as positive or negative, based on the sentiment classification (e.g., a product review). Furthermore, it does not examine or extract any details in the document. Here the document-level sentiment classification is nothing but a task and it is expressed in the form of a supervised learning problem having two classes (positive and negative) or the scores of rating [10].

In order to enforce the classification process, standard supervised learning techniques such as naive Bayesian classification and support vector machines (SVM) was utilized. Here we have another approach called unsupervised approaches for doing the document sentiment classification, which works according to the sentiment words and language schemes. It is highly evident that sentiment words (also known as opinion words) points to the positive or adverse sentiments (e.g., good and nice fall under positive sentiment words, and horrible and bad fall under negative sentiment words) and further it play a significant part in sentiment classification. The primary drawback here is recognizing the sentiment associated with the keywords. When other forms of the solutions concerning the sentiment, the problem in detection occurs in various scenarios, involving the stance detection on social platforms, until now, only limited research were done.

For improving the accuracy in evaluating the sentiments linked with the keywords, a proper stance identification algorithm has to be executed. The above addresses issue will be rectified by preprocessing is initially applied to filter repeated and noise content for the given dataset.

Then to select the prominent keywords, keyword extraction is performed and ELM helps to stance detection, where the posts were classified as neutral and non-neutral. Then IAHC algorithm helps the overall clustering accuracy for positive, negative and neutral posts. Enhancing the stance detection efficiency and clustering accuracy using proposed ELM with IACH algorithm significantly were the main contribution of this research.

II. LITERATURE REVIEW

In this new era, social networks like Twitter and Facebook were loaded with opinions and Rathi et al (2018) presented this information.

Twitter (commonly utilized micro-blogging platform) where people share their opinions as tweets and so it has emerged to be one of the best possible sources for carrying out sentimental analysis. Views in twitter are classified as good for positive category, bad for negative and neutral categories. Sentiment Analysis examines various opinions and group them in all these classifications. For classifying the emotions, the tweet was gathered.

Ensemble machine learning technique helps for maximizing the effectiveness and robustness of this technique, and for enhancing the classification outcomes in the field of sentiment analysis.

This research combines Support Vector Machine (SVM) and decision tree, their experimental analysis is performed with respect to f-measure and accuracy, and it confirms that the approach gives better classification results in contrast to individual classifiers.

A category of predefined polarity terms enforces a user rating, in addition to a sentiment analysis method for recommending the educational content in social environments and Karampiperis et al (2013) discussed it in [12].

Prior to making a recommendation of the content to another person, the text analyses have to be done and polarity is refereed through sentiments, since accurate result will be given by this technique. For filtering the content, user ratings will be used by the Recommender systems, which may not be a trustworthy source of filtering, since there always exists a probability of generating incorrect ratings because of the existence of content promoters.

Geetika et al (2014) presented the machine learning algorithms for sentiment analysis in [13] and the customers review classification support us to examine the information from Twitter dataset. However, it is extremely unorganized and is either positive or negative, or someplace in between these two.

So, initially step is pre-processing the dataset, next step is extracting the adjective from the dataset, which imply something meaningful and it is termed as feature vector. After this, the list of feature vector is chosen and machine learning based classification algorithms such as: Naive Bayes, Maximum entropy and SVM along with the semantic orientation based WordNet later enforced, which acquires the meanings and correspondence for the content feature. At last, with respect to recall, precision and accuracy, the classifier's performance is computed.

An efficient neutral stance detection model was utilized by Gao et al (2018) in [14], over multiple stages of a text, it considers the target and target towards information and it creates the target and aims towards information based highlights. The target information will be considered only the traditional attention-based neural network models. On the contrary, the model considers the benefit of the target in addition to the target towards information and it rectifies the stance detection task in a better manner.SemEval-2016 Task 6 dataset helps to do the experimental analysis and the results discloses that the model surpasses numerous strong baselines.

An approach was described by Vijayaraghavan et al (2016) in [15], for the identifying stance in tweets and makes use of the recent advances in short text categorization through the deep learning, for generating the create word-level and character-level models.

With the help of validation performance, the choice between word-level and character level models in each specific case was informed. The novel data augmentation techniques was enforced by this system, for expanding and diversifying training dataset, therefore making the system more robust. A macro-average precision, recall and F1-scores was accomplished by this approach.

Zhai et al (2012) established extreme Learning Machine (ELM), in [16] an da feature selection algorithm is used here, which incorporates a criterion for feature-ranking for computing the importance of a feature through the computation of the combined difference of the results of the probabilistic single-hidden-layer feed-forward neural networks (SLFN) with and with no feature.

Then the SLFN is trained using ELM, selects the weights of hidden layer in random and logically controls the weights of the output layer. The experimental analysis confirms that ELM method is effective and efficient.

Dias et al (2016) in [17] gave a weakly supervised technique for stance detection in tweets that are just content-based. Based on a group of heuristics the approach depend on automatic labelling of tweets with respective to stance, serving an objective that is twofold: a) a supervised learning algorithm helps for an automatic generation of a training corpus for establishing a predictive model and b) determining the stance of tweets in order to complement the predictive model .

The algorithm examines the performance of the technique taking six unique stance targets and a potential output will be accomplished yielding a weighted F-measure differing from 52% to 67%.

Lin et al (2013) in [18], mined book review text for recognizing the insignificant features of a set of identical books. This makes the comparisons among the books by searching for books having the same features; eventually carrying out clustering on the books present in this data set and it makes use of the same process of mining for recognizing a respective set of features in the users. At last, it computes the quality of techniques by analysing the association among the similarity metrics, and user ratings

Coban et al (2018) examined the applicability of "word2vec and clustering based text representation" method for Twitter sentiment analysis in [19]. It conducted experiments on two different datasets that are comprised of Turkish Twitter feeds from which one is subject-dependent and the other one is subject-independent.

Support Vector Machine (SVM) algorithm is utilized in classification phase. Experimental results confirms that the W2VC has been quite successful and has provided an incredible advantage with respect to time and performance as it decreases the feature space, but it does not provide enough success with respect to accuracy.

III. METHODOLOGY

To progress the stance detection performance more effectively and accurately, ELM with IAHC is proposed. For improving the accuracy, stance identification algorithm has to be executed for the given Facebook group by giving neutral and non neutral posts efficiently.

3.1 Preprocessing

Cleaning (basic step of pre-processing) eliminates the hyperlinks and repeated posts in the messages. Since it primarily put an effort to cluster the posts according to various aspects. The linked hash tag will be trimmed by the messages. Crawler will combine the Facebook data and it had various means of representing punctuations. Therefore, for replacing back the punctuations into the text, we need to change the data.

For stance detection, the pre-processed text is sent to the next stage

Stop word removal is another important process in pre-processing. This step confirms to get rid of the words without any meaning (like prepositions, article etc.). An online resource helps carrying out stop word elimination, which again is a module of the Stanford NLP resources [20]. The process of stemming is performed on the recognized tokens and in the user-created Facebook content words, which helps both the primary tokens and also the stemmed

tokens. We proceed the stemming process through well-accepted porter stemmer [21].

3.2 Keywords Extraction

Once after completing the pre-processing process the posts were fed into the keyword extraction process. Initial step is extracting the frequent terms and thereafter a set consisting of co-occurrences among every term and the frequent terms, i.e., occurrences within the same sentence, were produced. Co-occurrence distribution indicates the significance of a term. In case the probability distribution of co-occurrence between term *a* and the frequent terms is inclined towards a specific subset of frequent terms, then term *a* is likely to become a keyword. Through the extraction of the suitable keywords, it can simply selects the content of the document for reading and learning the correlation existing among documents [22]. "Automatic term recognition" is nothing but a comparable research topic and it termed as the usage context of computational linguistics and "automated indexing" or "automated keyword extraction" in the research field of information retrieval. Contents of the sentimental status differs the posts made in a network group. The post's status can be positive, negative or neutral. At keyword stage, we perform the assessment of the sentiment. This divides the posts that depend on the attitude shown towards the trends. A member's response reveals his/her attitude towards the subjects. An example below illustrates this:

User A says "I love ice cream. I Love to have it all time"

While User B says "I hate ice cream. I don´t want it at all "In this conversation, the keyword is ice-cream and the sentiment of the keyword with regard to user A is positive whereas it is negative for user B. Therefore, it examines the sentiments related to the primary keywords related to a group. The crucial words in a post will be formed by these keywords, for defining an article's subject.

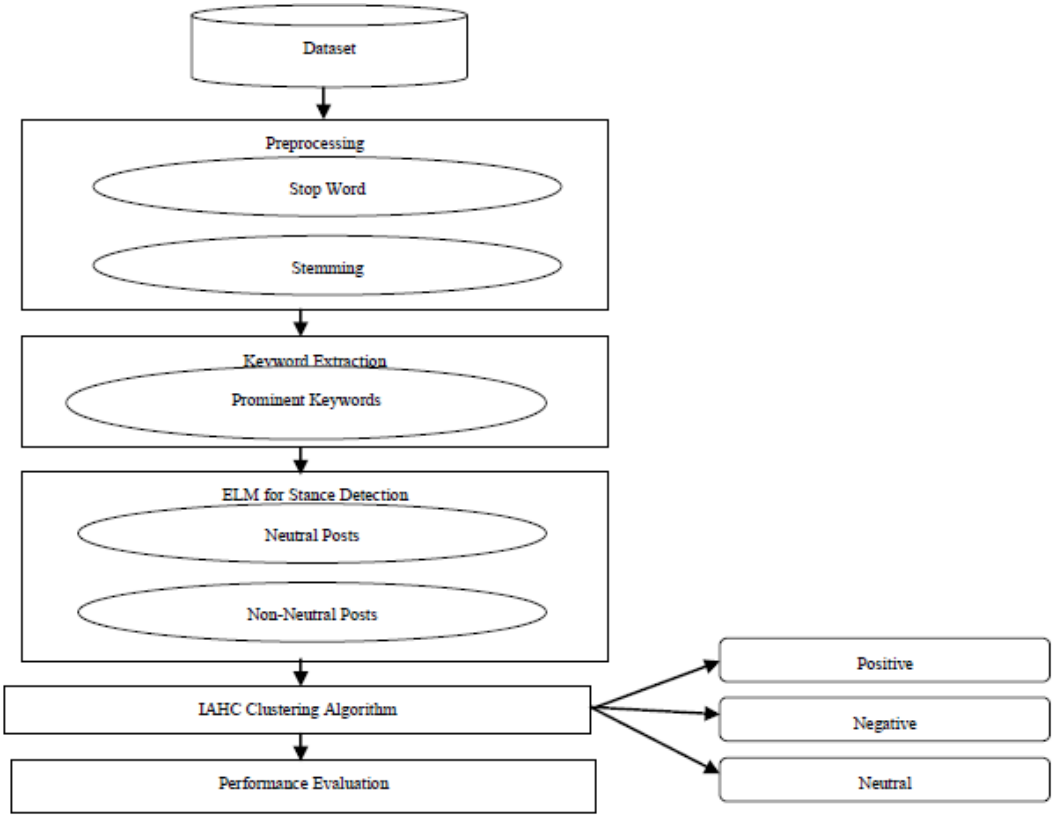

Figure 1: Overall configuration of the proposed system

The lists of keyword connected with all the posts are combined together. The finalized keyword list will be ready; by separating out those keywords, that has much smaller usage in posts. This is due to the fact that here the focus is much on the sentiments linked with the keywords that are found in all the posts. The variations observed in the sentimental scores related to the keywords can be an important factor, while considering the sentiment posts. The overall block diagram of the proposed system is shown in Fig 1.

In Algorithm 1 explains the process of extracting prominent keywords, which provides the output as the predominant keywords K. The threshold differs according to the dataset.

$$\chi^2 = \sum_{p\epsilon P} \frac{(freq\,(w,p) - n_w k_p}{n_w k_p} \quad (1)$$

$n_w k_p$ Indicates the expected frequency of co-occurrence

$(freq(w,p) - n_w k_p$ Indicates the difference between observed and expected frequencies

The phrase, words, sentences, concepts, and paragraphs are acquired and the consequence of recommended post is found through the term frequency (tf) and inverse term frequency (idf).

$$tf_{phrase,sentence,concept,paragrap\ h} = \frac{phrase_P, sentenc\ e_P, concep\ t_P, paragrap\ h_P}{phrase_{Total}, sentenc\ e_{Total}, concep\ t_{Total}, paragrap\ h_{Total}} \quad (2)$$

$phrase_P$Points to the frequency of occurrence of the provided phrase in the post, $sentence_P$ points to the frequency of occurrence of sentence provided in the post, $concept_P$points to the frequency of occurrence of the concept provided in the post and $paragraph_P$ points to the frequency of occurrence of given paragraph in the post. $phrase_{Total}$ points to the overall number of occurrence of phrase over the total group, $sentence_{Total}$ points to the overall number of occurrence of sentence over the whole group, $concept_{Total}$ points to the overall number of concept occurrence on the whole group and $paragraph_{Total}$ points to the overall number of occurrence of paragraph in the total group.

$$idf_{phrae,sentence,concept,paragrap\ h} = log_e \frac{P_{Total}}{P_{phrae,sentence,concept,paragrap\ h}} \quad (3)$$

P_{Total} refers to the overall number of posts existing within the group and $P_{phrae,sentence,concept,paragrap\ h}$ indicates the number of posts with the$phrae, sentence, concept, paragraph$. Given.

$$tfidf_{phrae,sentence,concept,paragrap\ h} = tf_{phrase,sentence,concept,paragrap\ h} \cdot idf_{phrae,sentence,concept,paragrap\ h} \quad (4)$$

Algorithm 1

1. Procedure: Major keyword extraction
2. P ← each post in a given group
3. K ←∅
4. For p∈ *P* do
5. p_keywords←Extract_keywords()
6. for p_key∈ K_keys then
7. K[p_key]+=1
8. Else insert (K, p_key)
9. For key∈ K do K[Key]< threshold
10. Eliminate (K, key)
11. Apply frequent keyword using eq (1) and (4)
12. Return k

The sentimental scores corresponding to all the posts to these keywords are decided, once after the prominent keywords have been extracted, and it gives information of the sentimental score (positive, negative, or neutral) related with every keyword for every post and the result help as the feature set for sentimental clustering.

A cautious policy mechanism is required for the filtering posts within the group, which validates the compatibility existing within a group. Simultaneously, the recommendation made to a group member with posts irrelevant to them and it needs ratings from other reliable group members. We can predict the status by the social attraction and relativity between the members present within a group.

3.3 ELM for Stance Detection

ELM's significant stage is Learning the parameters of hidden nodes, inclusive of the input weights and biases, which are assigned in random and it is not required to be tune, whereas the output weights can be mathematically decided applying the simple generalized inverse operation [23]. The only parameter was the number of hidden nodes, which has to be determined. ELM provides immensely faster speed of learning, better generalization performance and it enhances the training of single hidden-layer feed forward neural networks (SLFNs).

The dataset considers the lexical, syntactic, semantic and pragmatic challenges along with the Facebook comments; here the tasks were classified into two phases: robust and intuitive solution. The target of the initial phase is, recognize the comments with neutral stances in terms of the target subjects, and filters out the non-neutral (favour/against) stances out of the neutral ones. The aim of the next phase is to carry out a classification between the positive (favour) and negative (against) stances, and the non-neutral comments.

Given n distinct training features $(xi, ti) \in R^n \times R^m (i = 1,2, \ldots..N)$the output of a SLFN with $\bar{N}$hidden nodes (additive features) can be described by

$$o_j = \sum_{i=1}^{\bar{N}} \beta_i f_i(x_j) \quad (5)$$

$$\sum_{i=1}^{\bar{N}} \beta_i f(x_i; a, b), \qquad j = 1, \ldots. N \quad (6)$$

where$o_j$refers to the output vector of the SLFN in terms of the input sample $x_i.a_i = [a_{i1,}a_{i2,\ldots\ldots}a_{in}]^T$and $b_i$are learning parameters generated in random of the jth hidden node, correspondingly. $\beta_i = [\beta_{i1,}\beta_{i2,} \ldots. \beta_{im}]^T$is the link that connects the jth hidden node and the output nodes. $f(x_j; a_i, b_i)$ refers to the activation function of the actual ELM. Set $a_i.x_j$be the inner product of ai and xj. Equation (7) can be expressed in compact form as

$$h\beta = o \quad (7)$$

where

$$h = \begin{pmatrix} f(a_1.x_1+b_1 \ldots\ldots f(a_{\bar{N}}.x_1+b_{\bar{N}} \\ \vdots \\ f(a_1.x_N+b_1 \ldots\ldots f(a_{\bar{N}}.x_N+b_{\bar{N}} \end{pmatrix}_{N\times\bar{N}} \quad (8)$$

Here, H is known as the output matrix of the hidden layer.

ELM theories states that, without considering the input data, the hidden nodes' learning parameters ai and bi can be assigned in random, which, in turn, minimizes the network cost function $\|o - T\|$. Then, equation (2) tends to become a linear system, and the output weights βcan be analytically decided by recognizing a least square solution as given below

$$\hat{\beta} = h \times T \quad (9)$$

Any long training phrase will be eliminated, where the network parameters were iteratively adjusted with few suitable learning parameters (like the learning rate and iterations).

Algorithm 2

Input: $(xi, ti) \in R^n \times R^m (i = 1,2, \ldots..N)$
Activation function f, hidden node number $\bar{N}$
Output: the output weights β
Step 1. Randomly allocate the parameters of hidden nodes

$$(a_i, b_i) i = 1, \ldots \bar{N}$$

Step 2. Compute the output matrix of the hidden layer h.

Step 3: Re-compute

$$R(a_i) = \sum_{k=1}^{C} \sum_{x \in T} \left| p(w_k/x) - p'(w_k/x) \right| \quad (10)$$

Step 4. Compute the output weight

$$\beta: \beta = h \times T. \quad (11)$$

Thus, the algorithm provides neutral and non-neutral, for the given posts. $p(w_k/x)$and$p'(w_k/x)$ indicates the posterior probabilities between k[th] words and input posts (x) in the testing dataset. The performance achieved of stance detection was enhanced by taking the target information of the text features as well as target towards information into consideration. Text level stance detection targets to estimate the stance distributions based on their text information, target and also the target towards information.

Significant parts of the sentences were captured and it is obvious that not all words can give equal meaning to the sentence. Hence, it adopts a target and target towards attention strategy for the extraction of significant information of the sentence, in word level, rather than sending the hidden states to an average-pooling layer. Finally, it aggregates the representations of that particular information for representing the sentence.

3.4 IAHC Algorithm for Clustering

In this section, for deciding on the strongest cluster, IAHC algorithm is proposed on a social network of posts that is obtained. The cluster incorporates the neighboring points that are near enough to time and frequency. Every cluster makes for a probable candidate and it works according to the existence criterion. Agglomerative hierarchical clustering is a bottom-up technique under hierarchical clustering and here each cluster has sub-clustered. [24]. We can able to decide the cluster hierarchy, by choosing the last cluster number either by chopping the tree at an absolute cluster joining distance or at a jump in the join distance values. Complete-linkage clustering or the maximum distance between the components of every AHC cluster. The single-linkage clustering or minimum distance among the elements of every AHC cluster. The average linkage clustering is nothing but the average distance between components of every cluster. This algorithm generates less computational complexity and more clustering accuracy results.

If the clusters are at a much farther distance in order to be combined (distance criterion) or if there is an adequately less number of clusters (number criterion), the clustering process can be terminated, because of the distance between the cluster were high when distinguished with the previous agglomeration. In a single cluster, it begins with every single object (posts). Thereafter, in every sequential iteration, it accumulates the nearest pair of clusters by fulfilling some similarity condition, until entire data falls into one cluster. Smaller clusters were created, which helps to define the similarity between the two objects (posts). However larger dataset creates an issue with AHC and hence semantic similarity and frequent occurrence values, enhances this problem. Thus, it can generates an order for the posts, which might be meaningful for displaying the data and it gives higher accurate clustering results [25].

Discovering positive, negative and neutral posts for the given dataset, more efficiently and accurately, was the main aim of the proposed algorithm. According to their response to the posts, then the members of the group were categorized, which owes to different aspects. Elementary segment's sentences make sense for the algorithm. The paragraph is a common linguistic arrangement that indicates a uniform textual portion. Boundary found in the middle segment of the sentence is therefore contrasting to the objective of the author. Furthermore, for the proximity test, the size of a paragraph, comprises of adequate lexical information. The proximity test chooses the nearest pair of segments, which lead to define the events. The test depends on the repetition of words, a well-identified indicator for lexical structure.

$$proximity(w_i, w_i + 1) = \sum_{k=1}^{m} \frac{n(w_{k,i}).n(w_{k,i}+1)}{\|w_i\|\|w_i+1\|} \quad (12)$$

Where $n(w_{k,i})$refers to the number of words length on the given dataset and $w_{k,i} + 1$ is similar words in the same dataset

The algorithm recognizes similar kind's more than two posts and merges them into one cluster. By using minimum distance, the similar positive posts and similar negative posts were measured and the maximum distance, average distance and mean distance for the given dataset. Hence comparing the two nearby posts and it gives positive, negative posts effectively through IAHC algorithm.

Algorithm 3

Input: Dataset

Output: Positive, Negative and Neutral

Step 1: Read the given dataset

Step 2: Perform preprocessing with stop words and stemming

Step 3: for the three classes do

Step 4: for all pairs in every class do

Step 5: M← similarity between words from different cluster

Step 6: N← similarity between sentences from different cluster

Step 7: O←similarity between lines from different cluster

Step 7: Apply keyword matching between posts and obtain frequent terms

Step 8: Compute semantic similarity using wordnet W

Step 9: FinalSim←M+N+O+W

Step 10: end for

Step 11: do IAHC clustering $IHAC = (linkage = 'min', 'max', 'avg')$connectivtiy=none, nclusters, =clusters)

$$D(c1, c2) = \frac{1}{c1,c2} \sum_{x1,c1} \sum_{x2,c2} D(x1, x2) \quad (13)$$

Step 12: Compare FinalSim and D(c1,c2)

Step 13:Merge the two clusters and adjust the features of the new cluster

Step 14: Provides positive, negative and neutral posts

Algorithm 3 explains the rough outline of IAHC clustering technique. The semantic similarity computation is carried out for the extracted segments of entire features and then by summing all four value we can calculate the final similarity score. A clustering algorithm groups similar segments, once after acquiring the similarity matrix for each class. IAHC is considered as the successful technique for text and document clustering. Here, minimum, maximum, and average linkage proved

as a suitable one for stance identification. After the right clustering approach is found, then this condition is considered as the similarity matrices for the clustering element. The IAHC process iterate still it attains a pre-set threshold.

IV. EXPERIMENTAL RESULT

From Cheltenham Facebook Groups [26], the dataset has been collected and the experimental analysis was done on the online social networks to suite the real world environment. The three open groups performs the tests. The data set yield a rigorous structure in the aspect that data matrix acquired for like-share rating is small. Also, the available content for linguistic analysis is quite meagre. Here we proceed two set-ups process. First step confirms the efficiency of the proposed technique in rendering personalize recommendations of posts to yield amess free environment in the group. Second one confirms the suitability of proposed technique determines the admiration of members within the group. In this section, the existing decision tree and SVM algorithms are considered to calculate the performance metric against proposed ELM with IAHC algorithms.

Accuracy

Accuracy determines the overall accurateness of detection results and is considered as the addition of the classification parameters $(tp + tn)$ divided by enitre number of detection with classification parameters $(tp + tn + fp + fn)$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (14)$$

Where tp, $tnfp$ and fn are numbers of true positive, true negative, false positive and false negative correspondingly.

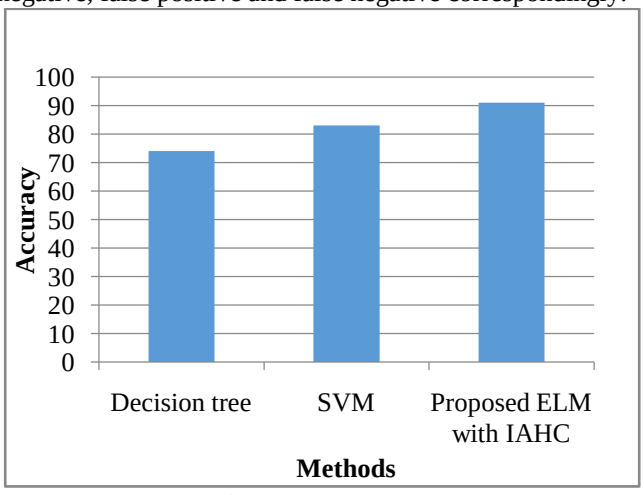

Figure 2: Accuracy

Fig 2, shows the comparison metric which is computed through the available and proposed technique in terms of accuracy. The methods were plotted along the x-axis and the accuracy value is plotted in y-axis. For the given dataset, the existing methods are such as SVM and decision tree algorithm provides lower accuracy whereas proposed ELM with IAHC approach provides higher accuracy. The result confirms that the proposed ELM with IAHC enhances the stance detection process by recognizing the neutral and non-neutral posts efficiently for the given dataset. The cluster method helps to maximize the classification accuracy and it gives accurate positive, negative and neutral posts for the given dataset.

Precision

By treating Posts as positive samples in the binary classification, precision is defined as,

$$precision = \frac{tp}{tp + fp} \quad (15)$$

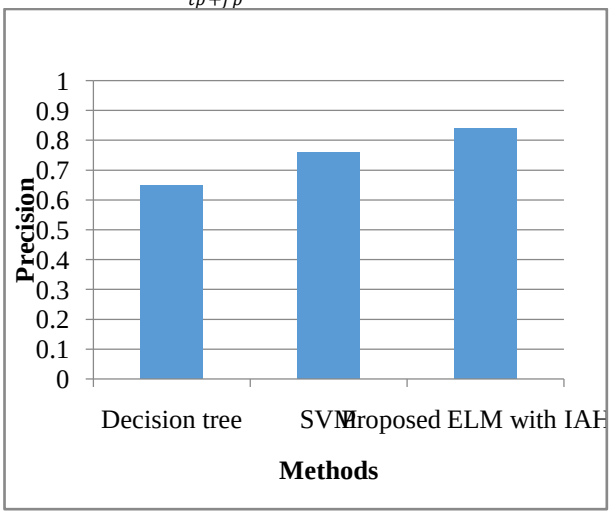

Figure 3: Precision

Fig 3, shows the comparison metric which is computed through the available and proposed techniques in terms of precision. The methods were plotted along the x-axis and the precision value is plotted in y-axis.

For the given dataset, the existing methods are such as SVM and decision tree algorithm provides lower precision whereas proposed ELM with IAHC approach gives higher precision. The output confirms that the proposed ELM with IAHC enhances the stance detection process by recognizing the neutral and non-neutral posts efficiently for the given dataset.

Recall

The calculation of the recall value is done as follows:

$$Recall = \frac{tp}{tp + fn} \quad (16)$$

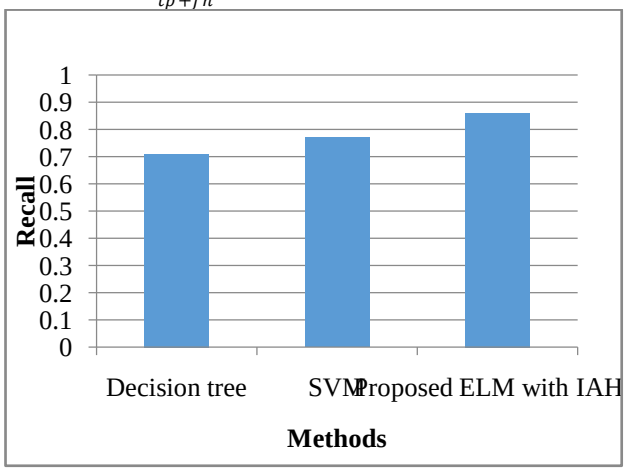

Figure 4: Recall

Fig 4, shows the comparison metric is computed through the available and proposed technique in terms of recall. The method was considered in x-axis and the recall value is plotted in y-axis. For the given dataset, the existing methods are such as SVM and decision tree algorithm provides lower recall whereas proposed ELM with IAHC approach provides higher recall. The output confirms that the proposed ELM with IAHC enhances the stance detection process by recognizing the neutral and non-neutral posts efficiently for the given dataset.

F-measure

$F1$-score is computed as:

$$\mathrm{F1-score} = \frac{2\times \mathrm{precision}\ \times \mathrm{recall}}{\mathrm{precision}\ + \mathrm{recall}} \quad (17)$$

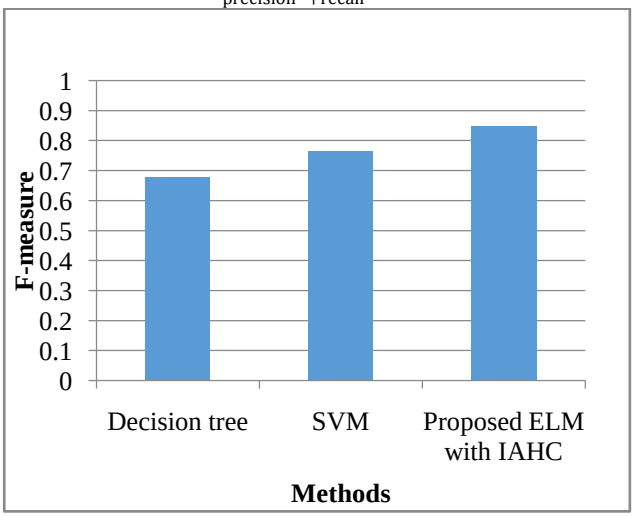

Figure 5: F-measure

Fig 5, shows the comparison metric is computed through the available and proposed technique in terms of F-measure. The techniques were considered in x-axis and the F-measure value is plotted in y-axis. For the given dataset, the existing methods are such as SVM and decision tree algorithm yields a much lesser F-measure while the newly introduced ELM with IAHC approach provides higher F-measure. The output confirms that the proposed ELM with IAHC enhances the stance detection process by identifying the neutral and non-neutral posts efficiently for the given dataset.

V. CONCLUSION

For efficient stance identification in the given Facebook group dataset, in this research work, ELM with IAHC clustering algorithm is proposed. Keyword extraction helps to mine the significant keywords through the frequent term measure. Hence, it chooses the prominent keywords from the dataset and further it helps to categorize the posts effectively.

ELM algorithm is proposed for stance detection and it categorize the posts as neutral and non-neutral posts. For accurate classification IAHC algorithm is enforced, whether the posts are found to be positive, negative or neutral in the given dataset. Thus, the proposed ELM with IAHC algorithm gives greater performance with regard to higher values of accuracy, precision, recall and f-measure. Hybrid optimization algorithm will be established as the future work for enhancing the various aspects of solution for stance detection.

REFERENCES

1. Guo, Liang, et al. "A hybrid social search model based on the user's online social networks." *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*. Vol. 2. IEEE, 2012.
2. Blair, Stuart J., Yaxin Bi, and Maurice D. Mulvenna. "Sentiment Classification of Social Media SContent with Features Generated Using Topic Models." *STAIRS*. 2016.
3. Kardan, Ahmad, Amin Omidvar, and FarzadFarahmandnia. "Expert finding on social network with link analysis approach." *2011 19th Iranian Conference on Electrical Engineering*. IEEE, 2011.
4. M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 71–79.
5. M. A. Walker, P. Anand, R. Abbott, and R. Grant, "Stance classification using dialogic properties of persuasion," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 592–596.
6. W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2016, pp. 1163–1168.
7. B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.
8. S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings, The International World Wide Web Conference Committee (IW3C2) Std., April 2010
9. B. Guc, "Information filtering on micro-blogging services," Master's thesis, Swiss Federal Institute of Technology Zurich, August 2010.
10. Narayanan R, Liu B, Choudhary A (2009) Sentiment analysis of conditional sentences. In: Proceedings of conference on empirical methods in natural language processing (EMNLP-2009), Singapore
11. Rathi, Megha, et al. "Sentiment Analysis of Tweets Using Machine Learning Approach." *2018 Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, 2018
12. P. Karampiperis, A. Koukourikos, and G. Stoitsi, "Collaborative filtering recommendation of educational content in social environments utilizing sentiment analysis techniques," Recommender Systems for Technology Enhanced Learning: Research Trends & Applications, vol. RecSysTEL Edited Volume, Spinger, 2013
13. GeetikaGautamDivakaryadav, "Department of Computer Science &Engg. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis" IEEE 2014
14. Gao, Wenqiang, Yujiu Yang, and Yi Liu. "Stance Detection with Target and Target Towards Attention." *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2018.
15. P. Vijayaraghavan, I. Sysoev, S. Vosoughi, and D. Roy, "Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns," pp. 413–419, 2016.
16. Zhai, Meng-Yao, et al. "Feature selection based on extreme learning machine." 2012 International Conference on Machine Learning and Cybernetics. Vol. 1. IEEE, 2012.
17. Dias, Marcelo, and Karin Becker. "An heuristics-based, weakly-supervised approach for classification of stance in tweets." *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2016
18. Lin, Eric, Shiaofen Fang, and Jie Wang. "Mining online book reviews for sentimental clustering." *2013 27th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2013.
19. Çoban, Önder, and GülşahTümüklüÖzyer. "Word2vec and Clustering based Twitter Sentiment Analysis." *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. IEEE, 2018.
20. Li, Weifeng, et al. "ELM combined with hybrid feature selection for classification." *2018 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2018.
21. Paik, Jiaul H., et al. "GRAS: An effective and efficient stemming algorithm for information retrieval." *ACM Transactions on Information Systems (TOIS)* 29.4 (2011): 19.

22. Batool, Rabia, et al. "Precise tweet classification and sentiment analysis." *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. IEEE, 2013.
23. Chaturvedi, Iti, et al. "Bayesian network based extreme learning machine for subjectivity detection." *Journal of The Franklin Institute* 355.4 (2018): 1780-1797.
24. Bouguettaya, Athman, et al. "Efficient agglomerative hierarchical clustering." *Expert Systems with Applications* 42.5 (2015): 2785-2797
25. Dai, Xiang-Ying, et al. "Online topic detection and tracking of financial news based on hierarchical clustering." *2010 International Conference on Machine Learning and Cybernetics*. Vol. 6. IEEE, 2010.
26. D.M.Blei, A.Y.Ng, M.I.Jordan, "Latent dirichletallocation",Journal of Machine Learning Research,2003,3, pp.993â˜A ,S1022.

AUTHORS PROFILE

Saini Jacob Soman holds M.Tech Degree in Computer and Information Technology. He is currently pursuing his research studies in Computer Science and Engineering at Department of Computer Science and Engineering, School of Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. He has an experience of 16 years in academic level. His area of interests are Data Mining and Machine Learning.

He has published nearly 10 research papers in various International Journals and conferences He has received a national award in 2017 for his contribution in academic field. He published 4 text books for kerala state higher secondary syllabus in computer science/ Applications.

Dr. R. Anandan possesses Doctoral degree in Computer Science and Engineering. He is currently working as Professor, Department of Computer Science and Engineering, School of Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India . He has vast experience in corporate and all levels of Academic in Computer Science and Engineering.

He associated as Member in many reputed International and National societies. He serves as Editorial Board Member / Technical Committee/ Reviewer in many International Journals. He has published more than 90 research papers in various International Journals and received 12 awards and filed 3 patents. He published 7 books in Computer Science and Engineering discipline.