Selecting Optimal Feature Subset in Patient Care Opinion Mining Using Extended Bag of Words

Keerthika R.[1*] and Nalini C.[2]
1. Department of Information Technology, Karpagam College of Engineering, Coimbatore 641032, Tamil Nadu, INDIA
2. Department of Information Technology, Kongu Engineering College, Perundurai (PO), Erode -638052, Tamil Nadu, INDIA
*keerthikait@gmail.com

Abstract
A branch of natural language processing or methods of machine learning is known as Sentiment analysis for identifying sentiments in text. Bag-Of-Words technique serves to be one of the most prevalently implemented technique, it has two weaknesses; use of a physical evaluating process in lexicon to identify the estimation of the words and the analysis of sentiment that have less accuracy owing to ignoring the impacts of grammar in words and ignoring its semantics. A new optimization algorithm called charged system search is used for improving the efficacy of the bag of words model. In this work, a novel feature selection technique using Charged System Search is proposed and is compared with MRMR, GA feature selection. The term frequency is used for the extraction of the features from the blogs that are available in many medical websites. The results prove the importance of feature selection in the improvement of accuracy of classification.

Keywords: Opinion Mining, Bag-of-Words (BoW), Minimum Redundancy and Maximum Relevance (MRMR), Genetic Algorithm (GA), Charged System Search (CSS), Cosine, Term Frequency.

Introduction

An Opinion Mining or a Sentiment Analysis includes building a system that is used for exploring the opinions of the users expressed in the posts of blogs, the comments, tweets or reviews about the products, events, services or policy. The aim is to identify the user attitude on some topic. Recently, the increase in the usage of the internet and user opinion exchange has been the motivating factor for Opinion Mining. The web is that huge repository of both unstructured data as well as structured data. Analyzing the information to extract the user opinion and its sentiment is quite challenging. An opinion is that quadruple that consists of a topic, a holder, a claim and a sentiment. The holder here trusts a claim on the topic and depicts it using a Sentiment [1]. Opinion mining is a recent research area in language processing.

This has quickly grown into a research area for economic reasons and this develops around a basic number of approaches. The approaches include the "bag-of-words" models and the sequence models (e.g. the HMMs), but have not been appropriate for all situations especially in the genres in which there are many potential opinion statements that are identified within the same stretch of text. This is a relevant problem in the sentiment analysis technique expansion to those areas like market prediction and social science that is based on corpus. In such areas, it is not sufficient for prediction or detection in the predefined areas and in cases of mining the locations of the opinions of the large corporations [2].

The Sentiment analysis has to generate big lexicons that take time so that necessary words are being searched. Few various problems faced by the Sentiment Analysis when analyzing the views, their meanings as well as grammar in a particular language or many languages. Specifically in cases which have many words with various meanings and polarities. So it involves the identification of the meaning of sentiment, the expressions, the polarity, the strength and the expressions. The linguistic resources has a volume that is enormous [3]. The various classes of sentiment analysis are:

Positive Sentiments: These denote the good words on the target considered. In case there are positive sentiments that are raised it is an indication of something good. For commodity reviews, if positive reviews are more it is bought my many customers.

Negative Sentiments: These denote the bad words on the target in deliberation. In case the negative sentiments are on the increase, it gets discarded from the optional list. For commodity reviews if there are more negative reviews no one will buy them.

Neutral Sentiments: These denote the words that are neither good or bad and so they are not adopted or depreciated.
A Sentiment classification is one essential task in the sentiment analysis that has the aim of classifying the given task's sentiment. The most familiar practice in this classification uses the technique in a conventional and topic based classification known as Bag-of-Words (BoW) that is used typically for representation of text. Many researches in this sentiment analysis tries to appreciate the BOW by means of consolidating the linguistic knowledge [4]. This BoW model represents the method used in processing of natural language for processing language and retrieval of information in which a text gets shown like a collection of words without regarding the grammar. The bag-of-words model is made use in the methods of classifying the document in which frequency every word is used as a training feature. The purpose of using bag-of-words to implement is for these reasons.

- For extracting words from a text or document or sentence easily

- To get the right weightage of words or features in documents, which can help identify the text category that aids in classification.

This further improves the classifier performance if combined with the classification known as naïve bayes to keep the text filtered for the word or feature of weightage that uses the BoW and later trains the classifier by using these words [5]. This BoW approach denotes a document that uses the frequency of occurrence of the words not considering the word order. Semantic similarity that exists between documents is captured by means of comparing the frequency of profiles or the representation vectors. But BoW does not capture similarity of synonyms [6].

Feature selection is that process of choosing a subset of features or attributes at the time of model construction. Normally a dataset can consist of various attributes which may not contribute significantly or be informative as the others. So it is a sensible choice to use only a handful of features belonging to the sets. Various approaches are adopted for the selection of features. The bag of words will hold a set of features that is chosen from a Labelled trained data [7]. This holds the set of features that are divided into methods that are lexicon based needing human annotation and statistical methods are used for the same.

These techniques treat the documents as BOW or string that retains the word sequence in the document. Owing to the simplicity of this process the BOW is regularly used. A commonly used selection step is removing the stop words and the stemming]8]. The aim is to identify a subset for sentiment classification and summarization of the problem using Particle Swarm Optimization (PSO). To make this as powerful as possible the PSO's performance has to be improved.

The PSO summarizes customer reviews and transfers the texts into the feature vectors and appropriate feature selection is performed. This is often seen in genetic algorithms (GA) and uses a fitness sum for the multi-objective function during the step of selection. The multi-objective function includes either minimizing principles or the maximizing principles. The features are then extracted by using a multi objective optimization for generating summary. The main advantage here is the diversity of the swarm and improve its search ability [9]. For this a technique known as Charged System Search in which section 2 shows related works, section 3 explains methods used, section 4 gives experimental results and section 5 gives the conclusion.

Related Works

Tu et al., [10] made a presentation of a technique to automatically detect the damaged roof-tops depending on bag-of-words technique. The damaged rooftop's building is initially split to various super-pixel regions. The BOW scheme is then used to frame the corresponding vectors and the un-damaged regions for every area. Lastly the damaged and the non-damaged parts are differentiated with the help of a Support Vector Machine. The results are reviewed based on experiments for a chosen site of the ruins of the Beichuan Earthquake in Sichuan, China, and hence proved that our technique was possible to detect the damaged roof-top regions.

Zhu et al., [11] made an introduction of a traditions bag of words and this was made use in an aircraft type of classification, and an optimized bag-of-words model that is used instead of the conventional bag-of-words that depend on the normal SIFT sampling and the K-means clustering that is presented depending on space partition of the SIFT sampling and the FCM clustering and is further use to recognize the type of aircraft. The experimental outcomes proved that the optimized bag-of-words is able to maintain a high rate of recognition in the classification type for the aircraft classification than that of the conventional bag-of-words and the Affine moments that originated the pictures of aircraft as well as adding noise to the pictures.

Manger et al., [12] addressed the problem of image retrieval to find images in large datasets which contain certain similar scenes or the objects that are given towards a certain questioned picture. Oftenly it is done using a familiar BoW technique that is quantized with certain local features like the SIFT for increasing the speed of the retrieval by using a scheme of indexing local features. The authors have focused on the limit of large datasets as the quantization of the descriptors of the individual features and their power of discrimination. As this context information that is quantized information has introduced another dimension in this BoW-Model, it can support the performance and its accuracy during retrieving process.

Montoliu et al., [13] made a presentation of a Bag of Words for testing an indoor positioning method that is based on a magnetic field. Solution to this problem is done as a problem of pattern identification in which every reference point is of a different class. The feature vectors are constructed by a bag-of-words that is simplified to allow user speed invariance. Many classifiers are used for getting promising results.

Fidalgo-Fernandes et al., [14] had addressed the need for using knowledge on the quality perceived by human and adding models of machine learning for the quality estimation objective. Another new technique has been proposed on the basis of splitting pictures to various units in which the average of the SSIM metric has been calculated. Another sliding window on a grid of cells which will split the picture and will defined an image descriptor group which have been aggregated with the help of bag-of-words. This model can enhance all particular values that are given using SSIM and has defined a recent way to apply the machine learning in evaluating the image quality.

Wu et al., [15] made a proposal for a stratification based forecasting method that uses wind power which was developed as a hybrid forecasting model using various stratifications with a charged system for search. This proposed model used the segmentation concept for the optimal stratification method for forecasting the short term outputs of wind power. In addition to this this method that has been proposed has elucidated different values of weighting for every individual model for different blocks of segmentation. On the basis of the results of forecasting this stratification based hybrid model was proposed to outperform the stand-alone and un-stratified models which are used for forecasting accuracy and that which verified the forecasting model for the wind power forecasting.

Kanagaraj et al., [16] further presented another Charged System Search (CSS) to solve the robotic drill and its path optimization for the Printed Circuit Board (PCB) in the manufacturing industries. With the increase in the number of holes the complexity also increases. The recently developed algorithm of CSS which is proposed for solving this problem that has minimal time of computation. The working of this CSS protocol has been examined using four case studies from literature. The experience of computation here is shown as the presented protocol that could find an optimum path for Printed Circuit Board and its hole making processes within a particular time in computing.

Aryan and Alizadeh [17] made a proposal for another protocol that depend on the Charged System Search. There is a manipulation of the searching scheme of this Charged System Search for containing an additional part that is dedicated in searching the promised areas based on recent best information of the traversed land area. The author further included a concept to reduce the influencing particles through application of a choice method. Research depicted that ideas put-forth could increase the exploitation rate of the Charged System Search and resulted in better accuracy.

Ergin and Isik [18] further assessed three methods of feature selection which were the Information Gain (IG), the Gini Index (GI), and the CHI square (CHI2) with the help of two classifiers the Artificial Neural Network (ANN) and the Decision Tree (DT), for classifying Turkish e-mails. These feature vectors have been framed using bag-of-words technique. This work concentrated on the Turkish language as it serves to be a much used agglutinative language in world. The outcomes revealed that the CHI2 and the GI techniques serve to be much effective compared to the IG technique in Turkish.

Khan et al., [19] presented another Semantics Based feature vector with part of speech (POS) made use for extracting the WordNet of the associated terms. This method is implemented in small documents which proved that it out-performed them in terms of frequency/ inverse document frequency (TF-IDF) in BOW method for classification. This sections shows the extraction of feature with Stemming/stop words/Term frequency or feature selection with the MRMR, the GA, the Charged System Search and the cosine are explained. These that are obtained from https://www.careopinion.org.uk/, (474 - positive, 236 - neutral and 382 - negative).

Methodology

In this section, the feature extraction using Stemming/stop words/Term frequency, feature selection using MRMR, GA, Charged System Search and cosine similarity are explained.

Dataset: The opinions were obtained from https://www.careopinion.org.uk/, (474 positive, 236 neutral and 382 negative).

Feature Extraction: The Feature extraction is one form of dimensionality reduction. The features of extraction from this document through weight evaluation in various domains. This feature extraction is at its preprocessing stage of the knowledge discovery. This step aims at the conversion of free text review sentences in the form of words and enrich their semantic meaning. There are three subtasks that are included here which are part of speech and its tagging, stemming and its meaningful word selection [20]. There are methods of statistical feature selection at document classification used for sentiment analysis requires a conversion of text into feature. These features are the sentiment holding terms that show high frequency of occurrence in data.

Stemming: Stemming is that approach that brings down the inflected words to their stem or their base or root. There are many words English that get reduced to their base like agreed, agreeing, agreement, disagreement and agreement all belonging to agree [21]. This is that process that is used for reducing all the derived words. The Stemming program is known as stemming algorithm or stemmer. This serves to be the scheme to determine the root-words or describing the relevant tokens inside one type. For instance, "He teach us in an interesting manner" This particular sentence once stemming will change to "teach interest manner" and with the help of stem (root) word comparing this sentence word with quantity of positive or negative words is done at ease. This porter stemming algorithm can remove the common morphological and also the flexional endings from the words.

The main use of this is a part of the process of normalization which is normally done at the time of setting up an Information Retrieval System [22]. There are several words from a sentence which have an opinion information which appear in bulged manner. Stemming is applied to those bulged words even prior in searching the appropriate lists. As there is a dearth for good Bengali stemmer, the cluster technique in stemming based on the Bengali stemmer is developed. This stemmer analyzes and also pre-fixes and suf-fixes every word pattern for a specific document. The words which are found will have to possess a similar root

manner that is combined into a definite cluster numbers that are identified having the rood at cluster center and the Porter Stemmer9 is used [29].

Stop words: Removing the stop words forms a normal step in pre-processing of step. The stop words like from, of, in and so on are much frequently repeated within a data set but doesn't give any specific knowledge in data analysing. Once such words are removed, much significant words are focused that also aid to reduce the dimension. To follow the document d for it to be feasible in set F(x) of stop-words-removed count of BOW x, every word v in the vocabulary has to come xv times in d, until v is a stop-word. In case v is a stop-word, it can occur many times and the document length n is not known.

This renders the search of A^* difficult. For example, any short document using only words in x can have a higher model of language and a score than that of a longer document using stop words. Also, a long document which keeps repeating a high level of probability and a phrase of stop word like and, in, the etc., can have a higher per word score than that of a document of moderate length using lesser stop-words. A per word score can also increase monotonically as per partial document length. For simplifying this and reducing the size of the set that is feasible, the document length form x and the feasible set to the documents are used. For constructing an estimator for a document length of n, given stop words-removed count BOW, the calculation of the average ratio of the document length (with stop-words) to the BOW length (without stop words) $\beta=\frac{1}{|D|}\sum_{d_i\in D}\frac{n_i}{1^{\bullet}x_i}$ on separate training set called D of documents, in which every document di will have a length n_i and stop words-removed BOW x_i, along with1 which is the all-one vector. The estimator of document-length, given BOW x with an unknown n and d, shall scale the BOW length by :

$$\beta:\hat{n}(x)=\beta 1^{\bullet}x$$

To identify the best document $d^*\in F(x)$, perform A^* search as earlier with the difference where successor states generate with appending words from remaining BOW with no replacement but from stop words list with replacement [23].

Term frequency: The Supervised machine learning (ML) and their algorithms need an appropriate representation to be a Features Vector. Many of the Machine Learning techniques makes use of the vector space model (VSM) where every document is shown like a vector of Weighted Features. There are various models of text representation that have been created for the tweets that depend on two characters of the BoW and the semantic ideas. The features are weighed by with the help of a Term Frequency Inverse Document Frequency (TF-IDF) weighting technique. It can help in bringing down the Features Weight which occurs in many data-set documents. This has been explained like [24]:

$$\text{TF-IDF}(f_n,d_i)=\text{TF}(f_n,d_i).\text{IDF}(f_n)$$

In which $\text{TF}(f_n,d_i)$ will be the frequency of feature f_n, $\text{IDF}(f_n)$ and is:

$$\text{IDF}(f_n)=\log\frac{|D|}{\text{DF}(f_n)}$$

In which $\text{DF}(f_n)$ indicates the number of documents within D which includes feature f_n. The $|D|$ denotes the number of documents within the data-set.

Feature selection: Many of the measurable elements and their selection schemes for an archive level grouping are used to analyze the conclusion. The factual methodology that is least difficult for highlighting determination is utilizing the commonly happening words in various corpuses as the extremity markers. Most methodology include a process having two stages [25].

- Identifying the record parts that contribute to positive and negative slants.
- Joining of these parts such a way that can frame chances of archive tending that is group to be one among the two polar classifications.

In the task of text classification, most techniques make use of Bag of Word features for representing documents that can lead to a large sized document vector or sentence vector in the feature space. The different methods of feature-selection have been used for selecting the useful features for reducing the feature space size and improve its efficiencies. The methods of feature selection are those techniques which choose a small feature set out of a particular set of features for capturing its relevant properties of the dataset classification. The Feature selection is viewed as a weighting form where some terms can get a zero weight and so it may be removed from the feature space. The idea 32 of the selection is giving more surprising features and lower weight for the expected features.

The hypothesis used here is that it has surprising features and if they are shared by two different vectors they are more discriminative of the similarities between the vector than the features that are less surprising. On the basis of the information theory another surprising feature that has a higher content of information than that of the expected feature. In the text domain another effective feature selection method has been essential for making the learning tasks that are accurate and efficient. In feature spaces of machine learning techniques text classification of the learning models

the size of the document vectors or the sentence vectors using Bag of word is big as it depends on the vocabulary size in the corpus [26].

Minimum Redundancy and Maximum Relevance (mRMR): This technique known as Minimum Redundancy and Maximum Relevance is a feature choice routine which is filter-based made use to select the noticeable characters in a class. The feature selection serves to be a major issue in Machine Learning. It duly determines the features' sub-sets which are much co-related as well as sufficient have strength for identifying the class which is termed as Maximum Relevance. The sub-sets of these characters normally have certain related characters which includes certain hidden characters. This mRMR technique of feature choosing makes an attempt in eliminating the redundant features known as Minimum Redundancy. If two of these related characters are redundant together the feature that is less important is dropped with no compromise to its classifier's performance. The mRMR scheme chooses prominent features as below:

- It chooses features correlated with the attribute of the class (maximum relevance).
- The features are chosen in a manner such that it remains less redundant and also much co-related with its class attributes (Minimum Redundancy)

The technique of mRMR feature selection makes use of Mutual Information for measuring co-relation among features and class parameters. Mutual Information takes a measurement of non-linear co-relation among two parameters. Considering a feature set $F = \{f_1, f_2, ..., f_n\}$ that contains n total features, with a class attribute C. Either the co-relation, or the relevance, of feature f_m having a class parameter C is shown with Mutual Information, i.e., joint probability distribution $P(f_m, C)$ and marginal probability distribution $P(f_m), P(C)$:

$$A = MI(f_m, C) = \sum_{m,C} P(f_m, C) \log \frac{P(f_m, C)}{P(f_m) P(C)}$$

The advantage of mRMR a Feature Selection technique serves to be the redundant characters are eliminated and it chooses only the relevant features which is not like other techniques like Information Gain (IG) and Mutual Information (MI). other techniques are used for sentiment analysis and focuses on choosing relevant feature without taking redundant information [27].

Genetic Algorithm (GA):

1. The GA is the method of stochastic general search which is capable of exploring effectively the large spaces of search that usually is needed in the cases of attribute selection. Also, as in the case of most search algorithm performing local and greedy search, this performs a global search. The gas will simulate the process in the natural evolution systems that are based on the "survival of the fittest' principle that was given by Charles Darwin. This genetic algorithm is composed of three operators:

2. reproduction
3. crossover
4. mutation

5. The reproduction chooses a good string (which is a subset of input attributes); with a crossover combination of good strings to generate better offspring's; the mutation alters the strings locally for attempting to create better strings. This string contains binary bits: 1 that represents selection of attributes or 0 to drop the same attribute. In every generation, there is an evaluation of population which is also tested for the termination. If the criterion for termination is not satisfied the population is duly operated based on GA operators and later reevaluated. This is repeated for a certain number of generations [28]. The GAs have been characterized by five basic components. A diagrammatic representation of the entire process [29].

6. Representation of Chromosomes for feasible solutions to the problems of optimization.
7. The Initial population of feasible solutions.
8. A function of fitness to evaluate the solutions.
9. The Genetic operators which generate new populations from among the present population.
10. The Control parameters like population size, number of generations and probability of genetic operators.

Charged System Search: Another effective optimizing protocol has been setup using the mentioned physics rules, termed as Charged System Search (CSS). In this, every solution candidate **X***i* that contains some decision variables (i.e. **X***i* = {*xi, j*}*)* which is assumed to be a charged particle. The charged particle has been effected due to various electrical fields of agents. The resultant force has a quantity which is dependent on the usage of electrostatics of law that is discussed and the movement's quality is found using Newtonian Mechanics laws. This means that the agent that has better outcomes that give a strong force compared to bad ones and the quantity of charge is explained taking the objective-function fit(i) into consideration. In order that Charged System Search is introduced, the rules below are formulated

Rule 1: Most natural evolution protocols have a solution population that serves to be emerged by selecting and altering randomly. Likewise, the CSS considers Charged Particles (CP). Each of this has a magnitude of charge *(qi)* that forms in its space an electric field. The magnitude of charge is explained based on the solution and its quality is as below:

$$q_i = \frac{fit(i) - fitword}{fitbest - fitworst}, \qquad i = 1, 2, ..., N$$

Where, the fitbest along with the fitworst serves to be the best as well as the worst fitnesses of all of the particles; fit*(i)*

indicates the objective-function and its value or the agent *i* and its fitness and *N* denotes the entire Charged Particles.

Rule 2: The first locations of the Charged Particles have been determined in random in-between the search space span.

$$x_{i,j}^{(o)} = x_{i,\min} + rand.\left(x_{i,\max} - x_{i,\min}\right), \qquad i = 1,2,...,n,$$

In which $x_{i,j}^{(o)}$ shows its initial value and of the *i* th parameter for *j* th Charged Particle; $x_{i,\min}$ and $x_{i,\max}$ which are the minimum as well as the maximum allowed range for *i* th variable; rand the random number in range [0,1]; and *n* the number of variables. The first velocities of the Charged Particles remain at zero

Rule 3: There are three conditions that are related to the attractive forces:

- Any of the CPs can affect the another; i.e., a bad Charged Particle is capable of affecting a good one and also vice versa *(pi j=1)*.
- A Charged Particle is capable of attracting another in case its electric charge amount (the fitness with a revise relation in bringing down problems) is found to be good compared to others which means a good Charged Particle can attract a bad one:
-

$$p_{ij} = \begin{cases} 1 & \text{fit}(j) > \text{fit}(i) \\ 0 & \text{else.} \end{cases}$$

- All the good CPs are capable of attracting the bad Charged Particles and only certain bad agents can attract good agents, keeping the probability functions below:

$$p_{ij} = \begin{cases} 1 & \dfrac{\text{fit}(i)\text{-fitbest}}{\text{fit}(j)\text{-fit}(i)} > rand \vee fit(j) > fit(i), \\ 0 & \text{else.} \end{cases}$$

In accordance with the rules mentioned above if a good agent attract a bad agent the capability of exploitation of this protocol is given and vice-versa in case a bad Charged Particle attract a good one. If a Charged Particle moves towards an agent that is good and the performance is improved and the principle of self adaptation is assured. By transferring a good Charged Particle towards that of a bad CP, previously obtained better solutions could be loast because it increases the cost of computation of finding a better solution. To solve this issue, the memory that holds the best of the so-far obtained is used [30].

Cosine Similarity: For this model, the text represents a vector of a set of real numbers, in which every component will correspond to a word which appear in the documents and the value serves to be the frequency in this document. It is called the bag-of-words expression. The likeness between two documents x and y will be the cos of angles between the two vectors $\vec{x}$ and $\vec{y}$ that represent x and y correspondingly, and is computed with the help of formulae below [31]:

$$similarity(x,y) = \frac{\vec{x}.\vec{y}}{|\vec{x}| \times |\vec{y}|}$$

In which $|\vec{x}|$ and $|\vec{y}|$ denote the norms of every document vector. The cosine distance has been explained as one minus cos of angles considered between vectors.

$$cosine\ distance(x,y) = 1 - cosine\ similarity(x,y)$$

1. The Cosine similarity will be a measurement of likeness that exists within the two vectors of the inner product space which will measure the cos of angles within these vectors which is not more than 90°.
2. During retrieving information and the text mining, every term is assigned various dimension along with a document which is defined by the vector in which every dimension's value will correspond to the actual frequency rate of appearance for this document.
3. This gives a measure of in which way two similar documents are in terms of its subject matter. The cosine which is part of two vectors is arrived at by the Euclidean dot product formula:

$$a.b = \|a\|\|b\|\cos\theta$$

4. For retrieval of information similarity of cosine of documents vary between 0 to 1 as term frequencies (TF-IDF weights) are not negative [32].

Results and Discussion

The experimental outcomes are compared with one another. Table 1 depicts the outcomes of the accuracy, positive predictive value and Hitrate using MRMR, GA and Charged System Search.

Table 1 and figure 1 depicts that the presented technique enhanced the accuracy in classifying by 7.4% when compared with MRMR and GA. Similarly, the proposed method improved classification accuracy by 2.4% when compared with GA and Charged System Search.

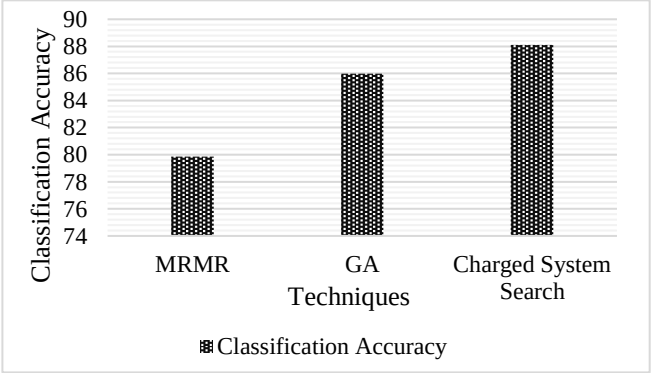

Figure 1: Classification Accuracy

Table 1
Summary of results

Techniques	MRMR		GA	Charged System Search
Classification Accuracy	79.85		85.99	88.1
Positive predictive value for Positive Opinion	0.8462		0.8911	0.9099
Positive predictive value for Neutral Opinion	0.7472		0.8333	0.8532
Positive predictive value for Negative Opinion	0.7792		0.841	0.8649
Hitrate for Positive Opinion	0.789		0.846	0.8734
Hitrate for Neutral Opinion	0.839		0.8898	0.911
Hitrate for Negative opinion	0.7853		0.8586	0.8717

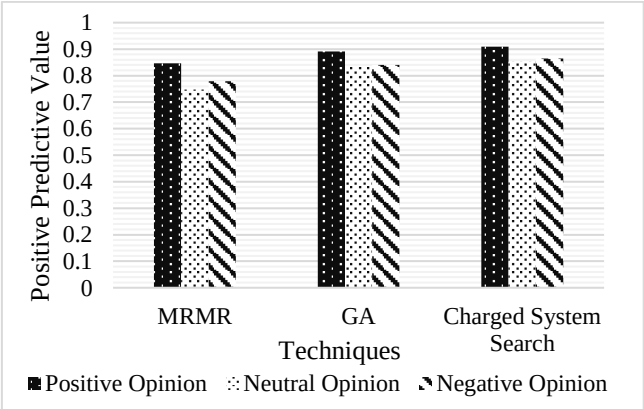

Figure 2: Positive Predictive Value

Table 1 and figure 2 shows that the proposed method improved positive predictive value by 5.17% when compared with MRMR and GA using positive opinion. Similarly, the proposed method improved positive predictive value by 2.8% when compared with GA and Charged System Search using negative opinion.

Table 1 and figure 3 shows that the proposed method improved hitrate by 6.97% when compared with MRMR and GA using positive opinion. Similarly, the proposed method improved hitrate by 2.35% when compared with GA and Charged System Search using neutral opinion.

Conclusion

The Sentiment analysis is that process which identifies customer sentiments and their emotional states. The customer feelings are expressed as either positive, negative or neutral. Most of the uses of opinion mining do not depend on bag of words that don't consider the context where Sentiment analysis is needed. The CSS consists of a number of agents that are charged particles. In most cases, the parts of speech are used as a feature for extracting the text and its sentiment for the text and it uses BoW as a dataset feature. The results of the experiments show that this method that is proposed has an improved accuracy of classification by about 7.4% which it is compared to the MRMR and the GA. Likewise, this proposed method has an improved accuracy of classification by 2.4% in comparison with the Charged System Search and the GA.

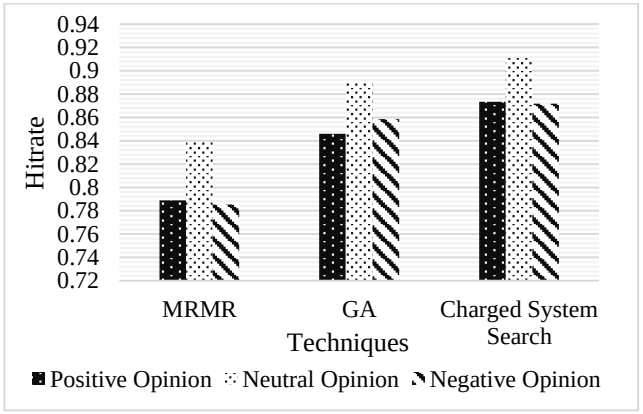

Figure 3: Hitrate

References

1. Chandrakala, S., & Sindhu, C. (2012). Opinion mining and sentiment classification: A survey. *ICTACT journal on soft computing*, *3*(1), 420-425.

2. Sayeed, A. B. Grammar-based approaches to opinion mining. 1-6.

3. El-Din, D. M. Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis. *International Journal of Advanced Computer Science & Applications*, *1*(7), 244-252.

4. Poornima, R., & BRupa, D. E. V. I. (2016). Sentiment Analysis for Two Sides of Review Using Dual Prediction. *International Journal of Technology and Engineering System (IJTES)*, 8(1), 65-69.

5. Mahendran, A., Duraiswamy, A., Reddy, A., & Gonsalves, C. (2013). Opinion Mining for text classification. *International Journal of Scientific Engineering and Technology*, *2*(6), 589-594.

6. Stalidis, P., Giatsoglou, M., Diamantaras, K., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). Machine Learning Sentiment Prediction based on Hybrid Document Representation. *arXiv preprint arXiv:1511.09107*.

7. Janardhana, D. R., & Mulimani, M. (2015). Sentiment Analysis and Opinion Mining using Machine Learning Techniques.

International Journal of Innovative Research in Computer and Communication Engineering, 3(10), 9321-9329.

8. Shirbhate, A. G., & Deshmukh, S. N. (2016). Feature Extraction for Sentiment Classification on Twitter Data. *International Journal of Science and Research (IJSR)*, 5(2), 2183-2189.

9. Suganya, B., & Priya, V. (2017). Particle Swarm Optimization Based Feature Selection and Summarization of Customer Reviews. *International Conference on Emerging trends in Engineering, Science and Sustainable Technology (ICETSST)*, 131-135.

10. Tu, J., Sui, H., Feng, W., Sun, K., & Hua, L. (2016). Detection of Damaged Rooftop Areas From High-Resolution Aerial Images Based on Visual Bag-of-Words Model. *IEEE Geoscience and Remote Sensing Letters*, *13*(12), 1817-1821.

11. Zhu, X., Ma, B., Guo, G., & Liu, G. (2016, August). Aircraft type classification based on an optimized Bag of Words Model. In *Guidance, Navigation and Control Conference (CGNCC), 2016 IEEE Chinese* (pp. 434-437). IEEE.

12. Manger, D., Herrmann, C., & Willersinn, D. (2016, November). Towards Extending Bag-of-Words-Models Using Context Features for an 2D Inverted Index. In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on* (pp. 1-5). IEEE.

13. Montoliu, R., Torres-Sospedra, J., & Belmonte, O. (2016, October). Magnetic field based Indoor positioning using the Bag of Words paradigm. In *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on* (pp. 1-7). IEEE.

14. Fidalgo-Fernandes, M., Bernardo, M. V., & Pinheiro, A. M. (2016, June). A bag of words description scheme based on SSIM for image quality assessment. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on* (pp. 1-6). IEEE.

15. Wu, Y. K., Su, P. E., & Hong, J. S. (2015, October). Stratification-based wind power forecasting in a high penetration wind power system using a hybrid model with charged system search algorithm. In *Industry Applications Society Annual Meeting, 2015 IEEE* (pp. 1-9). IEEE.

16. Kanagaraj, G., Ponnambalam, S. G., & Loo, C. K. (2015, August). Charged system search algorithm for robotic drill path optimization. In *Advanced Mechatronic Systems (ICAMechS), 2015 International Conference on* (pp. 125-130). IEEE.

17. Aryan, M., & Alizadeh, B. A. M. (2016, July). Bayesian charged system search: A hybrid method for multi-modal optimization problems. In *Evolutionary Computation (CEC), 2016 IEEE Congress on* (pp. 1501-1508). IEEE.

18. Ergin, S., & Isik, S. (2014, June). The assessment of feature selection methods on agglutinative language for spam email detection: A special case for Turkish. In *Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on* (pp. 122-125). IEEE.

19. Khan, A., Baharudin, B., & Khan, K. (2010, June). Semantic based features selection and weighting method for text classification. In *Information Technology (ITSim), 2010 International Symposium in* (Vol. 2, pp. 850-855). IEEE.

20. Vidhya, S., Singh, D. A. A. G., & Leavline, E. J. (2015). Feature Extraction for Document Classification. *International Journal of Innovative Research in Science, Engineering and Technology*, 4(6), 50-56.

21. Sekar, J. (2015). Social emotion prediction using opinion mining for various patterns. *Journal of Global Research in Computer Science*, *6*(1), 09-14.

22. Dhurve, R., & Seth, M. (2015). Weighted Sentiment Analysis Using Artificial Bee Colony Algorithm. *International Journal of Science and Research (IJSR)*, 4(8), 1717-1722.

23. Fillmore, N., Goldberg, A. B., & Zhu, X. (2008). *Document recovery from bag-of-word indices*. Technical report, University of Wisconsin–Madison.

24. Alowaidi, S., Saleh, M., & Abulnaja, O. (2017). Semantic Sentiment Analysis of Arabic Texts. *International Journal of Advanced Computer Science and Applications*, *8*(2), 256-262.

25. Tiwari, A., & Shrivas, R., & Dwivedi, N. (2015). Sentiment Analysis on Textual Reviews with Feature Reduction Using PCA Algorithm. *International Journal of Emerging Research in Management &Technology*, 4(12), 1-5.

26. Meng, Y. (2012). Sentiment analysis: A study on product features.

27. Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation*, *7*(4), 487-499.

28. Khare, P., & Burse, K. (2016). Feature Selection Using Genetic Algorithm and Classification using Weka for Ovarian Cancer. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 7 (1), 194-196.

29. Das, A., & Bandyopadhyay, S. (2010, August). Subjectivity detection using genetic algorithm. In *the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA10), Lisbon, Portugal*.

30. Kaveh, A., & Talatahari, S. (2010). A novel heuristic optimization method: charged system search. *Acta Mechanica*, *213*(3), 267-289.

31. Nahm, U. Y. (2004).Text mining with information extraction. The University of Texas at Austin.

32. Deshpande, S., Rathi, J., Gandhi, S., Shinde, M., & Deshmukh, V. (2015). Sentiment analysis tool using cosine and jaccard implementation. *International Journal of Computer Applications*, *115*(12)