Subjectivity Detection using Genetic Algorithm

Amitava Das[1] and **Sivaji Bandyopadhyay**[2]

Abstract. An opinion classification system on the notion of opinion subjectivity has been reported. The subjectivity classification system uses Genetic-Based Machine Learning (GBML) technique that considers subjectivity as a semantic problem using syntactic simple string co-occurrence rules that involves grammatical construction and linguistic features. Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. This is viewed as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. The proposed technique is tested for English and Bengali and for the news, movie review and blog domains. The system evaluation results show precision of 90.22%, and 93.00% respectively for English NEWS and Movie Review corpus and 87.65% and 90.6% for Bengali NEWS and Blog corpus.

1 Introduction

As a growing number of people use the Web as a medium for expressing their opinions, the Web is becoming a rich source of various opinions in the form of product reviews, travel advice, social issue discussions, consumer complaints, movie review, stock market predictions, real estate market predictions, etc. Present computational systems need to extend the power of understanding the sentiment/opinion expressed in an electronic text. The topic-document model of information retrieval has been studied for a long time and several systems are available publicly since last decade. On the contrary Opinion Mining/Sentiment Analysis is still an unsolved research problem. Although a few system like Bing[3] , Twitter Sentiment Analysis[4] Tool are available in World Wide Web since last few years still more research efforts are needed to match the user satisfaction level and social need.

The General Inquirer System by[1] IBM in the year of 1966 was probably the first milestone to identify textual sentiment. They called it a content analysis research problem in the behavioral science. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written and counting positive or negative emotion instances. Later on an opinion was defined as a private state that is not open to objective observation or verification [2]. During 1970-1995 various research activities ([3], [4]) proves the necessity of an automated system that can identify sentiment in electronic text.

In the year of 1999 Jaynce Wiebe [5] defined the term Subjectivity in Information Retrieval perspective. Sentences are categorized in two genres as Subjective and Objective. Objective Sentences are used to objectively present factual information and subjective sentences are used to present opinions and evaluations.

Researchers have experimented with several methods to solve the problem of subjectivity detection using SentiWordNet, Subjectivity Word List etc. as prior knowledge database. But Subjectivity Detection is a domain dependent and context dependent problem [6]. Hence building a prior knowledgebase for Subjectivity Detection will never end up with an adequate list. Moreover Sentiment/opinion changes its polarity orientation during time. For example, during 90's mobile phone users generally report in various online reviews about their color phones but in recent time color phone is not just enough. People are excited about their touch screen or various software installation facilities. Hence Subjectivity detection needs a most sophisticated algorithm to capture and effectively use the sentiment pragmatic knowledge. The algorithm should be customizable for any new domain and language.

Previous works in subjectivity identification have helped developing a large collection of subjectivity clues. These clues include words and phrases collected from manually developed annotated resources.

The clues from manually developed resources include entries from adjectives manually annotated for polarity [7], and subjectivity clues listed in [8]. Clues learned from annotated data include distributionally similar adjectives and verbs [9] and n-grams [10]. Low-frequency words are also used as clues. Such words are informative for subjectivity recognition.

The subjectivity detection task in Bengali has started only recently. Several syntactic and semantic feature ensembles with a rule base topic-base model is reported in [11], [12].

Genetic Algorithms (GAs) are probabilistic search methods ([13], [14]). GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions. Nowadays GAs have been applied to various domains that include timetable, scheduling, robot control, signature verification, image processing, packing, routing, pipeline control systems, machine learning, and information retrieval ([15], [16]).

Only a few attempt [17] in the literature uses Genetic Algorithm to solve the opinion mining problem. They developed the Entropy Weighted Genetic Algorithm (EWGA) for opinion feature selection. The features and techniques result in the creation of a sentiment analysis approach geared towards classification of web discourse sentiments in multiple languages. The EWGA has been applied for English and Arabic languages. The Entropy Weighted Ge-

[1] A. Das .Department of Computer Science and Engineering, Jadavpur University. Kolkata 700032, West Bengal, India. email: amitava.santu@gmail.com

[2] S. Bandyopadhyay.Department of Computer Science and Engineering, Jadavpur University. Kolkata 700032, West Bengal, India. email: sivaji_cse_ju@yahoo.com

[3] http://www.bing.com/

[4] http://twittersentiment.appspot.com/

netic Algorithm (EWGA) uses the information gain (IG) heuristic to weight the various opinion attributes.They compared their result with SVM based method and previous existing methods in literature. The EGWA method outperform compared to existing methods and achieved approximately 94.00% accuracy score on both the languages English and Arabic.

Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. Here we viewed this as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. In the present paper we use GBML to identify automatically best feature set based on the principle of *natural selection* and *survival of the fittest*. The identified fittest feature set is then optimized locally and global optimization is then obtained by multi-objective optimization technique. The local optimization identify the best range of feature values of a particular feature. the Global optimization technique identifies the best ranges of values of given multiple feature. The proposed technique is tested for English and Bengali and for the news, movie review and blog domains. The system evaluation results show precision of 90.22%, and 93.00% respectively for English NEWS and Movie Review corpus and 87.65% and 90.6% for Bengali NEWS and Blog corpus.

2 Resource Organization

Resource acquisition is one of the most challenging obstacles to work with resource constrained language Bengali, the fifth popular language in the World, second in India and the national language in Bangladesh. NLP research in Bengali has kicked off in recent times and resources like annotated corpus, various linguistic tools are still unavailable for this language. Hence in this section we mainly describe the corpus acquisition and development of tools used in feature extraction for Bengali. English is a resource rich language, the resources for English are collected from various publicly available resources, mentioned in detail in relevance sections.

2.1 Corpus

The subjectivity classification technique presented in this paper is based on Genetic-Based-Machine-Learning (GBML) methodology and hence annotated data preparation is necessary for system testing and evaluation. The technique has been applied on both English and Bengali language texts. In case of English, the MPQA[5] corpus is chosen which is well known for its high inter-annotator agreement score. In the MPQA corpus the phrase level private states are annotated that has been used in the sentence level opinion subjectivity annotation as described in [11]. Manually annotated Subjective data is available for English in the form of International Movie Database (IMDB)[6] among others.

For the present task we have used a Bengali NEWS corpus, developed from the archive of a leading Bengali NEWS paper available on the Web. A portion of the corpus from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section containing 28K wordforms have been manually annotated with sentence level subjectivity. Detailed reports about this news corpus development in Bengali can be found in [11] and a brief statistics is reported in Table 1.

Table 1. Bengali Corpus Statistics

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Avgerage number of sentences in a document	22	-
Total number of wordforms	28807	4675
Avgerage number of wordforms in a document	288	-
Total number of distinct wordforms	17176	1235

2.2 Feature Organization

The experimentation started with the complete collection of identified lexicon, syntactic, semantic and discourse level features. The best feature set selection has been carried out by the GBML technique. Various features and the linguistics tools used for features extraction are reported below. The GBML trained with all the features are summarized in Table 3.

2.2.1 Lexico-Semantic Features

- Part of Speech (POS)

Number of research activities like [18], [19] etc. have proved that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like the one presented in [20], are mostly based on adjective words. The Stanford Parser[7] has been used for identifying the POS tags in English. The POS tagger described in [11] has been used for Bengali.

- SentiWordNet

Words that are present in the SentiWordNet carry opinion information. The English SentiWordNet ([21]) has been used in the present task. The SentiWordNet (Bengali)[8] as described in [22] is used as an important feature during the learning process. These features are individual sentiment words or word n-grams (multiword entities) with strength measure as strong subjective or weak subjective. Strong and weak subjective measures are treated as a binary feature in the supervised classifier. Words which are collected directly from the SentiWordNet are tagged with positivity or negativity score. The subjectivity score of these words are calculated as:

$$E_s = |S_p| + |S_n|$$

where E_s is the resultant subjective measure and S_p, S_n are the positivity and negativity score respectively.

- Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document. After removal of function words and POS annotation, the system generates four separate high frequent word lists for the four POS categories: Adjective, Adverb, Verb and Noun. Word frequency values are effectively used as a crucial feature in the Subjectivity classifier.

[5] http://www.cs.pitt.edu/mpqa/databaserelease/

[6] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[7] http://nlp.Stanford.edu/software/lex-parser.shtml

[8] http://www.amitavadas.com/sentiwordnet.php

- Stemming

Several words in a sentence that carry opinion information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in the appropriate lists. Due to non availability of a good Bengali stemmer, a stemming cluster technique based Bengali stemmer [23] has been developed. The stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center. The Porter Stemmer[9] has been used for English.

2.2.2 *Syntactic Features*

- Chunk Label

Chunk level information is effectively used as a feature in the supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. A detailed empirical study [11] reveals that Subjectivity clue may be defined in terms of chunk tags. The Stanford Parser has been used for identifying the chunk labels in English. The Bengali chunker used in the present task is described in [11].

- Dependency Parser

Dependency feature is very useful to identify intra-chunk polarity relationship. It is very often a language phenomenon that modifiers or negation words are generally placed at a distance with evaluative polarity phrases. The Stanford Dependency Parser has been for English. A statistical parser [24] has been used for Bengali.

2.2.3 *Discourse Level Features*

- Positional Aspect

Depending upon the position of subjectivity clue, every document is divided into a number of zones. Various factors of this feature are Title of the document, the first paragraph and the last two sentences. A detailed study was done on the MPQA and Bengali corpus to identify the roles of the positional aspect (first paragraph, last two sentences) in the sentence level subjectivity detection task and these results are shown in the Table 2. Zone wise statistics could not be done for the IMDB corpus because the corpus is not presented as a document.

- Document Title

It has been observed that the Title of a document always carries some meaningful subjective information. Thus a Thematic expression bearing title words (words that are present in the title of the document) always get higher score as well as the sentences that contain those words.

- First Paragraph

People usually give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate or support their ideas with relevant reasoning or factual information. This first paragraph information is useful in the detection of subjective sentences bearing Thematic Expressions.

- Last Two Sentences

Table 2. Statistics on Positional Aspect.

Positional Factors	Percentage	
	MPQA	Bengali
First Paragraph	48.00%	56.80%
Last Two Sentences	64.00%	78.00%

It is a general practice of writing style that every document concludes with a summary of the opinions expressed in the document.

- Term Distribution Model

An alternative to the classical TF-IDF weighting mechanism of standard IR has been proposed as a model for the distribution of a word. The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. As discussed in [25] introduced the opinion distribution function feature to capture the overall opinion distributed in the corpus. Thus the objective is to estimate that measures the distribution pattern of the *k* occurrences of the word wi in a document *d*. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of topic-sentiment informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows: where *n*=number of sentences in a document with a particular theme word S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, is the positional id of current Theme word and is the positional id of the previous Theme word.

$$f_d\left(w_i\right) = \sum_{i=1}^{n} \frac{\left(S_i - S_{i-1}\right)}{n} + \sum_{i=1}^{n} \frac{\left(TW_i - TW_{i-1}\right)}{n}$$

where *n*=number of sentences in a document with a particular theme word S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

Distribution function for thematic words plays a crucial role during the Thematic Expression identification stage. The distance between any two occurrences of a thematic word measures its distribution value. Thematic words that are well distributed throughout the document are important thematic words. In the learning phase experiments are carried out using the MPQA Subjectivity word list distribution in the corpus and encouraging results are observed to identify the theme of a document. These distribution rules are identified after analyzing the English corpora and the same rules are applied to Bengali.

- Theme Words

[9] http://tartarus.org/~martin/PorterStemmer/

In the general practice of Information Retrieval term frequency plays a crucial role to identify document relevance. In many documents relevant words may not occur frequently and on the other hand irrelevant words may occur frequently. A rulebased Theme detection technique has been proposed in [9]. The theme of a document is described as a bag-of-words that describe the topic of the document.

In the general practice of Information Retrieval term frequency plays a crucial role to identify document relevance. In many documents relevant words may not occur frequently and on the other hand irrelevant words may occur frequently. A rule-based Theme detection technique has been proposed in [9]. The theme of a document is described as a bag-of-words that describe the topic of the document.

Table 3. Features.

Lexico-Syntactic Features
POS
SentiWordNet
Frequency
Stemming
Syntactic Features
Chunk Label
Dependency Parsing
Document Title
Discourse Level Features
First Paragraph
Term Distribution Model
Theme Word

3 Basic Principles of Genetic Algorithm

GAs are characterized by the five basic components as follows. Figure 1 displays a diagrammatic representation of the whole process.

1. Chromosome representation for the feasible solutions to the optimization problem.
2. Initial population of the feasible solutions.
3. A fitness function that evaluates each solution.
4. Genetic operators that generate a new population from the existing population.
5. Control parameters such as population size, probability of genetic operators, number of generation etc.

4 Proposed Technique

The experimentation starts with a large set of possible extractable set of syntactic, semantic and discourse level features. The fitness function calculates the accuracy of the subjectivity classifier based on the fittest feature set identified by natural selection through the process of crossover and mutation after each generation. The subjectivity classification problem can be viewed as a summation of the subjectivity probability of the set of possible features.

$$f_s = \sum_{i=0}^{N} f_i$$

Where is the resultant subjectivity function, to be calculated and is the ith feature function. If the present model is represented in a vector space model then the above function could be rewritten as:

$$f_s = \vec{f_i}.\vec{f_{i+1}}.\vec{f_{i+2}}.........\vec{f_n}$$

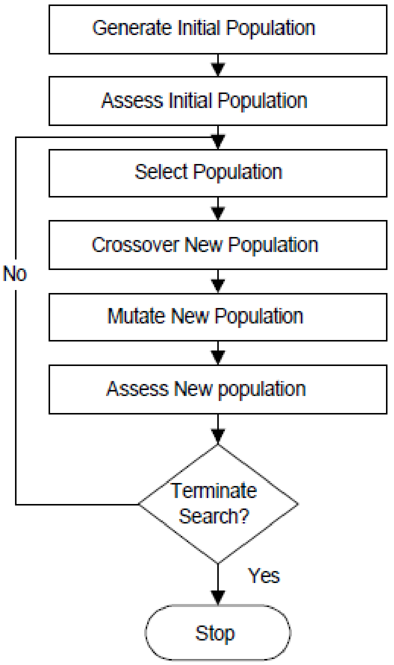

Figure 1. The Process of Genetic Algoriyhm

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a classification metric, since it is too sensitive to the absolute magnitudes of the various dimensions.

From the previous research it is already proven that particular features have their own range of tentative values (Instead of features identified by us; Syntactic Chunk Label and Discourse Level feature). As example some special types of POS category reflects sentiment very well, hence it is simpler to infer that frequent occurrence of those special types of POS category into a sentence can increase the subjectivity value of any sentence. Another example: occurrence low frequency word is a well established clue of subjectivity but a sentence with only low-frequent word may not subjective always. In a multiple feature or multiple vector spaced model desired optimal solution may found by finding out the optimal range (highest or lowest) of value of every feature vector. Hence it is obvious that in single-criterion optimization, the notion of optimality scarcely needs any explanation in this particular category of problem. We simply seek the best value of assumedly well-defined multi-objective (utility or cost) optimization function.

4.1 Problem Formulation

To maximize the subjectivity probability, the occurrence of low-frequency words (LFW), title words (TW), average distributed words (ADW) and theme words (TD) and their position in each sentence are calculated. The matrix representation for each sentence looks like: [x, y]= [frequency in the entire corpus, position in the sentence]

LFW= [5, 2]
TW= [34, 11]
ADW= [21, 10]
TD= [25, 5]

The above data are plotted as position versus frequency in the Figure 2.

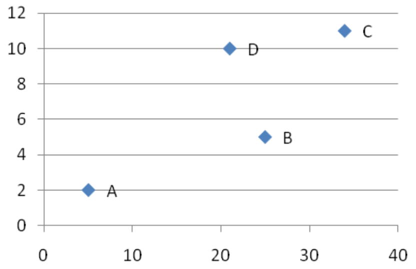

Figure 2. Position vs. Frequency plot of categorical words

Scanning the graph reveals that the best points are lower and to the right of the plot. In particular, scenarios A, B and C seem like good possible choices: even though none of the three points is best along both dimensions, we can see that there are tradeoffs from one of these three scenarios to another; there is gain along one dimension and loss along the other. In optimization terminology we say these three points are *nondominated* because there are no points better than these on all criteria.

The GBML provides the facility to search in the Pareto-optimal set of possible features. This Pareto-optimal set is being generated from crossover and mutation. To make the Pareto optimality mathematically more rigorous, we state that a feature vector x is partially less than feature vector y, symbolically x<p y, when the following condition holds:

$$(x < p\,y) \Leftrightarrow (\forall_i)\,(x_i \leq y_i) \bigwedge (\exists_i)\,(x_i < y_i)$$

This may be mapped to Pareto plane as shown in Figure 3, where Pareto front of *nondominated* points are highlighted in red color.

In the notion of Pareto optimality by multi-objective optimization we used GA in parallel fashion. The methodology used is as follows:

1. Generate chromosome for each feature.
2. Initialize population for each feature.
3. For i=1 to population size For j=1 to feature vector size Compute fitness value.
4. If termination condition satisfied go to Step 10.
5. Crossover.
6. Mutation.
7. Natural Selection.
8. Go to Step 3.
9. Output
10. End

The termination condition as mentioned in Step 4 is a pre-mature termination condition. If the fitness function wasn't improved for *n* consecutive generations then the iteration has been discarded. But there is no case of pre-mature termination case found during experimentation. Experimantally the threshold value of *n is 5.*

The parallelism is obtained here by generating n number of GA based subjectivity classifiers. Based on the principle of survival of the fittest, a few of the feature strings are selected. This parallelism provides the granularity for every feature. The GA based subjectivity classifiers are synchronous in nature. The n numbers of GA based subjectivity classifiers generate their population simultaneously. The fitness value is calculated after every iteration. The optimal solution is selected based on the theory of Pareto optimality. Pareto optimality helps to reach the fittest global solution from local best solution for each feature. The effectiveness of the present technique is observed in the experimental results.

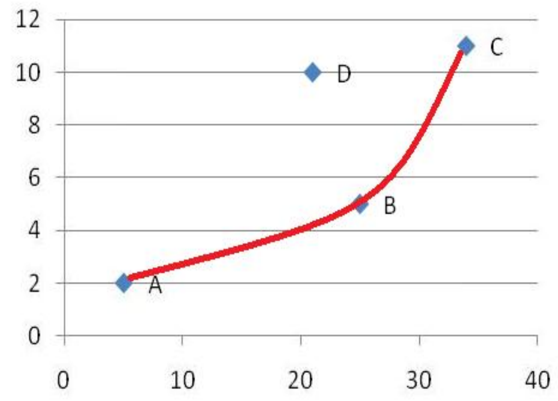

Figure 3. Pareto optimal plane

4.2 Chromosome Representation

The size of the chromosome for every feature varies according to the possible solution vector size. Tentative solutions are made of sequences of genes. Each gene corresponds to word sequence in the sentence to be tagged.

The chromosomes forming the initial population are created by randomly selecting from a dictionary one of the valid tags for each word. For the present task we have used real encoding. A sentence wise feature vector can be represented as.

Example. Imperialism/NNP is/VBZ the/DT source/NN of/IN war/NN and/CC the/DT disturber/NN of/IN peace/NN.

The encoded chromosome is represented in Table 5. The real values are the serial number of the corresponding tag from the POS Tag labeled dictionary. Table 4 reports how real values vary for every feature.

For POS feature values vary for languages as the tag set are different. There are 21 tags and 45 tags in the POS tagset for Bengali and English respectively. For sentiment words from SentiWordNet values are -1 for negative, 0 for neutral and +1 for positive words. For low frequency words features are considered as binary i.e. either a word is low-frequent or not. Any word occurring less than 5 times in the corpus has been considered as a low frequency word. This feature is encoded as a binary feature. Stems from the corpus are listed and the serial number of any stem within the list is used to encode the chro-

Table 4. Dimension of Chromosome Encoding.

Features	Real Values
POS	1-21 (Bengali) / 1-45 (English)
SentiWordNet	-1 to +1
Frequency	-1 to +1
Stemming	0 or 1
Chunk Label	1 to 17176/ 1 to 1235
Dependency Parsing	1-11 (Bengali) / 1-21 (English)
Title of the Document	1-30 (Bengali) / 1-55 (English)
First Paragraph	Varies document wise
Average Distribution	Varies document wise
Theme Word	Varies document wise

mosome. It is basically the range of unique wordforms in any corpus. Chunk label and Dependency parsing is encoded as the POS feature.

Discourse level features varies at each document level. For the three listed discourse level features three different dictionary of first paragraph word, eventually distributed words and theme words have been generated at each document level. Then index numbers from the dictionaries are used to generate the encoded chromosomes.

Table 5. Chromosome Representation.

NNP	VBZ	DT	NN	IN	NN	CC	DT	NN	IN	NN
1	12	6	2	18	2	4	6	2	18	2

4.3 Fitness Evaluation

Fitness function is a performance measure or reward function which evaluates how good each solution is. The following cost-to-fitness transformation is commonly used with GAs.

$$f\left(x\right)=C_{max}-g\left(x\right)when\,g\left(x\right)<C_{max}\quad or\;0\;Otherwise$$

There are variety of ways to choose the coefficient C_{max}. C_{max} may be taken as an input coefficient, as the largest g value observed thus far, as the largest g value in the current population, or the largest of the last *k* generation.

When the natural objective function formulation is a utility function we have no difficulty with the direction of the function: maximized desired profit or utility leads to desired performance. But still there is some problems with negative utility function as in the particular case it occurs during the fitness calculation of *n* number of features fitness evaluation. To overcome this, we simply transform fitness according to the equation:

$$f\left(x\right)=u\left(x\right)+C_{min}\quad When\,u\left(x\right)+C_{min}>0\quad or 0 Otherwise$$

For the present problem there is a single fitness function to select the best Pareto optimal plane.

4.4 Crossover

Crossover is the genetic operator that mixes two chromosomes together to form new offspring. Crossover occurs only with some probability (crossover probability). Chromosomes that are not subjected to crossover remain unmodified. The intuition behind crossover is the exploration of new solutions and exploitation of old solutions. GAs construct a better solution by mixing the good characteristic of chromosomes together. From the *n* solution strings in the population (simply n/2 pairs), certain adjacent string pairs are randomly selected for present crossover technique. In the standard GA, we use single-point crossover by selecting a pair of strings and swapping substrings at a randomly. No adaptive or probabilistic crossover technique has been used for current experimentation.

4.5 Mutation

Each chromosome undergoes mutation with a probability μ_m. The mutation probability is also selected adaptively for each chromosome as in [26]. The expression for mutation probability, μ_m, is given below:

μ_m= $k_2 \times \frac{(f_{max}-f)}{(f_{max}-\bar{f})}$ if $f > \bar{f}$,

μ_m= k_4 if $f > \bar{f}$, Here, values of k_2 and k_4 are kept equal to 0.5. This adaptive mutation helps GA to come out of local optimum. When GA converges to a local optimum, i.e., when $f_{max} - \bar{f}$ decreases, μ_c and μ_m both will be increased. As a result GA will come out of local optimum. It will also happen for the global optimum and may result in disruption of the near-optimal solutions. As a result GA will never converge to the global optimum. The μ_c and μ_m will get lower values for high fitness solutions and get higher values for low fitness solutions. While the high fitness solutions aid in the convergence of GA, the low fitness solutions prevent the GA from getting stuck at a local optimum. The use of elitism will also keep the best solution intact. For a solution with the maximum fitness value, μ_c and μ_m are both zero. The best solution in a population is transferred undisrupted into the next generation. Together with the selection mechanism, this may lead to an exponential growth of the solution in the population and may cause premature convergence.

Here, each position in a chromosome is mutated with probability μ_m in the following way. The value is replaced with a random variable drawn from a Laplacian distribution, $p\left(\epsilon\right)\alpha\;e^{-\frac{|\epsilon-\mu|}{\delta}}$, where the scaling factor δ sets the magnitude of perturbation. Here, μ is the value at the position which is to be perturbed. The scaling factor δ is chosen equal to 0.1. The old value at the position is replaced with the newly generated value. By generating a random variable using Laplacian distribution, there is a non-zero probability of generating any valid position from any other valid position while probability of generating a value near the old value is more.

4.6 Natural Selection

After we evaluate population's fitness, the next step is chromosome selection. Selection embodies the principle of 'survival of the fittest'. The mutant fittest chromosomes are selected for reproduction. A few poor chromosomes or lower fitness chromosomes may be selected. Each solution having a probability equal to its fitness score divided by the sum of the total solutions scores in the generation. The top *n* solutions at each generation automatically retained and carried over to the next generation. Roulette wheel selection is used to implement the proportional selection strategy.

5 Experimental Results

We have used Java API for Genetic Algorithm[10] application. Approximately 70% of every corpus has been used for training purpose and the rest 30% has been used for testing purpose. The following parameter values are used for the genetic algorithm: population size=50, number of generation=50.

[10] http://www.jaga.org/

Table 6. Results of final GA based classifier.

Languages	Domain	Precision	Recall
English	MPQA	90.22%	96.01%
	IMDB	93.00%	98.55%
Bengali	NEWS	87.65%	89.06%
	BLOG	90.6%	92.40%

The overall precision and recall values of the GBML based subjectivity classifier are shown in Table 6 for all the corpora selected for English and Bengali. It is observed that subjectivity detection is trivial for review corpus and blog corpus rather than for news corpus. In news corpus there is more factual information than review or blog corpus that generally contain people's opinion. Thus subjectivity classification task is domain dependent. But the proposed technique is domain adaptable through the use of natural selection. The difference of GA-based classifier with others statistical system is that a whole sentence could be encoded in GA and could be used as a feature. In other classifier system n-gram method has been followed. The fixed size of n in the n-gram does not fit into the variable string length of an input string.

5.0.1 Comparison

Present GBML sytem outperform than existing Subjectivity systems in literature. The CRF based subjectivity classification system as we reported previously in [12] perform experiment on same set of Bengali and English corpus and reported accuracy of the system was 72.16% and 74.6% for the news and blog domains respectively. In the previous Subjectivity Detection study the subjectivity problem was modeled as a text classification problem that classifies texts as either subjective or objective depending upon various experimentally choosen features. This paper illustrates a Conditional Random Field (CRF) based Subjectivity Detection approach tested on English and Bengali multiple domain corpus. Standard machine learning (ML) techniques needs rigorous permutation and combination wise experimentation to find out the best set of features for any particular problem definition. The GBML based methodology as we proposed here provide a best solution as natural selection method to overcome the classical feature engineering. The CRF based system was tested on the same dataset as reported in Table 7. Besides the novelty over feature engineering GBML technique is a better solution as it need no human interruption to find out best fetures and it choose the best fetures through *natural selection.*

Table 7. Results of final CRF-based subjectivity classifier.

Languages	Domain	Precision	Recall
English	MPQA	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	NEWS	72.16%	76.00%
	BLOG	74.6%	80.4%

In compare to the previous subjectivity classification systems on MPQA corpus the present GBML system has an increment of near about 4.0%. The reported highest accuracy on MPQA using Naive Bayse was 86.3% as reported in [27]. The authors used Naive Bayes sentence classifier and the reported accuracy was as reported in Table 8.

Table 8. Results of previous subjectivity classifier on MPQA.

Languages	Domain	Precision	Recall
English	MPQA	86.3%	71.3%

The accuracy of previous subjectivity detection on the same movie review corpus is 86.4% reported in [28]. The authors proposed a interesting machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions are then categorized using efcient techniques for finding minimum cuts in graphs to incorporate the cross-sentence contextual constraints. To capture cross-sentence contextuality we prefer the theme word features in present GBML based technique. Two standard machine learning (ML) techniques used in [28] as Naive Bayes (NB) and Support Vector Machine (SVM). The reported accuracy of the subjectivity system was as reported in Table 9.

Table 9. Results of previous subjectivity classifier on IMDB.

Classifier	Reported Accuracy
NB	86.40%
SVM	86.15%

6 Conclusion

Application of machine learning algorithms in NLP generally experiments with combination of various syntactic and semantic linguistic features to identify the most effective feature set. Here we viewed this as a multi-objective or multi-criteria optimization search problem. The experiments in the present task start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. In the present paper we use GBML to identify automatically best feature set based on the principle of natural selection and survival of the fittest. The identified fittest feature set is then optimized locally and global optimization is then obtained by multi-objective optimization technique. The local optimization identify the best range of feature values of a particular feature. the Global optimization technique identifies the best ranges of values of given multiple feature.

In the present experimental setup it harder to identify feature wise performance value. The GBML identifies the best feature set and their optimal range value by *natural selection.* The present experiment by us is to determine contribution of each feature to the overall subjectivity problem.

The performance of the present multiple objective optimization tecnique based GBML strategy easily estublised that it is worthy than available ML techniques so far used in NLP. The novelty of the present task is not only towards finding the better way to detect subjectivity moreover it depicts a generation change in ML techiques so far used in NLP.

7 Reference

1. Philip J. Stone. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, 1966.
2. Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. A comprehensive grammar of the English language. Longman, 1985.

3. Wilks Yorick and Bein Janusz. Beliefs, Points of View,and Multiple Environments. In Cognitive Science 7. pp. 95-119 . 1983.
4. Janyce M. Wiebe and William J. Rapaport. A computational theory of perspective and reference in narrative. In Proceedings of the Association for Computational Linguistics (ACL), pages 131–138, 1988.
5. Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold standard data set for subjectivity classifications. In Proceedings of the Association for Computational Linguistics (ACL), pages 246–253, 1999.
6. A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," In the Proceedings of Recent Advances in Natural Language Processing (RANLP), 2005.
7. Vasileios Hatzivassiloglou and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pages 174–181.
8. Janyce Wiebe. 1990. Recognizing Subjective Sentences: A Computational Investigation of Narrative Text. Ph.D. thesis, State University of New York at Buffalo.
9. Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 735–740.
10. Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.
11. A. Das and S. Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009b).
12. A. Das and S. Bandyopadhyay (2009a). Subjectivity Detection in English and Bengali: A CRF-based Approach., In Proceeding of ICON 2009, December 14th-17th, 2009, Hyderabad.
13. J. H. Holland. 1975. Adaptation in Natural and Artificial Systems. The University of Michigan Press, AnnArbor.
14. D. E. Goldberg. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, New York.
15. Kraft, D.H. et. al. "The Use of Genetic Programming to Build Queries for Information Retrieval." In Proceedings of the First IEEE Conference on Evolutional Computation. New York: IEEE Press. 1994. PP. 468-473.
16. Martin Bautista and M.J. "An Approach to An Adaptive Information Retrieval Agent using Genetic Algorithms with Fuzzy Set Genes." In Proceeding of the Sixth International Conference on Fuzzy Systems. New York: IEEE Press. 1997. PP.1227-1232.
17. Abbasi, A., Chen, H., and Salem, A. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Transactions on Information Systems, 26(3), 2008, no. 12.
18. Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 299-305, 2000.
19. Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 27–29, 2006.
20. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), pages 70-77, 2003.
21. Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of Language Resources and Evaluation (LREC), 2006.
22. A. Das and S. Bandyopadhyay (2010a). SentiWordNet for Bangla., In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary , February, 2010, Mysore.
23. A. Das and S. Bandyopadhyay (2010b). Morphological Stemming Cluster Identification for Bangla., In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators, January, 2010, Mysore.
24. A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay (2009). Dependency Parser for Bengali : the JU System at ICON 2009., In NLP Tool Contest ICON 2009, December 14th-17th, 2009a, Hyderabad.
25. Giuseppe Carenini, Raymond Ng, and Adam Pauls. Multi-document summarization of evaluative text. In Proceedings of the European Chapter of the As-sociation for Computational Linguistics (EACL), pages 305–312, 2006.
26. M. Srinivas and L. M. Patnaik. 1994. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Transactions on Systems, Man and Cybernatics, 24(4):656–667.
27. Janyce Wiebe and Ellen Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding of International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Pages 475–486, 2006.
28. Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the Association for Computational Linguistics (ACL), pp. 271–278, 2004.