ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

Sentiment Classification Using Feature Relation Network for N-Gram Data

Sanjeev Kumar Singh, S. R. Yadav

P.G. Student, Computer Science & Engineering, Millennium Institute of Technology, Bhopal, MP, India.
Assistant Professor & Head, Dept. of C.S.E,Millennium Institute of Technology, Bhopal, MP, India.

ABSTRACT; Text categorization is a fundamental task in document processing, allowing the automated handling of enormous streams of documents in electronic form. One difficulty in handling some classes of documents is the presence of different kinds of textual errors, such as spelling and grammatical errors in email, and character recognition errors in documents that come through OCR. Text categorization must work reliably on all input, and thus must tolerate some level of these kinds of problems. N-gram-based approach for text categorization is tolerant of textual errors.
A rule-based multivariate text feature selection method called Feature Relation Network (FRN) has been proposed that considers semantic information and also leverages the syntactic relationships between n-gram features. FRN is intended to efficiently enable the inclusion of extended sets of heterogeneous n-gram features for enhanced sentiment classification. Many experiments have been conducted on three online review test beds in comparison with methods used in sentiment classification. FRN has better performance in comparison with different univariate feature selection methods; it was able to select attributes resulting in significantly better classification accuracy irrespective of the feature subset sizes. Furthermore, by incorporating syntactic information about n-gram relations, FRN is able to select features in a more computationally efficient manner than many univariate techniques such as LL, IG, CHI Squared, WNG/LL BOW/LL.

KEYWORDS: OCR, FRN, LL, IG, N-gram, WNG, BOW.

I. INTRODUCTION

A major concern when incorporating large sets of diverse n-gram features for sentiment classification is the presence of noisy, irrelevant, and redundant attributes. These concerns can often make it difficult to harness the augmented discriminatory potential of extended feature sets. A text feature selection method has been assumed called Feature Relation Network (FRN) that accepts meaningful value and also facilitates the syntactic mappings among n-gram features. Enhanced sentiment classification needs the use of FRN for inclusion of large sets of heterogeneous n-gram features. A lot of observations has been done in comparison with methods used in previous opinion mining research work. FRN is most efficient as compared to univariate, multivariate, and hybrid feature selection processes; it efficiently selects features that results significantly better classification accuracy ignoring the attribute subset sizes. Syntactic information about n-gram relations is being incorporated using FRN in a more computationally efficient manner than many other feature selection techniques.

A. N-gram Data:

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles [1]. Using Latin numerical prefixes, an n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". English cardinal numbers are sometimes used, e.g., "four-gram", "five-gram", and so on. In computational biology, a polymer or oligomer of a known size is called a k-mer instead of an n-gram, with specific names using Greek numerical prefixes such as "monomer", "dimer", "trimer", "tetramer", "pentamer", etc., or English cardinal numbers, "one-mer", "two-mer", "three-mer", etc.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

B. Applications:

An *n*-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n-1)$–order Markov model. [2]*n*-gram models are now widely used in probability, communication theory, computational linguistics (for instance, statistical natural language processing), computational biology (for instance, biological sequence analysis), and data compression. Two benefits of *n*-gram models (and algorithms that use them) are simplicity and scalability – with larger *n*, a model can store more context with a well-understood space–time trade-off, enabling small experiments to scale up efficiently.

An N-gram is an N-character slice of a longer string. Although in the literature the term can include the notion of any co-occurring set of characters in a string (e.g., an N-gram made up of the first and third character of a word), in this paper the term for contiguous slices has been used only. Typically, one slices the string into a set of overlapping-grams. In this system, N-gramsof several different lengths has been used simultaneously. Blanks has been appended to the beginning and ending of the string in order to help with matchingbeginning-of-word and ending-of-word situations.(underscore characterwill be used ("_")to represent blanks.) Thus, the word "TEXT"would be composed of the following N-grams:bi-grams: _T, TE, EX, XT, T_tri-grams: _TE, TEX, EXT, XT_, T_ _quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _In general, a string of length *k*, padded with blanks, will have *k*+1 bi-grams, *k*+1tri-grams,*k*+1 quad-grams, and so on.N-gram-based matching has had some success in dealing with noisy ASCII input in other problem domains, such as in interpreting postal addresses ([4] and [5]), in text retrieval ([6] and[7]), and in a wide variety of other natural language processing applications [8]. The key benefit that N-gram-based matching provides derives from its very nature: since every string is decomposed into small parts, any errors that are presented to affect only a limited number of thoseparts, leaving the remainder intact. If N-grams has been counted that are common to two strings, a measure of their similarity obtained that is resistant to awide variety of textual errors.

C. Text Categorization Using N-GramFrequency Statistics

Human languages invariably have some words which occur more frequently than others. One of the most common ways of expressing this idea has become known as Zipf's Law [8], which can be re-state as follows: The *n*th most common word in a human language text occurs with a frequency inversely proportional to *n*.The implication of this law is that there is always a set of words which dominates most of the other words of the language in terms of frequency of use. This is true both of words in general, and of words that are specific to a particular subject. Furthermore, there is a smooth continuum of dominance from most frequent to least.The smooth nature of the frequency curves helpsin some ways because it implies there should not be worries too much about specific frequency thresholds. This same law holds, at least approximately, for other aspects of human languages. In particular, it is true for the frequency of occurrence of N-grams, both as inflection forms and as morpheme-like word components which carry meaning. (As shown in Figure 1 for an exampleof a Zipfian distribution of N-gram frequencies from a technical document.) Zipf's Lawimplies that classifying documents with N-gramfrequency statistics will not be very sensitive to cutting off the distributions at a particular rank. It also implies that if we are comparing documents from the same category they should have similar-gram frequency distributions.An experimental text categorization system has been built that uses this idea. Figure 2 illustrates the overall data flow for the system. In this scheme, a set of pre-existing text categories (such as subject domains) has been started for which samples are sized reasonably, say, of 10K to20K bytes each. From these, a set of N-gram frequency profiles has been generated to represent each of the categories. When a new document arrives for classification, the system first computes its N-gram frequency profile. It then compares this profile against the profiles for each ofthe categories using an easily calculated distance measure. The system classifies the document as belonging to the category having the smallest distance.

D. Feature Relation Network

For text n-grams, the relationship between n-gram categories can facilitate enhanced feature selection by considering relevance and redundancy, two factors critical to large-scale feature selection [9]. In order to efficiently remove redundant and irrelevant ones FRN has been used. Comparing all features within a feature set directly with one another can be an arduous endeavour. However, if the relationship between features can be utilized, thereby comparing only some logical subset of attributes, then the feature selection process can be made more efficient. Given large quantities of heterogeneous n-gram features, the FRN utilizes two important n-gram relations: Subsumption and parallel relations.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

These two relations enable intelligent comparison between features in a manner that facilitates enhanced removal of redundant and/or irrelevant n-grams.

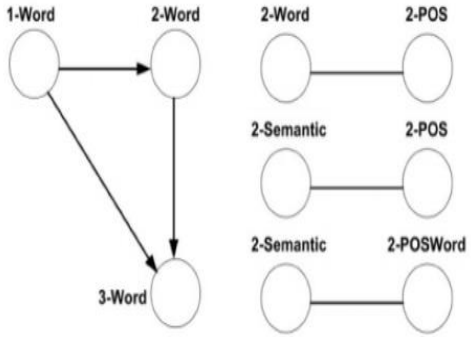

Fig. no. 1: (left) Subsumption relation between word n-grams and (right) parallel relation between various bigrams.

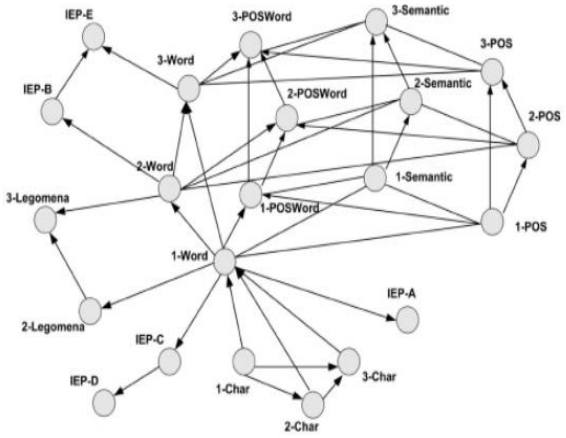

Fig. no. 2:The feature relation network.

1. Subsumption Relations

A subsumption relation occurs between two n-gram feature categories, where one category is a more general, lower order form of the other [10]. A subsumes BðA!BÞif B is a higher order n-gram category whose n-grams contain the lower order n-grams found in A. For example, word unigrams subsume word bigrams and trigrams, while word bigrams subsume word trigrams (as shown on the left side of Fig. 1). Given the sentence "I love chocolate," there are six-word n-grams: I, STUDY, MILLENNIUM, I STUDY, STUDYMILLENNIUM, and I STUDY MILLENNIUM COLLEGE. The unigram STUDY is obviously important, generally conveying positive sentiment. However, what about the bigrams and trigrams? It depends on their weight, as defined by some heuristic (e.g., log likelihood or information gain). We only wish to keep higher order n-grams if they are adding additional information greater than that conveyed by the unigram STUDY. Hence, given A! B, we keep features from category B if their weight exceeds that of their general lower order counterparts found in A by some threshold t [10]. For instance, the bigrams I STUDY and STUDYMILLENNIUM would only be retained if their weight exceeded that of the unigram STUDY by t (i.e., if they provided additional information over the more general unigram). Similarly, the trigram I STUDY MILLENNIUM would only be retained if its weight exceeded that of the unigram STUDY and any remaining bigrams (e.g., I STUDY and INMILLENNIUM) by t.

2. Parallel Relations

A parallel relation occurs where two heterogeneous same order n-gram feature groups may have some features with similar occurrences. For example, word unigrams (1-Word) can be associated with many POS tags (1-POS), and vice versa. However, certain word and POS tags' occurrences may be highly correlated. Similarly, some POS tags and

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

semantic class unigrams may be correlated if they are used to represent the same words. For example, the POS tag ADMIRE_VP and the semantic class SYN-Affection both represent words such as "like" and "love." Given two n-gram feature groups with potentially correlated attributes, A is considered to be parallel to B (A—B). If two features from these categories A and B, respectively, have a correlation coefficient greater than some threshold p, one of the attributes is removed to avoid redundancy. The right side of Fig. 1 shows some examples of bigram categories with parallel relations.

Correlation is a commonly used method for feature selection [11], [12]. However, correlation is generally used as a univariate method by comparing the occurrences of an attribute with the class labels, across instances [11]. Comparing attribute intercorrelation could remove redundancy, yet is computationally infeasible, often necessitating the use of search heuristics [12], [13]. FRN allows the incorporation of correlation information by only comparing select n-grams (ones from parallel relation categories within the FRN).

3. The Complete Network

Fig. 2 shows the entire FRN, consisted of the nodes previously described in Table 3. The network encompasses 22 n-gram feature category nodes and numerous subsumption and parallel relations between these nodes. The detailed list of relations is presented in Table 5. The order in which the relations are applied is important to ensure that redundant and irrelevant attributes are removed correctly. Subsumption relations are applied prior to parallel relations. Furthermore, subsumption relations between n-gram groups within a feature category are applied prior to across categoryrelations (i.e., 1-Word !2-Word isappliedpriorto1-Word !1-POSWord).

E. Sentiment Classification Using Feature Relation Network

Document level sentiment classification is a fundamental task in sentiment analysis and is crucial to understand user generated content in social networks or product reviews (Manning and Schutze, 1999; Jurafsky and Martin, 2000; Pang ¨ and Lee, 2008; Liu, 2012). The task calls for identifying the overall sentiment polarity of a document. In literature, dominant approaches follow (Pang et al., 2002) and exploit machine learn- ing algorithm to build sentiment classifier. Many of them focus on designing hand-crafted features (Qu et al., 2010; Paltoglou and Thelwall, 2010) or learning discriminate features from data, since the performance of a machine learner is heavily dependent on the choice of data representation (Bengio et al., 2015). Document level sentiment classification remains a significant challenge: how to encode the intrinsic (semantic or syntactic) relations between sentences in the semantic meaning of document.

This is crucial for sentiment classification because relations like "contrast" and "cause" have great influences on determining the meaning and the overall polarity of a document. However, existing studies typically fail to effectively capture such information. For example, Pang et al. (2002) and Wang and Manning (2012) represent documents with bag-of-ngrams features and build SVM classifier upon that. Although such feature-driven SVM is an extremely strong performer and hardly to be transcended, its "sparse" and "discrete" characteristics make it clumsy in taking into account of side information like relations between sentences. Recently, Le and Mikolov (2014) exploit neural networks to learn continuous document representation from data. Essentially, they use local ngram information and do not capture semantic relations between sentences. Furthermore, a person asked to do this task will naturally carry it out in a sequential, bottom-up fashion, analyse the meanings of sentences before considering semantic relations between them. This motivates us to develop an end-to-end and bottom-up algorithm to effectively model document representation.

In this paper, we introduce a neural network approach to learn continuous document representation for sentiment classification. The method is on the basis of the principle of compositionality (Frege, 1892), which states that the meaning of a longer expression (e.g. a sentence or a document) depends on the meanings of its constituents. Specifically, the approach models document representation in two steps.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 5 , May 2018

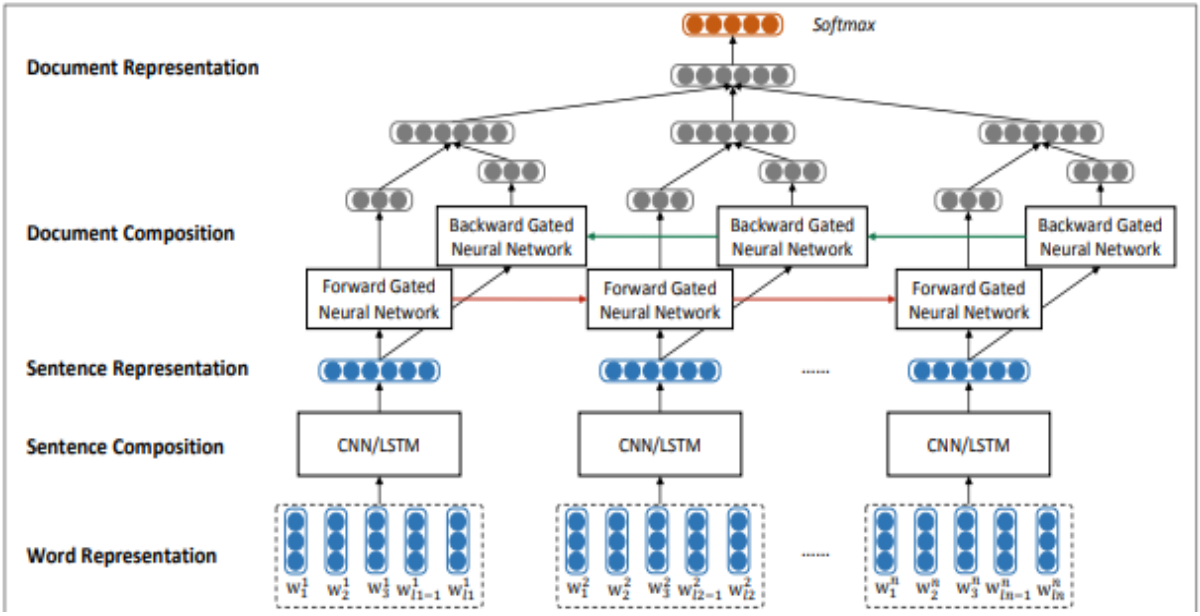

Figure 3: The neural network model for document level sentiment classification. w n i stands for the i-th word in the n-th sentence, ln is sentence length.

In the first step, it uses convolutional neural network (CNN) or long short-term memory (LSTM) to produce sentence representations from word representations. Afterwards, gated recurrent neural network is exploited to adaptively encode semantics of sentences and their inherent relations in document representations. These representations are naturally used as features to classify the sentiment label of each document. The entire model is trained end-to-end with stochastic gradient descent, where the loss function is the cross-entropy error of supervised sentiment classification2

We conduct document level sentiment classi- fication on four large-scale review datasets from IMDB3 and Yelp Dataset Challenge4. We compare to neural network models such as paragraph vector (Le and Mikolov, 2014), convolutional neural network, and baselines such as feature-based SVM (Pang et al., 2002), recommendation algorithm JMARS (Diao et al., 2014). Experimental results show that: (1) the proposed neural model shows superior performances over all baseline algorithms; (2) gated recurrent neural network dramatically outperforms standard recurrent neural network in document modelling. The main contributions of this work are as follows:

• We present a neural network approach to encode relations between sentences in document representation for sentiment classification.

• We report empirical results on four large-scale datasets and show that the approach outperforms state-of-the-art methods for document level sentiment classification.

• We report empirical results that traditional recurrent neural network is weak in modelling document composition, while adding neural gates dramatically improves the classification performance

II. RELATED WORK

Opinion mining involves several important tasks, including sentiment polarity and intensity assignment [26], [27]. Polarity assignment is concerned with determining whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment looks at heather the positive/negative sentiments are mild or strong. Given the two phrases “I don’t like you” and “I hate you,” both would be assigned a negative semantic orientation but the latter would be considered more intense. Effectively classifying sentiment polarities and intensities entails the use of classification methods applied to linguistic features. While several classification methods have been employed for opinion mining, Support Vector Machine (SVM) has outperformed various techniques including Naïve Bayes, Decision Trees, Winnow, etc. [28], [29], [39], [31]. The most popular class of features used for opinion mining is n-grams [28], [38]. Various n-gram categories have attained state-of-the-art results [32], [33]. Larger n-gram feature sets require the use of feature selection methods to extract appropriate attribute subsets. Next, we discuss these two areas: n-gram features and feature selection techniques used for sentiment analysis.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

A. N-Gram Features for Sentiment Analysis

N-gram features can be classified into two categories: fixed and variable. Fixed n-grams are exact sequences occurring at either the character or token level. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining, including word, part-of-speech (POS), character, legomena, syntactic, and semantic n-grams. Word n-grams include bag-of-words (BOWs) and higher order word n-grams (e.g., bigrams, trigrams). Word n-grams have been used effectively in several studies [28]. Typically, unigrams to trigrams are used [32], [33], though 4-grams have also been employed [35]. Word n-grams often provide a feature set foundation, with additional feature categories added to them [36], [33], [35], [37]. Given the pervasiveness of adjectives and adverbs in opinion-rich text, POS tag, n-grams are very useful for sentiment classification [38], [39]. Additionally, some studies have employed word plus part-of-speech (POSWord) n-grams. These n-grams consider a word along with its POS tag in order to overcome word-sense disambiguation in situations where a word may otherwise have several senses[37]. For example, the phrase "quality of the" can be represented with the POSWord trigram "quality-noun of- prep the-det." Character n-grams are letter sequences. For example, the word "like" can be represented with the following two and three letter sequences "li, ik, ke, lik, ike." While character n-grams were previously used mostly for style classification, they have recently been shown to be useful in related affect classification research attempting to identify emotions in text [29]. Legomena n-grams are collocations that replace once (hapax legomena) and twice occurring words (dis legome-na) with "HAPAX" and "DIS" tags [29], [37]. Hence, the trigram "I hate Jim" would be replaced with "I hate HAPAX" provided "Jim" only occurs once in the corpus. The intuition behind such collocations is to remove sparsely occurring words with tags that will allow the extracted n-grams to be more generalizable [40], [37].Syntactic phrase patterns are learned variable n-grams [35]. Riloff et al. [41] developed a set of syntactic templates and information extraction patterns (i.e., instantiations of those templates) reflective of subjective content. Given a set of predefined templates, patterns with the greatest occurrence difference across sentiment classes are extracted. For example, the template "<subj> passive-verb" may produce the pattern "<subj> was satisfied." Such phrase patterns can represent syntactic phenomena difficult to capture using fixed-word n-grams [39], [37].Semantic phrase patterns typically use an initial set of terms or phrases, which are manually or automatically filtered and coded sentiment polarity/intensity information. Many studies have used WordNet to automatically generate semantic lexicons [42], [43] or semantic word classes [44]. Riloff et al. [41] used a semi automated approach to construct sets of strong/weak subjectivity and objective nouns. Others have manually annotated or derived semantic phrases [36], [38].Table 1 provides a summary of n-gram features used for opinion classification. Based on the table, we can see that many n-gram categories have been used in prior opinion mining research. However, few studies have employed large sets of heterogeneous n-grams. As stated before, most studies utilized word n-grams in combination with one other category, such as POS tag, legomena, semantic, or syntactic n-grams, e.g., [28], [36], [33], [35], [37].

III. RESEARCH GAP

A. Research Gaps

Based on our review, we have identified appropriate gaps. Most studies have used limited sets of n-gram features, typically employing one or two categories [14], [15]. Larger n-gram feature sets introduce computational difficulties and potential performance degradation stemming from noisy feature sets. For instance, the popular 2,000 movie review testbed developed by Pang et al. [15] has over 49,000 bag-of-words [16]. Higher order n-gram feature spaces can be even larger, with hundreds of thousands of potential attributes. Feature selection methods are needed to help manage the large feature spaces created from the use of heterogeneous n-grams. As Riloff et al. [10] noted, using additional text features without appropriate selection mechanisms is analogous to "throwing the kitchen sink." However, large-scale feature selection requires addressing relevance and redundancy, something many existing methods fail to do [13].Redundancy is a big problem since there are a finite number of attributes that can be incorporated and n-grams tend to be highly redundant by nature. In the case of univariate methods, redundant features occupy valuable spots that may otherwise be utilized by attributes providing additional information and discriminatory potential. Powerful multivariate methods are capable of alleviating redundancy; however, they are often unsuitable for computational reasons. These methods have typically been applied to smaller feature sets, e.g., [17], [18]. It is unclear whether hybrid feature selection methods have the potential to overcome issues stemming from redundancy. More-ver, most of the feature election methods described are generic techniques that have been applied to a plethora of problems, since they assess attribute relevance solely based on the training data. Whenever possible, domain knowledge should be

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 5 , May 2018

incorporated into the feature selection process [19]. Existing lexicons and knowledge bases pertaining to the semantic and syntactic properties of n-grams could be exploited for enhanced assessment of relevance and redundancy associated with text attributes.

Table 1: N-Gram Feature Set

Label	Description	Examples	
N-Char	Character- level n-grams	1-Char	I, S, T, U, D, Y, I, N, M, I, L, L, E, N, N, I, U, M, C, O, L, L, E, G, E
		2-Char	IS, TU, DY, IN, MI, LL, EN, NI, UM,CO, LL, EG,
		3-Char	IST, DYI, NMI, LLE, NNI, UMC, OLL, EGE
N-Word	Word-level n-grams	1-Word	I, STUDY, IN MILLENNUM, COLLEGE
		2-Word	I STUDY, IN MILLENNIUM
		3-Word	I STUDY IN,
N-POS	Part-of-speech tag n-grams	1-POS	I, ADMIRE_VBP, NN
		2-POS	ADMIRE_VBP NN
		3-POS	I ADMIRE VBP NN
N-POSWord	Word and POS tag n-grams	1-POSWord	STUDY ADMIRE_VBP
		2-POSWord	I ISTUDY ADMIRE_VBP
		3-POSWord	I STUDY ADMIRE_VBP CHOCOLATE NN
N-Legomena	Hapax legomena and Dis legomena n-grams	2-Legomena	STUDY DIS
		3-Legomena	I STUDY DIS
N-Semantic	Semantic class n-grams	1-Semantic	SYN-Pronoun, SYN-Affection
		2-Semantic	SYN-Pronoun SYN-Affection
		3-Semantic	SYN-Pronoun SYN-Affection SYN-Candy
IEP-A/E	Information extraction patterns	IEP-A	<possessive> NP, <subj>AuxVP AdjP, <subj>AuxVPDobj, ActVP<dobj>, ActVP Prep <np>
		IEP-B	<subj>PassVP, InfVP Prep <np>, InfVP<dobj>
		IEP-C	<subj>ActVP
		IEP-D	<subj>ActVPDobj
		IEP-E	<subj>ActInfVP, <subj>PasslnfVP, ActInfVP<dobj>

IV. METHODOLOGY

A. Univariate Methods Used for Sentiment Classification

Multivariate methods consider attribute groups or subsets. These techniques sometimes use a wrapper model for attribute selection, where the accuracy of a target classifier is used as an evaluation metric for the predictive power of a particular feature subset [20]. Examples include decision tree models, recursive feature elimination, and genetic algorithms. By performing group-level evaluation, multivariate methods consider attribute interactions. Consequently, these techniques are also computationally expensive in relation to univariate methods. Decision tree models (DTMs) use a wrapper, where a DTM is built on the training data and features incorporated by the tree are included in the feature set [9]. Recursive feature elimination uses a wrapper model based on an SVM classifier [21]. During each iteration, the remaining features are ranked based on the absolute values of their SVM weights, and a certain number/percentage of these are retained [22], [23], [24]. Genetic algorithms (GAs) have been used to search for ideal subsets across the feature subspace in text classification problems such as style [13] and sentiment analysis [23]. A major pitfall associated with GA is that they can be computationally very expensive, since hundreds/thousands of solutions have to be evaluated using a classifier [23]. Feature subsumption hierarchies (FSHs) use the idea of performance-based feature subsumption to remove redundant or irrelevant higher order n-grams [25]. Only those word

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

bigrams and trigrams are retained, which provide additional information over the unigrams they encompass. Table 3 shows multivariate methods used for sentiment classification.

TABLE 2

Chi Squared [35]
$\chi^2(a,Y) = \sum_{a_{x_j} \in \{0,1\}} \sum_{i \in Y} \frac{\left(F(a_{x_j}, Y=i) - E(a_{x_j}, Y=i)\right)^2}{E(a_{x_j}, Y=i)}$ where : $\chi^2(a,Y)$ is the chi - squared value for feature a across classes Y $X = [x_1, x_2, \ldots x_m]$ are the training examples $a_{x_j} = 1$ if the training instance x_j contains feature a, $a_{x_j} = 0$ otherwise $F(a_{x_j}, Y=i)$ is the observed frequency of a_x, when $Y=i$ $E(a_{x_j}, Y=i) = \frac{p(a)\,p(Y=i)}{m}$ is the expected value of a_x, when $Y=i$, across X
Information Gain [2, 3, 13]
$IG(Y,a) = H(Y) - H(Y \mid a)$ where : $IG(Y,a)$ is the information gain for feature a $H(Y) = -\sum_{i \in Y} p(Y=i) \log_2 p(Y=i)$ is the entropy across classes Y $H(Y \mid a) = -\sum_{j \in a} p(a=j) \sum_{i \in Y} p(Y=i \mid a=j) \log_2 p(Y=i \mid a=j)$ is the entropy of $Y \mid a$
Log Likelihood Ratio [12, 27, 39]
$w(a) = \max_i \left(p(a \mid Y=i) \log \frac{p(a \mid Y=i)}{p(a \mid \neg Y=i)} \right)$ where : $w(a)$ is the log likelihood for feature a across classes Y

Multivariate methods consider attribute groups or subsets. These techniques sometimes use a wrapper model for attribute selection, where the accuracy of a target classifier is used as an evaluation metric for the predictive power of a particular feature subset [20]. Examples include decision tree models, recursive feature elimination, and genetic algorithms. By performing group-level evaluation, multivariate methods consider attribute interactions. Consequently, these techniques are also computationally expensive in relation to univariate methods. Decision tree models (DTMs) use a wrapper, where a DTM is built on the training data and features incorporated by the tree are included in the feature set [9]. Recursive feature elimination uses a wrapper model based on an SVM classifier [21]. During each iteration, the remaining features are ranked based on the absolute values of their SVM weights, and a certain number/percentage of these are retained [22], [23], [24]. Genetic algorithms (GAs) have been used to search for ideal subsets across the feature subspace in text classification problems such as style [13] and sentiment analysis [23]. A major pitfall associated with GA is that they can be computationally very expensive, since hundreds/thousands of solutions have to be evaluated using a classifier [23]. Feature subsumption hierarchies (FSHs) use the idea of performance-based feature subsumption to remove redundant or irrelevant higher order n-grams [25]. Only those word bigrams and trigrams are retained, which provide additional information over the unigrams they encompass. Table 3 shows multivariate methods used for sentiment classification across the feature subspace in text classification problems such as style [18] and sentiment analysis. A major pitfall associated with GA is that they can be computationally very expensive, since hundreds/thousands of solutions have to be evaluated using a classifier. Feature subsumption hierarchies (FSHs) use the idea of performance-based feature subsumption to remove redundant or irrelevant higher order n-grams [10]. Only those word bigrams and trigrams are retained, which provide additional information over the unigrams they encompass. Table 3 shows multivariate methods used for sentiment classification.

B. Generating N-Gram Frequency Profiles

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

The bubble in Figure 2 labelled "Generate Profile" is very simple. It merely reads incoming text and counts the occurrences of all N-grams. To do this, the system performs the following steps:

- Split the text into separate tokens consisting only of letters and apostrophes. Digits and punctuation are discarded. Pad the token with sufficient blanks before and after.
- Scan down each token, generating all possible N-grams, for N=1 to 5. Use positions that span the padding blanks, as well.
- Hash into a table to find the counter for the N-gram and increment it. The hash table uses a conventional collision handling mechanism to ensure that each N-gram gets its own counter.
- When done, output all N-grams and their counts.
- Sort those counts into reverse order by the number of occurrences. Keep just the N grams themselves, which are now in reverse order of frequency.

C. FRN Algorithm

Let $A=\{a_1,a_2,...a_n\}$ and $B=\{b_1,b_2,...b_m\}$ denote two sets of n - grams
//e.g., 1 - Word, 3 - Word, etc.

if $A \rightarrow B$ //A subsumes B
  For each a_x, where $a_x = (a_{x1},...a_{xd})$ denotes a tuple in A with $w(a_x) > 0$
  Let $C \subseteq B$, where $C=\{c_1,c_2,...c_y\}$
  And each $c_x = (c_{x1},...c_{xe})$ denotes a tuple in C with $w(c_x) > 0$
  Where the tuple a_x is a part of each c_x
    if $s(a_x)=s(c_x)$ //check the semantic orientation of the two features
      if $w(a_x) \geq w(c_x) - t$
        $w(c_x) = 0$

if $A - B$ //A is parallel to B
  For each a_x, where $a_x = (a_{x1},...a_{xd})$ denotes a tuple in A with $w(a_x) > 0$
  Let $C \subseteq B$, where $C=\{c_1,c_2,...c_y\}$
  And $c_x = (c_{x1},...c_{xe})$ denotes a tuple in C with $w(c_x) > 0$
  Where each c_x is potentially correlated with a_x
    if $Corr(a_x,c_x) \geq p$
      if $w(a_x) \geq w(c_x)$ then $w(c_x) = 0$

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

if $w(a_x) < w(c_x)$ then $w(a_x) = 0$

Where :

Corr(a,b) is the correlation coefficient for a and b across the m training instances :

$$\text{Corr}(a,b) = \frac{\sum_{x=1}^{m}(a_x - \bar{a})(b_x - \bar{b})}{\sqrt{\sum_{x=1}^{m}(a_x - \bar{a})^2 \sum_{x=1}^{m}(b_x - \bar{b})^2}}$$

$w(a_x)$ is the weight for feature a_x, computed as described in Fig. 3

$$s(a_x) = \arg\max_{v,w}\left(P(a_x \mid v)\log\left(\frac{P(a_x \mid v)}{P(a_x \mid w)}\right)\right)$$

t and p are predefined thresholds //we used $t = 0.05$ and $p = 0.90$

V. RESULTS AND ANALYSIS

We ran the FRN in comparison with LL, IG, and CHI. All four of these feature selection methods were run on the extended feature set described in Section 3.1, which encompassed the word, POS, POSWord, character, legomena, syntactic, and semantic n-grams. In order to assess the impact of using the extended feature set, we also compared two additional feature sets: bag-of-words and word n-grams. These feature sets were only run in conjunction with LL, resulting in two additional feature/feature selection combinations, BOW/LL and WNG/LL. BOW/LL constituted a baseline while WNG/LL was employed since it had performed well in prior opinion classification studies [14]. For the three feature sets (i.e., all n-grams, WNG, and BOW), we extracted all feature occurring at least three times, [15]. The extracted features were ranked using the aforementioned four feature selection methods on the training data for each of the five cross-validation folds. Hence, for each fold, the weights for all features occurring three times or more in the 1,600 training reviews were computed.

When comparing feature sets and selection methods, it is difficult to decide upon the number of features that should be included. Different feature set sizes can wield varying performance depending on the nature of the features and selection methods employed. In order to allow a fair comparison between feature selection methods, we evaluated the top 10,000 to 100,000 features (i.e., the highest weighted/ranked attributes), in 2,500 feature increments. Hence, 37 feature quantities were used for all three feature sets. The total number of BOW typically did not exceed 20,000, so only that many were evaluated. Such a setup is consistent with experimental designs used in prior research.

Fig. 5 shows the results for all six methods across the three testbeds. The table on the left of the figure shows the best percentage accuracy, the number of features used to attain these best results, pairwise t-test results using this number of features on random 90-10 training-testing splits (n=30), the area under the curve (AUC), and p-values for pairwise t-tests across the different feature subset sizes (n=37). The first t-test was intended to measure the significance of the best results, while the second measured the overall effectiveness across feature subset sizes. BOW/LL was not compared on the second t-test, since it did not have enough features to generate a sufficient number of feature subsets. The charts on the right show the results forall 37 feature subsets (using the top 10,000 to 100,000 features). Looking at the left side of Fig. 6, FRN outperformed LL, IG, CHI, WNG/LL, and BOW/LL on all three testbeds in terms of best accuracy and AUC. FRN's best accuracy values were 3-4 percent better than any of the comparison techniques across all three testbeds. Based on the pair wise t-test results, FRN significantly outperformed the comparison methods, with all p-values significant at alpha =0.05.

Feature selection methods, using between 10,000 and 100,000 features. FRN outperformed LL, CHI, WNG/LL, and BOW/LL on all three testbeds by a wide margin, with considerably better accuracy on virtually all feature subset sizes. It has been also outperformed IG on all but one feature subset size on the movie and automobile review data sets. However, IG had slightly better accuracy on a few of the 37 feature subset sizes on the digital camera testbed. Nevertheless, FRN had a higher AUC and its best accuracy was 2 percent greater than that of IG.Looking at the results by feature set, techniques that utilized the extended feature set (i.e., LL, IG, and CHI) outperformed WNG/LL and the BOW/LL baseline on the digital camera data sets. They also had slightly better performance on the automobile data set. However, WNG/LL had better performance on the movie testbed. Overall, the extended feature set did not provide a significant performance increase over word n-grams when using univariate feature selection methods. This is not surprising since the extended feature set includes many redundant attributes across the various categories, which univariate feature selection methods are unable to remove. Consequently, the univariate methods require more attributes from the extended feature set to get the necessary depth required for enhanced opinion classification

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 5 , May 2018

accuracy; only a subset of the highest weighted features is truly providing additional discriminatory potential. This is evidenced by the general upward slope of LL, CHI, and IG as the feature subset sizes increase. The results emphasize the need to combine rich, extended feature sets with more powerful feature selection techniques.

Table [3]: Digital Camera (Epinions)

Selection Methods	Best Acc.	T-test p-values	# Feat.	AUC	T-test p-value
FRN	55.6	0	100	897.35	0
IG	53.32	0.001	40	885.4	0.027
CHI	53.05	0.001	85	881.15	0.001
LL	49.51	0.001	90	876.2	0.001
WNG/LL	46.78	0.001	65	810.26	0.001
BOW/LL	42.07	0.001	10	757.26	0

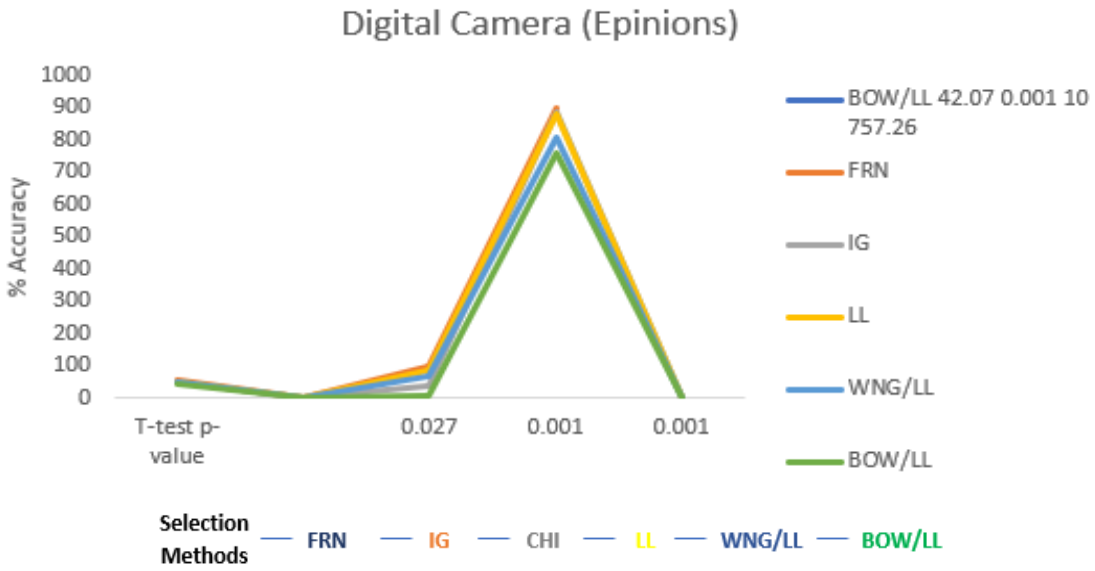

Fig. [5]:Digital Camera (Epinions)

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 5 , May 2018

Table [3]: Automobiles (Edmunds)

Selection Methods	Best Acc.	T-test p-values	# Feat.	AUC	T-test p-value
FRN	56.81	0	100	890.35	0
IG	54.32	0.001	40	875.4	0.027
CHI	52.46	0.001	85	870.15	0.001
LL	49.31	0.001	90	865.2	0.001
WNG/LL	48.26	0.001	65	811.26	0.001
BOW/LL	45.4	0.001	10	752.26	0

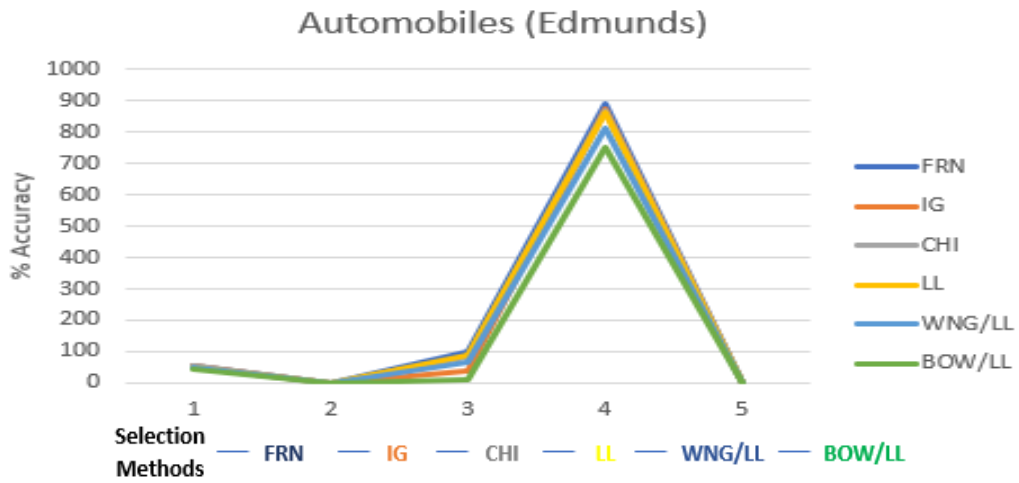

Fig. [6]: Automobiles (Edmunds)

Table [4]: Movie Reviews (Rotten Tomatoes)

Selection methos		Best Acc.	T-test p-values	# Feat.	AUC	T-test p-value
FRN		75.81	0	100	990.35	0
IG		70.32	0.001	40	950.4	0.027
CHI		64.46	0.001	85	930.15	0.001
LL		60.31	0.001	90	910.2	0.001
WNG/LL		57.26	0.001	65	880.26	0.001
BOW/LL		54.4	0.001	10	850.26	0

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 5 , May 2018

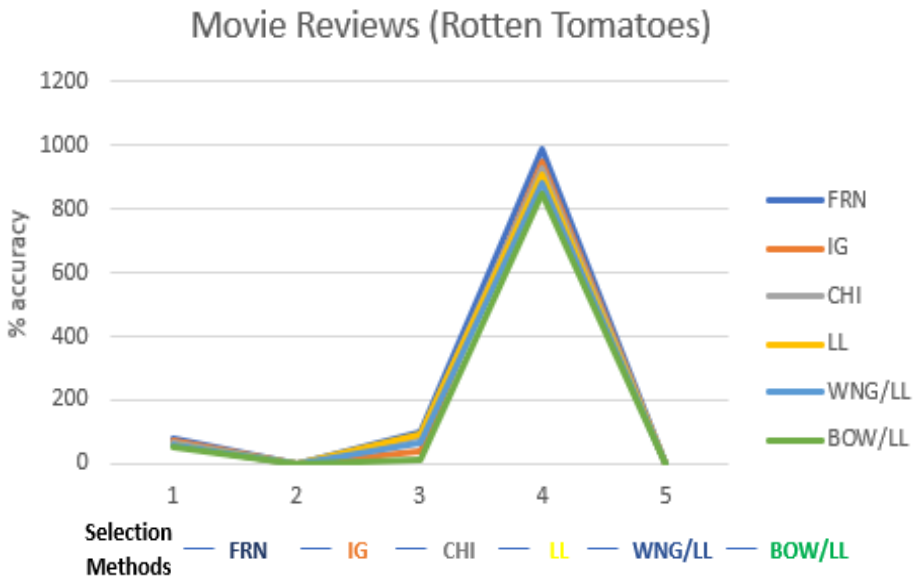

Fig. [7]: Movie Reviews (Rotten Tomatoes)

VI. CONCLUSION AND SCOPE OF FUTURE WORK

In this study use of FRN for improved selection of text attributes for Enhanced sentiment classification has been done. Feature Relation Network's use of syntactic relation and semantic information regarding n- gram enabled it to achieve improved result over various univariate feature selection methods Based on the result obtained in this study a few research directions have been identified. FRN may be suitable for other text classification problems where semantic information available. It is also intended to explore the additional potential feature occurrence measurement. In this study feature present vector has been used, other would be added resulting in multidimensional FRN alternate semantic weighting mechanism would also be explore.

REFERENCES

[1]. Jump up^ Broder, Andrei Z.; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey (1997). "Syntactic clustering of the web". Computer Networks and ISDN Systems. 29 (8): 1157–1166. doi:10.1016/s0169-7552(97)00031-7.
[2].Jump up^ https://www.coursera.org/learn/natural-language-processing/lecture/UnEHs/07-01-noisy-channel-model-8-33
[3]. Cavnar, William B. and Vayda, Alan J., "Using superimposed coding of N-gram lists for Efficient Inexact Matching", Proceedings of the Fifth USPS Advanced Technology Conference, Washington D.C., 1992.
[4] Cavnar,William B. and Vayda, Alan J., "Ngram-based matching for multi-field database access in postal applications", Proceedings of the 1993 Symposium On Document Analysis and Information Retrieval, University of Nevada, Las Vegas.
[5] Cavnar, William B., "N-Gram-Based Text Filtering For TREC-2," to appear in the proceedings of The Second Text Retrieval Conference (TREC-2), ed. by, Harman, D.K., NIST, Gaithersburg, Maryland, 1993.
[6] Kimbrell, R.E., "Searching for Text? Send and N-gram!" Byte, May 1988, pp. 297- 312.
[7] Suen, Ching Y., "N-Gram Statistics for Natural Language Understanding and Text Processing," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI- 1, No. 2, April 1979, pp.164-172.
[8] Zipf, George K., Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology, Addison-Wesley, Reading, Mass.,
[9] H. Liu and H. Motada, Feature Extraction, Construction, and Selection—Data Mining Perspective. Kluwer Academic Publishers, 1998.
[10] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.
[11].Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley Longman, Boston, MA, USA.
[12].Etemadpour, R., Olk, B., and Linsen, L. (2014d). Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. In 5th International Conference on Information Visualization Theory and Applications (IVAPP), pages 233-246, Lisbon, Portugal
[13] J. Li, R. Zheng, and H. Chen, "From Fingerprint to Writeprint," Comm. ACM, vol. 49, no. 4, pp. 76-82, 2006
[14] H. Cui, V. Mittal, and M. Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," Proc. 21st AAAI Conf. Artificial Intelligence, pp. 1265-1270, 2006.
[15] S.R. Das and M.Y. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," Management Science, vol. 53, no. 9, pp. 1375-1388, 2007.
[16]. Inselberg, A. Parallel Coordinates: Visual Multidimensional Geometry and Its Applications; Springer: New York, NY, USA, 2009.
[17]. Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., and Mller, T. (2010). Dimstiller: Workows for dimensional analysis and reduction. In IEEE VAST, pages 3-10. IEEE.
[18]. Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. IEEE Trans. Visualization and Computer Graphics (TVCG) (Proc. InfoVis), 19(12):2376-2385.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

[19].Rensink, R. A. and Baldridge, G. (2010). The perception of correlation in scatterplots. Comput. Graph. Forum, 29(3):1203-1210.
[20] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157- 1182, 2003.
[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, 2002.
[22] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.
[23] G. Mishne, "Experiments with Mood Classification," Proc. Stylistic Analysis of Text for Information Access Workshop, 2005.
[24] D. Mladenic, J. Brank, M. Grobelnik, and N. Milic - Frayling, "Feature Selection Using Linear Classifier Weights: Interaction with Classification Models," Proc. ACM SIGIR, pp. 234-241, 2004.
[25] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 440-448, 2006.
[26] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. ACM SIGKDD, pp. 168-177, 2004.
[27] A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Human Language Technology, Empirical Methods in Natural Language Processing, pp. 339-346, 2005.
[28] A. Abbasi and H. Chen, "CyberGate: A System and Design Framework for Text Analysis of Computer Mediated Communication," MIS Quarterly, vol. 32, no. 4, pp. 811-837, 2008.
[29] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1168-1180, Sept. 2008.
[30] H. Cui, V. Mittal, and M. Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," Proc. 21st AAAI Conf. Artificial Intelligence, pp. 1265-1270, 2006.
[31] B. Pang and L. Lee, "A Sentimental Education: Sentimental Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proc.42ndAnn.MeetingoftheAssoc. ComputationalLinguistics, pp. 271-278, 2004.
[32] A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Trans. Information Systems, vol. 26, no. 3, article no. 12, 2008.
[33] V. Ng, S. Dasgupta, and S.M.N. Arifin, "Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews," Proc. Conf. Computational Linguistics, Assoc. for Computational Linguistics, pp. 611-618, 2006.
[34] B. Pang, L. Lee, and S. Vaithyanathain, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 79-86, 2002.
[35] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 440-448, 2006.
[36] S. Argamon, C. Whitelaw, P. Chase, S.R. Hota, N. Garg, and S. Levitan, "Stylistic Text Classification Using Functional Lexical Features," J. Am. Soc. Information Science and Technology, vol. 58, no. 6, pp. 802-822, 2008.
[37] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning Subjective Language," Computational Linguistics, vol. 30, no. 3, pp. 277-308, 2004.
[38] Z. Fei, J. Liu, and G. Wu, "Sentiment Classification Using Phrase Patterns," Proc. Fourth IEEE Int'l Conf. Computer Information Technology, pp. 1147-1152, 2004.
[39] M. Gamon, "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis," Proc. 20th Int'l Conf. Computational Linguistics, pp. 841- 847, 2004.
[40] J. Wiebe, T. Wilson, and M. Bell, "Identifying Collocations for Recognizing Opinions," Proc. Assoc. for Computational Linguistics, European Chapter of the Assoc. for Computational Linguistics Workshop Collocation, 2001.
[41] E. Riloff, J. Wiebe, and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," Proc. Seventh Conf. Natural Language Learning, pp. 25-32, 2003.
[42] S. Kim and E. Hovy, "Determining the Sentiment of Opinions," Proc. 20th Int'l Conf. Computational Linguistics, pp. 1367-1373, 2004.
[43] G. Mishne, "Experiments with Mood Classification," Proc. Stylistic Analysis of Text for Information Access Workshop, 2005.
[44] A. Burgun and O. Bodenreider, "Comparing Terms, Concepts, and Semantic Classes in WordNet and the Unified Medical Language System," Proc. North Am. Assoc. Computational Linguistics Workshop, pp. 77-82, 2001.