Extracting Emotion Topics from Blog Sentences – Use of Voting from Multi-Engine Supervised Classifiers

Dipankar Das
Department of Computer Science and Eng.,
Jadavpur University
188, Raja S.C. Mullick Road, Kolkata-700032
(+91) 033-24146648
dip_nil2004@yahoo.co.in

Sivaji Bandyopadhyay
Department of Computer Science and Eng.,
Jadavpur University
188, Raja S.C. Mullick Road, Kolkata-700032
(+91) 033-24146648
sivaji_cse_ju@yahoo.com

ABSTRACT
This paper presents a supervised multi-engine classifier approach followed by voting to identify emotion topic(s) from English blog sentences. Manual annotation of the English blog sentences in the training set has shown a satisfactory agreement with *kappa* (κ) measure of 0.85 and MASI (Measure of Agreement on Set-valued Items) measure of 0.82 for emotion topic spans. The baseline system based on object related dependency relations includes the topic oriented thematic roles present in the verb based syntactic frame of the sentences. In contrast, the supervised approach consists of three classifiers, Conditional Random Field (CRF), Support Vector Machine (SVM) and a Fuzzy Classifier (FC). The important features are incorporated based on the ablation study of all features and Information Gain Based Pruning (IGBP) on the development set. One or more emotion topics associated with focused target span are identified based on the majority voting of the classifiers. The supervised multi-engine classifier system has been evaluated with average *F-scores* of 70.51% and 90.44% for emotion topic and target span identification respectively on 500 gold standard test sentences and has outperformed the baseline system.

Categories and Subject Descriptors: H.3.4 [**Context**]: Textual Context, Text Span

General Terms: Algorithms, Measurement, Performance, Design, Experimentation, Human Factors, Languages.

Keywords: Emotion topic, Target, CRF, SVM, Fuzzy Classifier, Voting.

1. INTRODUCTION
The Topic is the real world object, event, or abstract entity that is the primary subject of the opinion as intended by the opinion holder [23]. Topic span associated with an opinion expression is the closest minimal span of text that mentions the topic and target span is the text span that covers the syntactic surface form comprising the contents of the opinion [23]. In our present task, the same definition and terminology are used for identifying topic span and target span. Let us consider the following examples:

SMUC'10, October 30, 2010, Toronto, Ontario, Canada.

Example 1.
"*He first cried up the* ***toy car***".
Example 2.
"*Max ignored the* ***issues of sports*** *as well as* ***politics*** ".
Example 3.
"{I *enjoyed* the **summer vacation**} [because I had a golden chance to **play cricket** in that period].
Example 4.
"{I am currently *angry*}[because I want **Jarred** to take me for **DQ Blizzard**].

In Example 1, the sentence contains the topic span "*toy car*" with respect to the emotion expression "*cried up*" and the emotion holder "*he*". In Example 2, the sentence contains multiple emotion topics as shown in bold face associated with the underlined target span. The topics are related to the emotional expression "*ignore*". The identification of topic span is difficult within the single target span of the opinion as there are multiple potential topics, each identified with its own topic span [23] "[Wilson, personal communication]", "[Wiebe, personal communication]".

Although the identification of topic spans is difficult, the information of emotion topics is useful for the domain of Question Answering (QA), Information Retrieval (IR), product reviews, social media, stock markets, and customer relationship management [6]. Major studies on Opinion Mining and Sentiment Analyses have been attempted with more focused perspectives rather than fine-grained emotions. Especially, the blog posts contain instant views, updated views or influenced views regarding single or multiple emotion topics. Thus the present task deals with the identification of emotion topics from English blog sentences.

In the present task, topics related to the emotional expressions are identified from the sentences of an English blog corpus [1]. Each of the sentences in the blog corpus is annotated with emotional expressions, sentential emotion tags and intensities but not annotated with emotion topics. Hence, three annotators have carried out the annotation of topic and target spans in 1800 blog sentences. The agreement of the annotated topic spans and target spans is measured using Cohen's *kappa* (κ) [4] and measure of agreement on set-valued items (MASI) [20]. The average agreements of 0.85 and 0.83 are obtained for topic span and 0.97 and 0.93 for target span annotation using *kappa* and MASI measures respectively. Both the results show acceptable agreements.

The baseline system is developed based on the parsed constituents of the *object* related dependency relations with the additional clues from *Thematic Roles* of *Topic* type. The phrase segments containing *Topic* as the *Thematic Role* are extracted from the verb based syntactical argument structures of the sentences. The argument structures are acquired from VerbNet [11]. The error analysis shows that the argument structures fail to capture the topic spans if multiple potential emotion topics are present in a sentence. In addition to that, the baseline system also suffers from identifying the scope of each emotion topic.

Topics are generally distributed in different text spans of writer's text. The writer's direct as well as indirect emotional intentions are reflected in the target span in a sentence by mentioning one or more emotional topics. In Example 1, the emotional expression (*cried up*) and target span (*the toy car*) (in this case it is the also the topic span) are both present in a single clause. But, there are some cases where the emotion topics occur in two different text spans. In Example 3, the sentence contains two potential emotion topics "*summer vacation*" and "*play cricket*" in two different clauses but both the topics point to the single emotional expression "*enjoy*" directly and indirectly.

Hence, a supervised approach is adopted to identify multiple emotion topics along with their topic and target spans from each of the blog sentences. The supervised system consists of three different classifiers. Conditional Random Field (CRF) [17], Support Vector Machine (SVM) [9] and Fuzzy Classifier (FC) [13] are employed to identify the target span and topic span considering various features and their combinations. The annotated emotional expressions along with *direct* and *transitive* dependencies, *causal verbs*, *discourse markers, Emotion Holder*, *Named Entities* and four types of similarity measures like *Structural Similarity, Sentiment Similarity, Syntactic Similarity* and *Semantic Similarity* are incorporated as crucial features based on the ablation study conducted on 300 development sentences. Information Gain Based Pruning (IGBP) is carried out for removal of unnecessary non-emotional and non-topical words (e.g. *gather*, *seem*) from the sentences to highlight the emotional as well as topical words in the sentences. The special feature, *Structural Similarity* is based on the Rhetorical Structure Theory (RST) that describes about various parts of a text, how they can be arranged and connected to form a whole text [2]. The theory maintains that consecutive discourse elements, termed *text spans*, which can be in the form of clauses, sentences, or units larger than sentences, are related by a relatively small set (20–25) of *rhetorical relations* [14], [15] between *nucleus* and *satellite*. The primary goal of the writer is termed as *nucleus* whereas the part that provides supplementary material is termed as *satellite.* The portions of *nucleus* and *satellite* are marked using "{}" and "[ ]" notations in the earlier examples.

Here, we have combined the three classifiers to build a multi-engine framework. Emotion topics and targets are identified based on majority voting on the classifier outputs or on the output from the classifier having highest *F-Score* (obtained using cross-validation techniques). The supervised multi-engine system followed by voting technique achieves the *F-Scores* of 70.51% and 90.44% for topic and target span identification respectively. The error analysis shows that in few cases the supervised system fails to distinguish emotion topics from other potential non-emotion topics. The emotion topics represented as metaphors or unstructured texts create problem in the identification task.

The rest of the paper is organized as follows. Section 2 describes the related work. The emotion topic and target annotation is discussed in Section 3. The baseline system is described in Section 4. The supervised framework with features is discussed in Section 5. Evaluation results along with feature analysis and error reducing mechanisms are specified in Section 6. Finally Section 7 concludes the paper

2. RELATED WORK

In the related area of opinion topic extraction, different researchers contributed their efforts. Some of the works are mentioned in [12], [22], [26]. But, all these works are based on lexicon look up and are applied on the domain of product reviews. The topic annotation task on the MPQA corpus is described in [23]. The authors have pointed out that the target spans alone are insufficient for many applications as they neither contain information indicating which opinions are about the same topic, nor provide a concise textual representation of the topics. The introduction of rhetorical structure in our present task helps in identifying more focused target span associated with relevant topics related to the emotional expressions.

The method of identifying an opinion with its holder and topic from online news is described in [10]. The model extracts opinion topics for subjective expressions signaled by verbs and adjectives. They have extracted the topics associated with a specific argument position based on verb or adjective. Similarly, the verb based argument extraction and associated topic identification is followed in the present baseline system. But, the incorporation of four types of similarity features in training and use of multi-engine novel voting technique contributes significantly in emotion topic and target span identification.

Opinion topic identification differs from topic segmentation [3]. The opinion topics are not necessarily spatially coherent as there may be two opinions in the same sentence on different topics, as well as opinions that are on the same topic separated by opinions that do not share that topic [23]. The hypothesis is established by applying the technique of co-reference classification for topic annotation. The building of fine-grained topic knowledge based on rhetorical structure and segmentation of topics using four types of similarity features substantially reduces the problem of emotion topic distinction in our present supervised framework.

The knowledge of Rhetorical Structure Theory in the text structure was used in [19] to improve the identification of topical words in a text document. The similarity between a text and its title is used to identify the text structure. But, the work done at document level was not aimed for opinion topic. The present technique is also applied to identify multiple emotion topics in a sentence. The use of *causal verbs, Emotion Holders, discourse markers* and *rhetorical structures* discover relations among topical entities that appear in the target span.

3. ANNOTATION

One of the major problems of emotion topic extraction is the lack of appropriately annotated corpora. The blog corpus [1] tagged with any of the Ekman's [7] six emotion types at sentence level is

considered in our present task. Emotional expressions and sentential intensities are also annotated in the corpus. But, the corpus does not contain any information related to emotion topic. Three annotators presented as A1, A2 and A3 have used an open source graphical tool (http://gate.ac.uk/gate/doc/releases.html) to carry out the annotation on 1200 sentences. As an individual emotion topic consists of a single word or a string of successive words, the annotation task is conducted to identify the scope of the topic spans in a sentence. To accomplish the goal, we have used two standard metrics for measuring inter-annotator agreement.

Firstly, we have used Cohen's *kappa* coefficient (κ) [4]. It is a statistical measure of inter-rater agreement for qualitative (categorical) items. It measures the agreement between two raters who separately classify items into some mutually exclusive categories. Secondly, we have chosen the measure of agreement on set-valued items (MASI) that was used for measuring agreement on co-reference annotation [20] and in the evaluation of automatic summarization [21]. MASI is a distance between sets whose value is 1 for identical sets and 0 for disjoint sets. For sets A and B it is defined as:

MASI = J * M, where the Jaccard metric (*J*) is

$$J = |A \cap B| / |A \cup B|$$

Monotonicity (*M*) is defined as,

$$1, if A = B$$

$$2/3, if A \subset B or B \subset A$$

$$1/3, if A \cap B \neq \phi, A - B \neq \phi, and B - A \neq \phi$$

$$0, if A \cap B = \phi$$

It is observed that in both strategies, the agreement for annotating target span is (≈ 0.9) signifying highly moderate annotation. But, the disagreement occurs in topic span annotation. The selection of emotion topic from other relevant topics causes the disagreement. It is found that the average number of emotion topics in sentences containing multiple topics is 2~3. The inter-annotator agreement results of the two strategies with respect to all emotion classes are shown in Table 1.

The low agreements in topic annotation show the problem in identifying the lexical scopes or spans for each of the emotion topics in a sentence. It is to be mentioned that the agreement in identifying emotion topics in emotional sentences containing single emotion topic is more than the agreement in identifying emotion topics in sentences containing multiple emotion topics. Total 1979 emotion topics are annotated for 1821 target spans as a single target span may contain more than one emotion topic. It is decided to form the gold standard set of 1800 sentences if at least two out of three annotations matches in case of *kappa* or MASI.

4. BASELINE FRAMEWORK

The baseline model is developed based on the *Object* information present in the dependency relations of parsed emotional sentences. Stanford Parser [16], a probabilistic lexicalized parser containing 45 different part of speech (POS) tags of Pen Tree bank is used to get the parsed sentences with dependency relations. The dependency relations are checked for the predicates "*dobj*" so that the *object* related components in the "*dobj*" predicate is considered as the probable candidate for emotion topic. It is observed that only the *dobj* based dependency relations fail to capture the topic spans inscribed in a text. Thus we have turned our focus towards syntax or argument structure acquisition framework.

Table 1. Inter-Annotator Agreement using *kappa* and MASI

Category	A1-A2	A2-A3	A1-A3	Average
Topic Span (kappa)	0.85	0.85	0.84	0.85
Topic Span (MASI)	0.80	0.82	0.81	0.82
Target Span (kappa)	0.97	0.96	0.97	0.97
Target Span (MASI)	0.93	0.94	0.92	0.93

The verb specific syntactic argument structure or subcategorization information plays an important role to identify the emotion topics. The approach is related to some earlier works [10], [5] done for emotion holder and topic identification using VerbNet information. VerbNet [11] associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information such as *Thematic Roles*, semantic predicates, with syntactic frames and selectional restrictions. Verb members in the same VerbNet class share common syntactic frames and thus they are believed to have the same syntactic behaviour. The VerbNet files containing verbs with their possible subcategorization frames and membership information are stored in XML file format. The XML files of VerbNet are pre-processed to build up a general list that contains all member verbs and their available syntactic frames with topic related *thematic* information (e.g. *Topic*, *Theme*, *Event* etc.). The pre-processed list is searched to acquire the syntactic frames of each verb.

On the other hand, the parsed emotional sentences are passed through a rule based *phrasal-head* extraction module to identify the phrase level argument structure of the sentences with respect to their verbs. The extracted *head part* of every phrase from the well-structured bracketed parsed data is considered as the component of the argument structure. The acquired argument structures are compared against the extracted VerbNet frame syntax. If the acquired argument structure matches with any of the extracted VerbNet frame syntaxes that contain topic type *thematic roles*, the phrase associated in the topic slot of the VerbNet frame syntax is mapped to the corresponding phrase in the argument structure. The topic *thematic roles* such as *Topic*, *Theme*, *Event* etc. as specified in the VerbNet are properly tagged in the correct position of the sentences.

For Example 1, the Parse tree, dependency relations, acquired argument structure and VerbNet Frame Syntax for the verb *cry* are as follows.

Parse Tree: (ROOT (S (**NP** (PRP He))(ADVP (RB first))(**VP** (VBD cried)(PRT (RP up)) (**NP** (DT the)(NN toy)(NN car)))(. .)))

Dependency Relations: nsubj(cry-3, He-1), advmod(cry-3, first-2), prt(cry-3, up-4), det(car-7, the-5), nn(car-7, toy-6), ***dobj***(**cry**-3, **car**-7)

Acquired Argument Structure: [NP VP NP]

Simplified Extracted VerbNet Frame Syntax: [<NP value="*Agent*" ></VERB><NP-***topic***>]

The baseline system considers the predicate ***dobj***(**cry**-3, **car**-7) whereas the *phrasal heads* are extracted from the parse tree to form the argument structure. Overall, the baseline system achieves the *F-Score* of 56.75% for topic identification. The components of the dependency relations that are directly linked with the verb forms the target span for the baseline system. It is observed that the matching target spans are obtained for the active simple sentences rather than complex, compound and passive occurrences. 63.09% *F-Score* is achieved by the baseline system for target span identification. The system does not capture all the topical phrases that are related with the emotional expressions and fail to identify the individual scope of multiple topics in the sentences.

5. SUPERVISED FRAMEWORK

A set of standard preprocessing techniques is carried out, *viz.*, *tokenizing, stemming* and *stop word removal*. Tools provided by *Rapidminer's text plugin* (http://rapid-i.com/content/blogcategory/38/69/) were used for these tasks. WordNet's (Miller, 1990) morphological analyzer performed stemming. The training and classification processes for SVM have been carried out by YamCha toolkit and TinySVM-0.07 (http://chasen.org/~taku/software/TinySVM/) respectively. On the other hand, CRF++-0.51 (http://crfpp.sourceforge.net/) package and a Fuzzy Classifier (http://www.autonlab.org/autonweb/10522) are used for training and classification purposes. The selection of best feature set for each classifier has been identified based on the performance of the classifier in terms of *F-Score* on 300 development sentences. Information Gain Based Pruning (IGBP) is carried out to remove the words (e.g. *game*, *gather*, *seem* etc.) that do not play any contributory role in the classification. Rest 1000 and 500 sentences have been respectively used for training and testing of the classifiers.

5.1 Features

Feature plays a crucial role in any machine-learning framework. By manually reviewing the blog data and different language specific characteristics, the following features have been selected for our classification task. The words belonging to annotated target spans and topic spans are tagged with target and topic tags respectively. The aim of our method is to provide the whole sentence and classify the words into target or non-target words as well as emotion topic or non-emotion topic words. The successive words that are tagged with similar classes in a sentence are collected to form the spans of target or topic. Each word associated with the following features is represented as the feature vector.

Structural Similarity (StrucSim)*:* Instead of identifying rhetorical relations, the present task acquires the rhetorical components such as *locus*, *nucleus* and *satellite* from a sentence as these rhetoric clues help in identifying the individual topic spans associated in a target span of the sentences. The topic of an opinion depends on the context in which its associated opinion expression occurs [24]. The part of the text span containing annotated emotional expression is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation markers (,) (!) (?). Frequently used *causal keywords* (*as, because, that, while, whether etc*), *discourse markers* and *causal verbs* are also the useful clues if they are explicitly specified in the text. In Example 3 and Example 4, the separation of *nucleus* and *satellite* is done based on *causal* keyword "*because*".

The identification of *discourse markers* from written text itself is a research area. Hence, the present task aims to identify only the explicit *discourse markers* that are tagged by *conjunctive_*() or *mark_*() type dependency relations of the parsed constituents. The dependency relations containing *conjunctive* markers (*conj_and*(), *conj_or*(), *conj_but*()) are considered for separating *nucleus* from *satellite* if the markers are present in between two successive clauses that are tagged as *S* or *SBAR* in the output of the parse tree. Otherwise, the component contained in *mark_*() type dependency relation is considered as a *discourse marker*.

The list of *causal* verbs is prepared by processing the XML files of English VerbNet [11]. If any VerbNet class file contains any frame with semantic type as *Cause,* we collect the member verbs of that XML class file and termed the member verbs as *causal verbs*. If any clause tagged as *S* or *SBAR* in the parse tree contain any *causal verb*, the clause is considered as the *nucleus* and the rest of the clauses are denoted as *satellite.* The list contains a total number of 250 *causal verbs*. (e.g. "{They *cause* tears to run down my cheeks} [that in turn make me want to fall to my knees.]"). Not only separating the clauses but also the phrases of a single sentence into *nucleus* and *satellite* ("{I feel ***really alone*** right now} [*because* it's Friday.]") have been conducted.

If any word in the annotated emotional expression co-occurs with any word element of the *nucleus* or *satellite* in direct dependency relation*,* the feature is considered as *common similarity* whereas if they occur in transitive dependency relation, the feature is considered as *distinctive similarity*. (e.g. *common similarity* "{I ***enjoyed*** **summer-vacation**}[…]", ***dobj***(enjoy-2, **summer-vacation**-3)). This features aims to separate emotion topics from non-emotion topics as well as to separate the overlapping possibilities of discrete emotion topic spans from non-topical contiguous regions.

Sentiment Similarity (SentiSim)*:* The phrase level chunks extracted from the parsed sentences are used to calculate *Sentiment Similarity*. The *positive* or *negative* valence of each word (*pretty*, *good*) in a chunked phrase is measured using *SentiWordNet* [8]. If any word contained in the chunked phrase is present in the *SentiWordNet,* the corresponding feature entails that the phrase contains *Sentiment Similarity.* Any non-sentiment word (*tournament*) belonging to that chunked phrase is considered as the candidate of target or topic (e.g. "*overall it was a* { ***pretty good*** *tournament* }").

Syntactic Similarity (SynSim)*:* The syntactic similarity feature is identified from the parsed sentences with the help of a context window containing POS level argument structure present between the verb and the emotional expression*.* Only the chunked phrases containing verb, noun and preposition are considered. This feature is identified from the extracted argument structures of the sentences (discussed in Section 4). If any word of a phrase is already defined as a *Theme, Topic* or *Event* in the baseline argument extraction module, the word is considered as emotion

topic and all the words of the chunked phrase is then selected as target words. (e.g. "*He first cried up the **toy car***").

Semantic Similarity (SemSim): The semantic similarity is identified with the help of three WordNet (Miller, 1990) features identified between words and emotional expressions. The features are defined as follows.

WordNet Synonymy: If any word and emotional expression present in any synset of WordNet, emotion topic feature is assigned for that word (e.g. "*I **won** the financial **profit**.*").

WordNet Hypernymy: If any word is defined as *event, topic, theme, subject, issue* or *matter* in its *hypernym* tree, the corresponding word is considered as the probable candidate for target and emotion topic (e.g. "*you at least suffered the **circumstances***").

WordNet SenseID: If any word and the emotional expression both share at least a common *SenseID*, the feature is assigned for that word by considering it as the candidate word for target and emotion topic (e.g. "*He can **enjoy** his **love** with freedom.*")

Dependency Relations (DR): Two types of dependency relations are considered as features. The *direct* dependency is identified based on the simultaneous presence of both the words in the same dependency relation whereas the *transitive* dependencies are verified if the words are connected via one or more intermediate dependency relations (e.g. In the dependency relations of Example 1 in Section 4, the relations between *cry* and *car* are of direct dependency whereas *he* and *toy* is of transitive dependency). Transitive dependency feature aims for identifying the expanded span of target or topic while identifying exact spans is controlled by direct dependency feature.

Emotion Holder (EH): The emotion holder identification module described in [5] is used in the present task to annotate the emotion *holders*. The emotion *holder* information not only aims to identify the focused target span but also contributes in the emotion topic classification technique. If any *direct* or *transitive* dependency relation holds between any word, emotion *holder* and emotional expression, the word is considered as the candidate of the target and topic span. (Example 1, the relations are *nsubj(cry-3, **He**-1)*, *dobj(cry-3, **car**-7), nn(**car**-7, **toy**-6)* that contain a chain from emotion holder "*he*" to topic "toy car".

Named Entity (NE): Each of the sentences is passed through a Stanford Named Entity Recognizer (http://nlp.stanford.edu/software/CRF-NER.shtml) for identifying the named entities. If any word is tagged as a named entity and present in *satellite* and not tagged with *Emotion Holder* (*EH*) feature, the word is selected as a potential candidate for target and topic (e.g. "{*I forgot*} [*how demeaning **BME** classes are.*]").
Different unigram and bi-gram context features (word and POS tag level) and their combinations were generated from the training corpus. The features of topic and target also contribute mutually to train the classifier.

6. EVALUATION

The importance of incorporating different features is identified based on their performance on the development set of 300 sentences. The best feature set for each of the classifiers is obtained by parameter estimation and threshold determination (we varied the threshold for the classification). The *F-scores* of the important features and their combinations for each of the classifiers are shown in Table 2. All machine learning methods were tested via 10-fold cross validation.

CRF also assigns the tag sequence to the word sequence that is responsible for target and topic span. But, CRF and SVM suffer from inadmissible tag sequence (target and topic tags occur in alternative sequence rather than consecutive) and label bias problem (uneven distribution between emotion and non-emotion target and topic tags). To eliminate inadmissible sequences, we define a transition probability between word classes P ($ci \mid cj$) to be equal to 1 if the sequence is admissible, and 0 otherwise. The probability of the classes c_1,….., cn is assigned to the words in a sentence "s" in an topic or target class D is defined as follows:

$$P(c_1 \ldots, c_n \mid s, D) = \prod_{i=1}^{n} P(c_i|s,D) * P(c_i|c_{i-1})$$

where, ($ci \mid s$, D) is determined by the CRF classifier.

One solution to the unbalanced class distribution or label bias problem is to split the 'non-emotion target/topic' (emo_ntrl_target/topic) classes into several subclasses effectively. That is, given a POS tagset *POS*, we generate new emotion target/topic classes, 'emo_ntrl_target/topic-C'|C $\in$ *POS*. We have forty-five (45) subclasses in the English POS tagset, which correspond, to non-emotion target and topic regions such as 'emo_ntrl_target/topic-NN'(common noun), 'emo_ntrl_target/topic-VFM' (verb finite main) etc.

The Fuzzy Classifier (FC) assigns the topic or target tag to the words that occupy successive positions. Tuning different control parameters e.g., *min_ball_width* parameter is fixed as 0.0103 (its reference threshold is 1e-007), the best achievable feature set and results are obtained accordingly. The fuzzy technique helps to identify all the emotion topic or target words in a sentence associated with different distinguishable weightings.

But, SVM identifies emotion topics and targets significantly better than other two classifiers. We have used both *one vs. rest* and *pair wise* multi-class decision methods for extending the binary classification task into the multi-class classification. In SVM, various degrees of the *polynomial kernel function* have been also used. During CRF and SVM-based training phase, the current token word with three previous and three next words and their corresponding POS were selected as context feature for that word.

6.1 Feature Analysis

It is observed that the features like, *Dependency Relations*, *Emotion Holder*, *Named Entity* are performed similarly for all the classifiers. *Emotion Holder* feature performs well in case of the sentences containing multiple clauses and long distance dependencies. *Structural Similarity* feature containing rhetoric knowledge of the sentences along with *Syntactic Similarity* helps in identifying the words that are responsible for the target span.

CRF and Fuzzy Classifier perform better in case of *Structural Similarity* and *Syntactic Similarity* whereas SVM additionally performs well in case of *Emotion Holder*, *Sentiment Similarity* and *Semantic Similarity*. The combination of *Structural Similarity, Sentiment Similarity* and *Semantic Similarity* reasonably improves the performance of the classifiers. *Transitive dependency* performs better in target span identification rather than topic span whereas *direct dependency* shows the improvement in topic span. The combined features that depend on emotional expressions (e.g., *Structural Similarity* and *Semantic Similarity*) and *Emotion Holder* significantly improve the performance of the classifiers. Only some important features and their combined performance for each of the classifiers are shown in Table 2. For lack of space, the figures are only shown for emotion topic identification but it is observed that the target identification always produces comparatively better results for all the classifiers. It is observed that all three supervised classifiers outperform the baseline system significantly.

Table 2. *F-Scores* (in %) of different features on the development set for three classifiers

Feature(s)	CRF	SVM	FC
Emotion Holder (*EH*)	23.03	24.02	21.58
Named Entity (*NE*)	22.10	22.75	25.19
StrucSim	46.92	44.67	45.77
SentiSim	20.04	24.34	21.09
SynSim	43.78	42.33	46.12
SemSim	37.76	40.08	33.54
EH+StrucSim	54.82	56.90	53.11
EH+SentiSim	44.63	47.05	42.94
EH+SynSim	53.18	52.73	52.22
EH+SemSim	51.04	52.88	50.79
EH+DR+StrucSim+ SentiSim	57.70	61.07	60.26
EH+DR+StrucSim+ SemSim	56.87	57.03	57.47
EH+DR+StrucSim+ SentiSim+SemSim	58.22	59.45	58.10
EH+DR+SynSim+ StrucSim+SentiSim	58.76	60.92	60.88
EH+NE+DR+ FourSims	59.98	61.79	61.15
EH+NE+FourSims+ DR+Context Features	61.55	63.21	62.32

6.2 Information Gain Based Pruning

The importance of incorporating the attributes/features is examined through Information Gain measure. This decision technique is used to measure the importance of an attribute/feature (X) with respect to the class attribute (Y). Formally, information gain of a feature X with respect to a class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X.

$$InfoGain(Y;X) = entropy(Y) - entropy(Y|X)$$

where X and Y are discrete variables taking values {x1, x2,....,xm} and {y1,y2,....,yn} respectively. The *Entropy* (Y) is defined as:

$$\text{Entropy}(Y) = -\sum_{i=1}^{n} P(Y = y_i)\ log_2\ P(Y = y_i)$$

The conditional entropy of Y given X is defined as

$$\text{Entropy}(Y|X) = -\sum_{i=1}^{m} P(X = x_j)\text{Entropy}(Y|X = x_j)$$

Features with high Information Gain reduce the uncertainty about the class to the maximum. In our experiment on the development set, all the words except non-emotional words (e.g. *seem*, *gather*) achieve high Information Gain threshold (>50%). Information Gain Based Pruning (IGBP) shows significant improvement in all the classifiers.

Table 3. *F-Scores* (in %) of three classifiers before and after Information Gain Based Pruning (IGBP)

Category	CRF	SVM	FC
Topic Span	61.55	63.21	62.32
Topic Span (IGBP)	64.43	65.56	64.14
Target Span	72.34	73.67	72.08
Target Span (IGBP)	74.22	75.19	73.67

6.3 Multi-Engine with Voting

A multi-engine approach has been proposed in order to achieve better performance for emotion topic and target identification. We have considered the CRF, SVM and Fuzzy Classifier based systems as these yielded the overall *F-Score* values that are close to each other. A close investigation to the evaluation results on the development set have shown that a large number of emotion topics and targets, classified wrongly by any classifier, are correctly classified by another classifier. This situation points to the development of a multi-engine supervised system using voting scheme.

The main philosophy behind this is to give importance to various classifiers in order to determine the final emotion topic tag and target tag to a particular word. Appropriate voting technique may be effective to develop any multi-engine system. The three classifiers have been evaluated separately. But before applying weighted voting, we need to decide the weights to be given to each individual classifier. The following two types of weighting methods are applied for voting.

Majority Voting (M*voting*): We have assigned the same voting weight to all the systems. The multi-engine (ME) system selects the classifications, which are proposed by majority of the models. If the three outputs are different, then the output of the SVM system is selected as this system yielded the highest performance between the CRF, SVM and Fuzzy Classifiers for the given development datasets.

Cross Validation Total F-Score Values (CVTFV): The training data is divided into N portions. We train each system by using N−1 portions, and then evaluate them on the remaining portion. This is repeated N times. In each of the iterations, we have evaluated the individual system. At the end, we get N number of *F-Score* values for each of the system. Final voting weight for a system is given by the average of these N number of *F-Score* values. Here, we set the value of N to be 10. We have defined total *F-Score* as follows: we have assigned the overall average *F-Score* of any classifier as the weight for it. For example, the CRF, SVM and Fuzzy classifiers will consider their overall average *F-Score* value as their corresponding weight. The classification of any word *w* is determined by the following function:

$$C(w) = \sum_{i=1}^{n} a_i * Out(Model_i)$$

where, *C*(*w*) is the voted output tag to be assigned to the word *w*, *ai* is the overall average *F-Score* of the i^{th} system (CRF/SVM/Fuzzy) and *Out*(*Modeli*) is the output tag (one of the topic/target tag or non-topic/non-target tag) predicted by the i^{th} system for the word *w*. Finally, the tag with the highest coefficient value (i.e., the largest value of *ai*) is selected as the final output of the voted system.

6.4 Results

The results of the baseline system, combination of classifiers and multi-engine classifiers with and without two types of voting methods on 500 gold standard test sentences are shown in Table 4. It is observed that the performance of topic span identification is less in comparison with target span. But, the *recall* of the system is better than the *precision* for topic and target span identification as the system is strong enough to identify one or more potential emotion topics and targets from the sentences but suffers in determining the exactly matched scopes of the emotion topics to some extent.

The error occurs mostly for metaphoric usages, unstructured sentences (e.g. "*Really starting to lose it.*") and the sentences containing typographic errors (e.g. "*she's feeling very goooood about herself.*"). The system fails to capture the scopes of some emotional topic words (*exam results*) from the non-emotional topics (*cricket matches, new films*) e.g. "*The success was aimed for the **exam results**, except **cricket matches** or **new films**.*"

Table 4. Test set *Precision (P), Recall (R) and F-Score (F)* of baseline, and supervised multi-engine classifiers

Classifiers	Target			Topic		
	P	R	F	P	R	F
Baseline	.61	.67	.63	.52	.60	.56
CRF+SVM	.75	.85	.80	.62	.75	.67
CRF+FC	.76	.80	.78	.62	.71	.66
SVM+FC	.87	.80	.83	.64	.70	.67
ME	.86	.90	.88	.65	.73	.69
ME+ M*voting*	.87	.93	.90	.66	.74	.70
ME-CVTFV	.88	.92	.90	.65	.75	.70

7. CONCLUSION

We have built the emotion topic and target identification system that makes use of lexical (word, POS), *syntactic, semantic, rhetoric* and *sentiment knowledge* provided by a set of rich lexical resources in the form of *WordNet*, *SentiWordNet*, *dependency parser*, *VerbNet*. The performance of the system on the blog domain is satisfactory except for low precision (65.49%) obtained for metaphoric words or some non-emotional topics depicted as emotional topics. Investigation on causes of errors points to the classic limitations of long distance dependency among the relevant emotional topic constituents in a sentence.

8. REFERENCES

[1] Aman, S. and Szpakowicz, S..2007. Identifying Expressions of Emotion in Text. *V. Matoušek and P. Mautner (Eds.):TSD 2007, LNAI 4629*, 196–205.

[2] Azar M. 1999. Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. *Argumentation*, vol (13), pp. 97–114, 1999.

[3] Choi, F. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*.

[4] Cohen, J. 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, vol. 20, pp. 37–46.

[5] Das D. and Bandyopadhyay, S. 2010. Emotion Holder for Emotional Verbs – The role of Subject and Syntax. *CICLing-2010, A. Gelbukh (Ed.), LNCS 6008*, pp. 385-393, Romania

[6] Das. D and Bandyopadhyay, S. 2010. Sentence Level Emotion Tagging on Blog and News Corpora. *Journal of Intelligent System (JIS)*, vol. 19(2), pp. 125-134.

[7] Ekman, P.1993. Facial expression and emotion. *American Psychologist*, vol. 48(4) 384–392.

[8] Esuli, A and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.*LREC-06*.

[9] Joachims, T.1998. Text Categorization with Support Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML)*,137-142

[10] Kim, S. and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*.

[11] Kipper-Schuler, K.2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia,PA

[12] Kobayashi, N., K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. 2004. Collecting evaluative expressions for opinion extraction. *IJCNLP*.

[13] Liu, T, Moore, A., Gray , A. and Yang, K. 2004. An Investigation of Practical Approximate Nearest Neighbor Algorithms. *LIU-NIPS 04*.

[14] Mann, W. and Thompson, S. 1987. Rhetorical Structure Theory: Description and Construction of Text Structure, In *G. Kempen (ed.), Natural Language Generation, Martinus Nijhoff*, The Hague, pp. 85–96.

[15] Mann, W. C. and Thompson, S. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *TEXT* vol. 8, 243–281.

[16] Marneffe Marie-Catherine de, MacCartney, B. and Manning, C. D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *5th International Conference on Language Resources and Evaluation*.

[17] McCallum, A., Pereira, F. and Lafferty, J. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data. *ISBN*, pp. 282 – 289

[18] Miller, G. A. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235–312

[19] Nomoto, T. and Matsumoto, Y. 1996. Exploiting Text Structure for Topic Identification. In *Proceedings of the 4th Workshop on Very Large Corpora*, pp.101-112.

[20] Passonneau, R. 2004. Computing reliability for coreference annotation, In *Proc. International Conf. on Language Resources and Evaluation*.

[21] Passonneau, R.J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation, In *Proc. 5th International Conference on Language Resources and Evaluation*.

[22] Popescu, A. and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

[23] Stoyanov, V. and C. Cardie. 2008. Annotating topics of opinions. In *Proceedings of LREC*

[24] Stoyanov, V. and Cardie, C. 2008. Topic Identification for Fine-Grained Opinion Analysis. *Coling 2008*, pp. 817–824

[25] Wiebe, J., Wilson, T. and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 1(2).

[26] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of ICDM*.