Sentiment Analysis in Twitter with Lightweight Discourse Analysis

Subhabrata Mukherjee[†], *Pushpak Bhattacharyya*[‡]
[†]IBM India Research Lab
[‡]Dept. of Computer Science and Engineering, IIT Bombay
subhabmu@in.ibm.com, pb@cse.iitb.ac.in

ABSTRACT

We propose a lightweight method for using discourse relations for polarity detection of *tweets*. This method is targeted towards the web-based applications that deal with *noisy, unstructured* text, like the *tweets*, and cannot afford to use heavy linguistic resources like *parsing* due to frequent failure of the parsers to handle noisy data. Most of the works in micro-blogs, like *Twitter*, use a bag-of-words model that ignores the discourse particles like *but, since, although etc*. In this work, we show how the discourse relations like the *connectives* and *conditionals* can be used to incorporate discourse information in any bag-of-words model, to improve sentiment classification accuracy. We also probe the influence of the semantic operators like *modals* and *negations* on the discourse relations that affect the sentiment of a sentence. Discourse relations and corresponding rules are identified with minimal processing - just a list look up. We first give a linguistic description of the various discourse relations which leads to conditions in rules and features in SVM. We show that our discourse-based bag-of-words model performs well in a noisy medium (*Twitter*), where it performs better than an existing Twitter-based application. Furthermore, we show that our approach is beneficial to structured reviews as well, where we achieve a better accuracy than a state-of-the-art system in the *travel review* domain. Our system compares favorably with the state-of-the-art systems and has the additional attractiveness of being less resource intensive.

KEYWORDS : Sentiment Analysis, Discourse, Twitter, Connectives, Micro-blogs

Proceedings of COLING 2012: Technical Papers, pages 1847–1864,
COLING 2012, Mumbai, December 2012.

1 INTRODUCTION

An essential phenomenon in natural language processing is the use of discourse relations to establish a coherent relation, linking phrases and clauses in a text. The presence of linguistic constructs like *connectives, modals, conditionals* and *negation* can alter sentiment at the sentence level as well as the clausal or phrasal level. Consider the example, "*@user share 'em! i'm* **quite excited** *about Tintin, despite* **not really liking** *original comics. Probably because Joe Cornish had a hand in."* The overall sentiment of this example is *positive*, although there is equal number of positive and negative words. This is due to the connective *despite* which gives more weight to the previous discourse segment. Any bag-of-words model would be unable to classify this sentence without considering the discourse marker. Consider another example, "*Think i'll stay with the whole 'sci-fi'* **shit***. but this time...a* **classic** *movie."* The overall sentiment is again *positive* due to the connective *but*, which gives more weight to the following discourse segment. Thus it is of utmost importance to capture all these phenomena in a computational model.

Traditional works in *discourse analysis* use a discourse parser (Marcu 2000; Zirn *et al.*, 2011, Wellner *et al.*; 2007; Pitler *et al.*, 2009; Elwell *et al.*, 2008) or a dependency parser (Vincent *et al.*, 2006). Many of these works and some other works in discourse (Taboada *et al.*, 2008; Zhou et al., 2011) build on the Rhetorical Structure Theory (RTS) proposed by Mann *et al.* (1988) which tries to identify the relations between the nucleus and satellite in the sentence.

Most of these theories are well-founded for *structured text*, and *structured* discourse annotated corpora are available to train the models. However, using these methods for micro-blog discourse analysis pose some fundamental difficulties:

1. Micro-blogs, like *Twitter*, do not have any restriction on the form and content of the user posts. Users do not use formal language to communicate in the micro-blogs. As a result, there are abundant *spelling mistakes, abbreviations, slangs, discontinuities* and *grammatical errors*. This can be observed in the given examples from real-life *tweets*. The errors cause natural language processing tools like *parsers* and *taggers* to fail frequently (Dey *et al.*, 2009). As the tools are generally trained on structured text, they are unable to handle the noisy and unstructured text in this medium. Hence most of the discourse-based methods, based on RST or parsing of some form, will be unable to perform very well in micro-blog data.

2. The web-based applications require a fast response time. Using a heavy linguistic resource like *parsing* increases the processing time and slows down the application.

Most of the works in micro-blogs, like *Twitter*, (Alec *et al.*, 2009; Read *et al.*, 2005; Pak *et al.*, 2010; Gonzalez *et al.*, 2011) use a bag-of-words model with features like part-of-speech information, unigrams, bigrams *etc.* along with other domain-specific, specialized features like *emoticons, hashtags etc.* In *most* of these works, the *connectives, modals* and *conditionals* are simply ignored as stop words during feature vector creation. Hence, the discourse information that can be harnessed from these elements is completely discarded. In this work, we show how

the *connectives*, *modals*, *conditionals* and *negation* based discourse information can be incorporated in a bag-of-words model to give better sentiment classification accuracy.

The roadmap for the rest of the paper is as follows: Related work is presented in Section 2. Section 3 presents a comprehensive view of the different discourse relations. Section 4 studies the effect of these relations on sentiment analysis and identifies the critical ones. Section 5 discusses the influence of some semantic operators on discourse relations for sentiment analysis. We develop techniques for using the discourse relations to create feature vectors in Section 6. Lexicon based classification and supervised classification systems are presented in Section 7 to classify the feature vectors. Experimental results are presented in Section 8, where we use *three* different datasets, from the *Twitter* and *Travel review* domain, to validate our claim. The results are discussed in Section 9, followed by conclusions.

2 RELATED WORK

2.1 Discourse Based Works

Maru (2000) discussed probabilistic models for identifying elementary discourse units at clausal level and generating trees at the sentence level, using lexical and syntactic information from discourse-annotated corpus of RST. Wellner *et al.* (2007) considers the problem of automatically identifying arguments of discourse connectives in the PDTB. They model the problem as a predicate-argument identification where the predicates are discourse connectives and arguments serve as anchors for discourse segments. Wolf *et al.* (2005) presents a set of discourse structure relations and ways to code or represent them. The relations were based on Hobbs (1985). They report a method for annotating discourse coherent structures and found different kinds of crossed dependencies.

In the work, *Contextual Valence Shifters* (Polanyi *et al.*, 2004), the authors investigate the effect of *intensifiers*, *negatives*, *modals* and *connectors* that changes the prior polarity or valence of the words and brings out a new meaning or perspective. They also talk about pre-suppositional items and irony and present a simple weighting scheme to deal with them.

Somasundaran *et al*. (2009) and Asher *et al.* (2008) discuss some discourse-based supervised and unsupervised approaches to opinion analysis. Zhou *et al.* (2011) present an approach to identify discourse relations as identified by RST. Instead of depending on cue-phrase based methods to identify discourse relations, they leverage it to adopt an unsupervised approach that would generate semantic sequential representations (SSRs) without cue phrases.

Taboada *et al*. (2008) leverage discourse to identify relevant sentences in the text for sentiment analysis. However, they narrow their focus to adjectives alone in the relevant portions of the text while ignoring the remaining parts of speech of the text.

Most of these discourse based works make use of a *discourse parser* or a *dependency parser* to identify the scope of the discourse relations and the opinion frames. As said before, the parsers fare poorly in the presence of noisy text like *ungrammatical sentences* and *spelling mistakes* (Dey *et al.*, 2009). In addition, the use of parsing slows down any real-time interactive system due to

increased processing time. For this reason, the micro-blog applications mostly build on a bag-of-words model.

2.2 Twitter Based Works

Twitter is a micro-blogging website and ranks second amongst the present social media websites (Prelovac, 2010). A micro-blog allows users to exchange small elements of content such as short sentences, individual pages, or video links. Alec *et al.* (2009) provide one of the first studies on sentiment analysis on micro-blogging websites. Barbosa *et al.* (2010) and Bermingham *et al.* (2010) both cite noisy data as one of the biggest hurdles in analyzing text in such media.

Alec *et al.* (2009) describe a distant supervision-based approach for sentiment classification. They use *hashtags* in tweets to create training data and implement a multi-class classifier with topic-dependent clusters. "*The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages*"[1].

Barbosa *et al.* (2010) propose an approach to sentiment analysis in Twitter using POS-tagged n-gram features and some Twitter specific features like *hashtags.* Joshi *et al.* (2011) propose a rule-based system, *C-Feel-It*, which classifies a tweet as positive or negative based on the opinion words present in it. It uses sentiment lexicons for classification and twitter-specific features like *emoticons, slangs, hashtags etc.* Use of *emoticons* is common in social media and micro-blogging sites, where the users express their sentiment in the form of accepted symbols. Example: ☺ *(happy)*, ☹ *(sad).*

Read *et al.*, (2005) and Pak *et al.* (2010) propose a method to automatically create a training corpus using micro-blog specific features like *emoticons*, which is subsequently used to train a classifier. Gonzalez *et al.* (2011) discuss an approach to identify sarcasm in tweets. To create a corpus of *sarcastic, positive* and *negative* tweets, they rely on the user provided information in the form of *hashtags.* They claim that the author is the best judge for determining whether the tweet is sarcastic or not, which is indicated by the *hashtags* used by the author in the post.

Our work builds on the discourse-related works of Polanyi *et al.* (2004), Wolf *et al.* (2005) and Taboada *et al.* (2008) and carries the idea further in the sentiment analysis of micro-blogs. We exploit the various features discussed in the Twitter specific works to develop a bag-of-words model, in which the discourse features are incorporated to give better sentiment classification accuracy.

We evaluate our system on three datasets using lexicon-based classification as well as a supervised classifier (SVM). We use a manually labeled tweet set of 8,507 tweets and an automatically annotated set of 15,204 tweets using *hashtags*, to establish our claim. We, further, use a dataset from the *travel review* domain by Balamurali *et al.* (2011) to show that our method is beneficial to structured reviews as well, which is indicated by the improved classification accuracy over the compared work.

[1] https://support.twitter.com/articles/49309

3 CATEGORIZATION OF DISCOURSE RELATIONS

"An important component of language comprehension in most natural language contexts involves connecting clauses and phrases together in order to establish a coherent discourse" (Wolf *et al.,* 2004). A coherently structured discourse is a collection of sentences having some relation with each other. A coherent relation reflects how different discourse segments interact (Hobbs 1985; Marcu 2000; Webber *et al.,* 1999). Discourse segments are non-overlapping spans of text. The interaction relations between discourse segments may be of various forms as listed in *Table 1*.

Coherence Relations	Conjunctions
Cause-effect	because; and so
Violated Expectations	although; but; while
Condition	if…(then); as long as; while
Similarity	and; (and) similarly
Contrast	by contrast; but
Temporal Sequence	(and) then; first, second, … before; after; while
Attribution	according to …; …said; claim that …; maintain that …; stated
Example	for example; for instance
Elaboration	also; furthermore; in addition; note (furthermore) that; (for , in, with) which; who; (for, in, on, against, with) whom
Generalization	in general

Table 1: Contentful Conjunctions used to illustrate Coherence Relations (Wolf *et al.* 2005)

Our work, almost entirely, rests on this platform, where we identify the relations from *Table 1*, which can affect the analysis of opinions most in a discourse segment. *Table 2* provides some examples, taken from *Twitter*, to illustrate the effect of conjunctions in discourse analysis. These examples are similar to the ones in Polanyi *et al.* (2004) and Taboada *et al.* (2008). The words in *bold* connect the discourse segment in brackets. The relations are broadly classified in ten categories in *Table 2*.

4 DISCOURSE RELATIONS CRITICAL FOR SENTIMENT ANALYSIS

Not all of the discourse relations are significant from the point of view of sentiment analysis. This section examines the role of the critical ones in SA.

1. Violated Expectations and Contrast - A simple bag-of-words model will classify *Example 2* (*Table 2*) as neutral. This is because it has one positive term *excited* and one negative phrase *not*

really liking. However, it represents a positive emotion of the opinion holder, due to the segment after the connective *despite*. In *Example 5*, *brightened* is positive and *poorly* is negative. Hence the overall polarity is un-decided. But it should have been *positive*, since the segment following *but* gives the overall impression of the opinion-holder which is positive.

Violating expectation conjunctions oppose or refute the neighboring discourse segment. We further categorize them into the following *two* sub-categories: *Conj_Fol* and *Conj_Prev.*

Conj_Fol is the set of conjunctions that give more importance to the discourse segment that follows them. *Conj_Prev* is the set of conjunctions that give more importance to the previous discourse segment.

Thus, in *Example 5* of *Table 2*, the discourse segment following *but* should be given more weight. In *Example 2*, the discourse segment preceding *despite* should be given more weight.

1. Cause-effect: *(YES! I hope she goes with Chris)* **so** *(I can freak out like I did with Emmy Awards.)*

2. Violated Expectations: *(i'm quite excited about Tintin),* **despite** *(not really liking original comics.)*

3. Condition: **If** ***(****MicroMax improved its battery life), (it wud hv been a gr8 product).*

4. Similarity: *(I lyk Nokia)* **and** *(Samsung as well).*

5. Contrast: *(my daughter is off school very poorly),* **but** *(brightened up when we saw you on gmtv today).*

6. Temporal Sequence: *(The film got boring)* **after a while**.

7. Attribution: ***(****Parliament is a sausage-machine: the world)* **according to** *(Kenneth Clarke).*

8. Example: *(Dhoni made so many mistakes…)* **for instance**, *(he shud've let Ishant bowl wn he was peaking).*

9. Elaboration: **In addition** *(to the worthless direction), (the story lacked depth too).*

10. Generalization**: In general**,*(movies made under the RGV banner) (are not worth a penny).*

Table 2: Examples of Discourse Coherent Relations

2. Conclusive or Inferential Conjunctions - These are the set of conjunctions, *Conj_infer*, that tend to draw a conclusion or inference. Hence, the discourse segment following them (*subsequently* in *Example 11*) should be given more weight.

Example 11: @User *I was nt much satisfied with ur so-called gud phone and* **subsequently** *decided to reject it.*

3. Conditionals - In *Example 3 (Table 2),* both *improve* and *gr8* represent a *high degree* of positive sentiment. But the presence of *if* tones down the final polarity as it introduces a hypothetical situation in the context. The *if…then…else* constructs depict situations which may or may not happen subject to certain conditions.

In our work, the polarity of the discourse segment in a conditional statement is toned down, in *lexicon-based classification*. In *supervised classifiers*, the conditionals are marked as features. Such statements are not completely ignored as objective, as they bear some sentiment polarity.

4. Other Discourse Relations - Sentences under *Cause-Effect, Similarity, Temporal Sequence, Attribution, Example, Generalization and Elaboration,* provide no contrasting, conflicting or hypothetical information. They can be handled by taking a simple majority valence of the individual terms, as in a bag-of-words model.

Table 3 lists all the essential discourse relations discussed in this section. The relations have been compiled from Hobbs (1985), Polanyi *et al.* (2004) and Taboada *et al.* (2008).

5 SEMANTIC OPERATORS INFLUENCING DISCOURSE RELATIONS

There are some semantic operators that influence the discourse relations connecting the phrases. In the sentence *You may like the movie despite the bad reviews,* the connective *despite* gives more weightage to the discourse segment preceding it and hence, *like* is weighed up. But the uncertainty resulting from *may* that pulls down the weightage is completely ignored. Similarly, in the sentence *He put a lot of effort for the finals, but still it was not good enough to win the match,* the connective *but* upweights *good* and *win* ignoring the negation operator *not*. In this section, we consider the semantic operators like the *modals* and *negation*, ignoring which results in an incorrect interpretation of the sentiment.

Relations	Attributes
Conj_Fol	*but, however, nevertheless, otherwise, yet, still, nonetheless*
Conj_Prev	*till, until, despite, in spite, though, although*
Conj_Infer	*therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*
Conditionals	*If*
Strong_Mod	*might, could, can, would, may*
Weak_Mod	*should, ought to, need not, shall, will, must*
Neg	*not, neither, never, no, nor*

Table 3: Discourse Relations and Semantic Operators Essential for Sentiment Analysis

1. Modals - Events that have happened, events that are happening or events that are certain to occur are called *realis events.* Events that have possibly occurred or have some probability to occur in the distant future are called *irrealis events.* Thus, it is important to distinguish between

real situations and *hypothetical* ones. The modals (*might, may, could, should, would etc.*) depict *irrealis events. Example 3 (Table 2)* does not necessarily talk of MicroMax being *great*, but talks of its *possibility* of being *great* subject to certain conditions (its *battery life*). These constructs cannot be handled by taking a simple majority valence of terms.

We further divide the modals into *two* sub-categories: *Strong_Mod* and *Weak_Mod*.

Strong_Mod is the set of modals that express a higher degree of uncertainty in any situation.

Weak_Mod is the set of modals that express lesser degree of uncertainty and more emphasis on certain events or situations.

In our work, the polarity of the discourse segment neighboring a *strong modal* is toned down in *lexicon-based classification*, similar to the *conditionals*, as it expresses a higher degree of uncertainty. In *supervised classifiers*, the modals are marked as features.

Example 12 (Strong Modals): *Unless I missed the announcement their God is now featured on postage stamps, it* **might** *be a hard sell.*

He **may** *be a rising star.*

Example 13 (Weak Modals): *G.E 12* **must** *be the* most *deadly General Election for politicians ever.*

Our civil service **should** *work without TD interference.*

2. Negation - The negation operator (Example: *not, neither, nor, nothing etc.*) inverts the sentiment of the word following it. The usual way of handling negation in SA is to consider a window of size *n* (typically 3-5) and reverse the polarity of all the words in the window. In *Example 14*, negating all the words in a window of size 5 reverses the polarity of "*like*" for Samsung as well; this is undesirable. We consider a negation window of size 5 and reverse all the words in the window, till either the window size exceeds or a *violating expectation* (*or a contrast*) conjunction is encountered. Hence, the scope of reversing polarity is limited to the appearance of *but* in the given example.

Example 14 (Negation): *I do not like Nokia but I like Samsung.*

6 ALGORITHM TO HARNESS DISCOURSE INFORMATION

The discourse relations and the semantic operators (identified in *Section 4* and *Section 5*) are used to create a feature vector, according to the following algorithm.

Let a user post R consist of 'm' sentences s_i $(i=1…m)$, where each s_i consist of n_i words w_{ij} $(i=1…m,\ j=1…n_i)$. Let f_{ij} be the weight of the word w_{ij} in sentence s_i, initialized to 1. Let A be the set of all discourse relations in *Table 3*. Let $flip_{ij}$ be a variable which indicates whether the polarity of w_{ij} should be flipped or not. Let hyp_{ij} be a variable which indicates the presence of a *conditional or a strong modal* in s_i.

for $i=1…m$
  for $j=1…n_i$
    f_{ij}=1;
    hyp_{ij}=0;
1. **if** $w_{ij} \in$ Conditionals *or* $w_{ij} \in$ Strong_Mod
      hyp_{ij}=1;
    end if
  end for

  for $j=1…n_i$
    $flip_{ij}$=1;
2. **if** $w_{ij} \in$Conj_Fol *or* $w_{ij} \in$ Conj_Infer
      for $k=j+1…n_i$ *and* $w_{ij} \notin A$
        f_{ik}+=1;
      end for
    end if
3. **else if** $w_{ij} \in$ Conj_Prev
      for $k=1…j-1$ && $w_{ij} \notin A$
        f_{ik}+=1;
      end for
    end if
4. **else if** $w_{ij} \in$ *Neg*
      for $k=1…$*Neg_Window and* $w_{ik} \notin$ *Conj_Prev and* $w_{ik} \notin$ *Conj_Fol*
        $flip_{i,j+k}$=-1;
      end for
    end if
  end for
end for
Input: Review *R*
Output: w_{ij}, f_{ij}, $flip_{ij}$, hyp_{ij}

Algorithm 1: Using the Discourse Relations to Create the Feature Vector

In *Step 1*, we mark all the *conditionals* and *strong modals* which are handled separately by the lexicon-based classifier and the supervised classifier.

In *Step 2* and *Step 3*, the weight of any word appearing before *Conj_Prev* and after *Conj_Fol* or *Conj_Infer* is incremented by 1.

In *Step 4*, the polarity of all the words appearing within a window (*Neg_Window* is taken as 5), from the occurrence of a negation operator and before the occurrence of a *violating expectation* conjunction, are reversed.

Finally, we get the feature vector $\{w_{ij}, f_{ij}, flip_{ij}$ and $hyp_{ij}\}$ for all the words in the review.

Here, the assumption is that the effect of any conjunction is restricted to continuous spans of text till another conjunction or the sentence boundary.

7 FEATURE VECTOR CLASSIFICATION

We devised a lexicon based system as well as a supervised system for feature vector classification.

7.1 Lexicon Based Classification

The Bing Liu sentiment lexicon (Hu *et al.*, 2004) is used to find the polarity $pol(w_{ij})$ of a word w_{ij}. It contains around 6800 words which are manually polarity labeled. The polarity of the review (*pos* or *neg*) is given by,

$$sign(\sum_{i=1}^{m}\sum_{j=1}^{n_i} f_{ij} \times flip_{ij} \times p(w_{ij}))$$

$$where\ p(w_{ij}) = pol(w_{ij})\ \ if\ hyp_{ij} = 0$$
$$= \frac{pol(w_{ij})}{2} \quad if\ hyp_{ij} = 1 \qquad \ldots\ Equation\ 1$$

Equation 1 finds the weighted, signed polarity of a review. The polarity of each word, $pol(w_{ij})$ *being +1 or -1*, is multiplied with its discourse-weight f_{ij} (assigned by *Algorithm 1*), and all the weighted polarities are added. $Flip_{ij}$ indicates if the polarity of w_{ij} is to be negated.

In case there is any *conditional* or *strong modal* in the sentence (indicated by $hyp_{ij} = 1$), then the polarity of every word in the sentence is toned down, by considering half of its assigned polarity ($\frac{+1}{2}$ *or* $\frac{-1}{2}$).

Thus, if *good* occurs in the user post twice, it will contribute a polarity of $+1 \times 2 = +2$ to the overall review polarity, *if* $hyp_{ij} = 0$. In the presence of a *strong modal* or *conditional*, it will contribute a polarity of $\frac{+1}{2} \times 2 = +1$.

All the *stop words*, *discourse connectives* and *modals* are ignored during the classification phase, as they have a zero polarity in the lexicon.

7.2 Supervised Classification

The Support Vector Machines have been found to outperform other classifiers, like *Naïve Bayes* and *Maximum Entropy*, in sentiment classification (Pang *et al.*, 2002). Hence, in our work, SVM's are used to classify the set of feature vectors $\{flip_{ij}, w_{ij}, f_{ij}$ and $hyp_{ij}\}$.

Features used in the Support Vector Machines:

N-grams – *Unigrams* along with *Bigrams* are used.

Stop Words – All the stop words (like *a, an, the, is etc.*) and discourse connectives are discarded.

Feature Weight – In the *baseline bag-of-words* model, the feature weight has been taken as the feature frequency *i.e.* the number of times the unigram or bigram appears in the text. In the *discourse-based bag-of-words* model, the *discourse-weighted frequency* of a word is considered. *Algorithm 1* assigns a weight f_{ij} to every occurrence of a word w_{ij} in the post. If the same word occurs multiple times, the weights from its multiple occurrences will be added and used as a feature weight for the word.

Modal and Conditional Indicator – This is a Boolean variable which indicates the presence of a strong modal or conditional in the sentence (i.e. $hyp_{ij}=1$).

Stemming – All the words are stemmed in the text so that "*acting*" and "*action*" have a single entry corresponding to "*act*".

Negation – A Boolean variable ($flip_{ij}$) is appended to each word (w_{ij}) to indicate whether it is to be negated or not (i.e. $flip_{ij}=1$ or $flip_{ij}=0$).

Emoticons – An emoticon dictionary is used to map each emoticon to a *positive* or *negative* class. Subsequently, the emoticon class information is used in place of the emoticon.

Part-of-Speech Information – The part-of-speech information is also used with a word.

Feature Space - In the *lexeme feature space* individual words are used as features; whereas in the *sense space*, the sense of the word (synset-id) is used in place of the word. A synset is a set of synonyms that collectively disambiguate each other to give a unique sense to the set, identifiable by the *synset-id*. This is beneficial in distinguishing between the various senses of a word.

For example, the word *bank* has 18 senses (10 Noun senses and 8 Verb Senses). Consider the two senses of a *bank* – 1) *Bank* in the sense of "*a financial institution*", identifiable by the synset "*depository financial institution, bank, banking concern, banking company"*, and 2) *Bank* in the sense of *relying*, identifiable by the synset "*trust, swear, rely, bank"*. Now, the first sense has an objective polarity whereas the second sense has a positive polarity. This distinction cannot be made in the lexeme feature space, where we consider only the *first* sense of the word.

8 EVALUATION

8.1 Dataset

We performed experiments on three different datasets to validate our approach:

Dataset 1: Twitter is crawled using a Twitter API and 8507 tweets are collected based on a total of around 2000 different entities from 20 different domains. The following domains are used for crawling data: *Movie, Restaurant, Television, Politics, Sports, Education, Philosophy, Travel, Books Technology, Banking & Finance, Business, Music, Environment, Computers, Automobiles,*

Cosmetics brands, Amusement parks and *Eatables and History.* These are manually annotated by 4 annotators into four classes - *positive, negative, objective-not-spam* and *objective-spam.* The objective-not-spam category contains tweets which are objective in nature but are not spams. The objective-spam category contains spam tweets which are subsequently ignored during evaluation.

Dataset 2: Following the works of Read *et al.* (2005), Alec *et al.* (2009), Pak *et al.* (2010) and Gonzalez *et al.* (2011) we create an artificial dataset using *hashtags*. The Twitter API is used to collect another set of 15,214 tweets based on *hashtags*. Hashtags *#positive, #joy, #excited, #happy etc.* are used to collect tweets bearing positive sentiment, whereas hashtags like *#negative, #sad, #depressed, #gloomy, #disappointed etc.* are used to collect negative tweets. *Hashtag keywords are subsequently removed from the tweets.*

Dataset 3: *Travel Review* Dataset in Balamurali *et al.* (2011) contains 595 polarity-tagged documents for each of the positive and negative classes. All the words in the travel review documents are automatically tagged with their corresponding synset-id's using *Iterative Word Sense Disambiguation* algorithm (Khapra *et al.*, 2010).

8.2 Evaluation on the Twitter Dataset 1 and 2

The crawled tweets are pre-processed before evaluation. All the *links* (urls) in the tweets are replaced by *#link*. All the *user id's* in the tweets are replaced by *#user*.

Manually Annotated Dataset 1				
#Positive	#Negative	#Objective Not Spam	#Objective Spam	Total
2548	1209	2757	1993	8507
Auto Annotated Dataset 2				
#Positive		#Negative		Total
7348		7866		15214

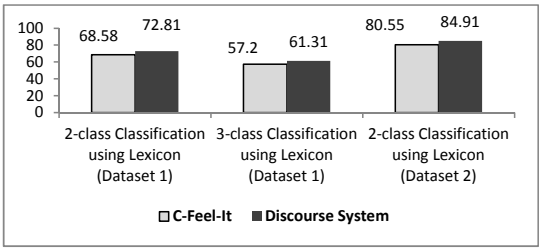
Table 4: Twitter Datasets 1 and 2 Statistics

Evaluations are performed in *Dataset 1* and *2* under a *2-class* and a *3-class* classification setting. In the 2-class setting, only *positive* and *negative* tweets are considered; whereas in the 3-class setting *positive, negative* and *objective-not-spam* tweets are considered. All the experiments in these two datasets are performed in the *lexeme feature space* using *lexicon-based classification* as well as *supervised classification*.

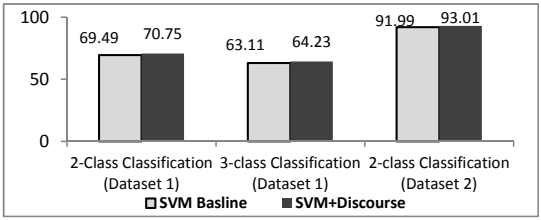
The baseline system, for this part of the evaluation, is taken as *C-Feel-It* (Joshi *et al.*, 2011). It is a rule-based system which implements a bag-of-words model using lexicon-based classification. The accuracy comparisons between C-Feel-It and the discourse system are performed under identical settings. The only difference between the two systems is the handling of *connectives, modals, conditionals* and *negation*, as indicated by *Algorithm 1*.

Graph 1 shows the accuracy comparison between C-Feel-It and the discourse system, in Datasets 1 and 2, using lexicon-based classification under a 2-class and a 3-class setting. *Graph 2* shows

the accuracy comparison between baseline SVM and SVM integrated with discourse features, in Datasets 1 and 2, under a 2-class and a 3-class setting. All the SVM features discussed in *Section 7.2*, except the discourse features arising out of the incorporation of *discourse weighting*, *modal and conditional indicator* and *negation*, are used in the baseline SVM model. A linear kernel, with default parameters, is used in the SVM (Chang *et al.*, 2011) with 10-fold cross-validation.



Graph 1: Accuracy Comparison between C-Feel-It and Discourse System using Lexicon in Datasets 1 and 2



Graph 2: Accuracy Comparison between Baseline SVM and SVM with Discourse in Datasets 1 and 2

8.3 Evaluation on the Travel Review Dataset 3

The *travel review* dataset (Balamurali *et al.*, 2011) is used to determine whether our discourse-based approach performs well for structured text as well. This evaluation is done under a 2-class classification setting in the *lexeme space* as well as the *sense space*. *Table 5* shows the accuracy comparison between the baseline bag-of-words model and bag-of-words model integrated with discourse features using lexicon-based classification, in dataset 3, under a 2-class setting.

Sentiment Evaluation Criterion	**Accuracy**
Baseline Bag-of-Words Model	69.62
Bag-of-Words Model + Discourse	**71.78**

Table 5: Accuracy Comparison between Bag-of-Words and Discourse System using Lexicon in **Dataset 3**

An automatic word sense disambiguation algorithm, *Iterative Word Sense Disambiguation* (Khapra *et al.*, 2010), has been used to auto-annotate the words in the review with their corresponding synset-id's. The same dataset is used in this work. A linear kernel, with default parameters, is used in the SVM with 5-fold cross-validation, similar to the compared system (Balamurali *et al.*, 2011). *Table 6* shows the performance of the discourse system along with the compared system using different features, on Dataset 3, using supervised classification. The features used in the SVM, for this part of the evaluation, include *stop word removal*, *no stemming*, *part-of-speech information* and *unigrams*. These features are used in all the systems in *Table 6*, including the discourse one. *Table 6* shows the system accuracy under four scenarios:

1. When only unigrams are used
2. When only sense of unigrams are used
3. When unigrams are used along with their senses (synset-id's)
4. When unigrams are used with senses and discourse features

9 DISCUSSIONS

Accuracy improvements over the baseline and the compared systems in all the datasets clearly signify the effectiveness of incorporating discourse information for sentiment classification. The bag-of-words model integrated *with discourse information* outperforms the bag-of-words model, *without this information*, under all the settings; although, the performance improvements vary in different settings. Statistical tests have been performed and all the accuracy improvements have been found to be statistically significant with *99%* level of confidence.

Systems	Accuracy (%)
Only Unigrams	84.90
Only IWSD Sense of Unigrams [26]	85.48
Unigrams + IWSD Sense of Unigrams [26]	86.08
Unigrams + IWSD Sense of Unigrams + Discourse Features	**88.13**

Table 6: Accuracy Comparisons in Travel Review Dataset 3

9.1 Accuracy Comparison between C-Feel-It and Discourse System

These comparisons are performed under a 2-class and a 3-class classification setting, using lexicon-based classification, in the lexeme space under identical settings - the only difference being the incorporation of discourse features. In *Dataset 1*, there is an accuracy improvement of around *4%* over C-Feel-It for both 2-class and 3-class classification. The discourse system accuracy at *72.81%* for 2-class classification is higher than that of the 3-class classification accuracy of *61.31%*. This shows that 3-class classification of tweets is much more difficult than 2-class classification.

9.2 Accuracy Comparison between Baseline SVM and Discourse System

These comparisons are performed under a 2-class and a 3-class classification setting, using supervised classification, in the lexeme space. A similar feature set, except the discourse features, is used for both the systems. In *Dataset 1*, there is an accuracy improvement of *1%* in both the 2-class and 3-class classification, which has been found to be statistically significant. In *Dataset 2*, there is an accuracy improvement of *2%* over baseline SVM for 2-class classification. It is observed that in the 2-class setting, the discourse system performs better in the lexicon-based classification with an accuracy of *72.81%* compared to the supervised classification accuracy of *70.75%*. This is contrary to the common scenario in text classification, where the supervised classification system always performs much better than the lexicon-based classification. This may be due to the very sparse feature space, owing to the length limit of tweets (140 characters).

9.3 Accuracy Comparison in Dataset 3

In the *Travel review* dataset, lexicon-based classification yielded an accuracy improvement of *2%* for the discourse model over simple bag-of-words model, in the *lexeme space*. In the SVM classification, in the *sense space,* under a 2-class setting, the discourse system achieved an accuracy of *88%* compared to *86%* accuracy of Balamurali *et al.* (2011). A similar feature set has been used in both the models, which indicates that the performance improvement is due to the incorporation of discourse features in SVM.

9.4 Drawbacks

The lexicon-based classification suffers from the usage of a generic lexicon in the *lexeme space*, where it cannot distinguish between the various senses of a word. The lexicons do not have entries for the interjections like *wow*, *duh etc.* which are strong indicators of sentiment. The frequent spelling mistakes, abbreviations and slangs used in the tweets do not have entry in the lexicons. For example, *love* and *great* are frequently written as *luv* and *gr8* respectively, which will not be detected. A spell-checker may help the system in this regard.

The supervised system suffers from a sparse feature space due to very short contexts. A concept expansion approach, to expand the feature vectors, may prove to be useful. This is due to the extensive world knowledge embedded in the tweets. For example, the tweet "*He is a Frankenstein*" is tagged as objective. The knowledge that *Frankenstein* is a negative concept is not present in the lexicon. The IWSD algorithm for automatic sense annotation has an F-Score of *70%*, which means many of the word-senses were wrongly tagged. In case a better WSD algorithm is used, higher system accuracy can be achieved in the travel review dataset.

In the absence of *parsing* and *tagging* information, due to the noisy nature of the tweets, the scope of the discourse marker has been heuristically taken till the sentence boundary or till the next discourse marker. Consider the sentence, "*I wanted to follow my dreams and ambitions despite all the obstacles, but I did not succeed.*" Here *want* and *ambition* will get the polarity *+2 each*, as they appear before *despite*; *obstacle* will get a polarity *-1* and *not succeed* will get a polarity *-2*. Thus the overall polarity is *+1*, whereas the overall sentiment should be *negative*.

This is because we do not consider the *positional importance* of a discourse marker in the sentence and consider all the discourse markers to be equally important. A better method is to give a ranking to the discourse markers based on their *positional* and *pragmatic* importance.

10 FUTURE WORKS AND CONCLUSIONS

In this work, we showed that the incorporation of discourse markers in a bag-of-words model improves the sentiment classification accuracy by *2 - 4%*. This approach is particularly beneficial for - 1) applications dealing with noisy text where *parsing* and *tagging* do not perform very well, and 2) applications, requiring a fast response time, where employing a heavy linguistic tool like *parsing* will be detrimental to its performance due to the increased processing time.

Most of the works in micro-blogs, like *Twitter*, build on a bag-of-words model that ignores the discourse markers. We demonstrated an approach to incorporate discourse information to improve their performance, retaining the simplicity of the bag-of-words model. We validated this claim on two different datasets (manually and automatically annotated) from *Twitter*, where we achieved an accuracy improvement of *4%* for lexicon-based classification over an existing application (Joshi *et al.*, 2011), and 2% for supervised classification over the baseline SVM with advanced features. We also showed that our method fares well for structured reviews as well, where we achieved similar accuracy improvements over Balamurali *et al.*, 2011.

References

Alec, G.; Lei, H.; and Richa, B. Twitter sentiment classification using distant supervision. Technical report, Standford University. 2009

AR, Balamurali and Joshi, Aditya and Bhattacharyya, Pushpak. Harnessing WordNet Senses for Supervised Sentiment Classification. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP). 2011

Asher, Nicholas and Benamara, Farah and Mathieu, Yvette Yannick. Distilling opinion in discourse: A preliminary study. In Proceedings of Computational Linguistics (CoLing). 2008

Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In 23rd International Conference on Computational Linguistics: Posters, 36–44. 2010

Bermingham, A., and Smeaton, A. Classifying sentiment in microblogs: is brevity an advantage ACM 1833–1836. 2010

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. 2011

Dey, Lipika and Haque, Sk. Opinion Mining from Noisy Text Data. International Journal on Document Analysis and Recognition 12(3). pp 205-226. 2009

Elwell, Robert and Baldridge, Jason. Discourse Connective Argument Identification with Connective Specific Rankers. In Proceedings of IEEE International Conference on Semantic Computing. 2008

Gonzalez-Ibanez, Roberto and Muresan, Smaranda and Wacholder, Nina. Identifying sarcasm in Twitter: a closer look, In Proceedings of ACL: short paper. 2011

Hobbs, Jerry R., and Michael Agar.The Coherence of Incoherent Discourse, Journal of Language and Social Psychology, vol. 4, nos. 3-4, pp. 213-232. 1985

Joshi, A.; Balamurali, A. R.; Bhattacharyya, P.; and Mohanty, R. C-feel-it: a sentiment analyzer for microblogs. In Proceedings of ACL: Systems Demonstrations, HLT '11, 127–132. 2011

Mann, William C. and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3), 243-281. 1988

Marcu, Daniel. 2000. The Theory and Practice of Discourse Parsing and Summarization, MIT Press, Cambridge, MA.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD. 2004

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. Domain-specific word sense disambiguation combining corpus basedand wordnet based parameters. In Proceedings of GWC'10, Mumbai, India. 2010

Ng, Vincent and Dasgupta, Sajib and Arifin, S. M. Niaz. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In Proceedings of the COLING/ACL on Main conference poster sessions. 2006

Pak, Alexander and Paroubek, Patrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of LREC. 2010

Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical Methods in Natural Language. 2002

Pitler, Emily and Louis, Annie and Nenkova, Ani. Automatic Sense Prediction for Implicit Discourse Relations in Text. In Proceedings of ACL and IJCNLP. 2009

Polanyi, Livia and Zaenen, Annie. Contextual valence shifters. In Exploring Attitude and Affect in Text: Theories and Applications. AAAI Spring Symposium Series. 2004

Prelovac, V. 2010. Top social media sites. Web.

Read, Jonathon. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of Association for Computer Linguistics (ACL). 2005

Somasundaran, Swapna. Discourse-level relations for Opinion Analysis. PhD Thesis, University of Pittsburgh. 2010

Taboada, Maite and Brooke, Julian and Tofiloski, Milan and Voll, Kimberly and Stede, Manfred. Lexicon-based methods for sentiment analysis. Computational Linguistics. 2011

Taboada, Maite and Brooke, Julian and Voll, Kimberly. Extracting Sentiment as a Function of

Discourse Structure and Topicality. Simon Fraser Univeristy School of Computing Science Technical Report. 2008

Webber, Bonnie and Knott, Alistair and Stone, Matthew and Joshi, Aravind. Discourse relations: A structural and presuppositional account using lexicalized tag. In Proceedings of ACL. 1999

Wellner, Ben and Pustejovsky, James. Automatically identifying the arguments of discourse connectives. In Proceedings of The Joint Conference on EMNLP-CoNLL, pp. 92-101. 2007

Wolf, Florian and Gibson, Edward and Desmet, Timothy. Discourse coherence and pronoun resolution, Language and Cognitive Processes, 19(6), pp. 665–675. 2004

Wolf, Florian and Gibson, Edward. Representing discourse coherence: A corpus-based study, Computational Linguistics, 31(2), pp. 249–287. 2005

Zhou, Lanjun and Li, Binyang and Gao, Wei and Wei, Zhongyu and Wong, Kam-Fai. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In Proceedings of EMNLP. 2011

Zirn, C\"{a}cilia and Niepert, Mathias and Stuckenschmidt, Heiner and Strube, Michael. Fine-Grained Sentiment Analysis with Structural Features. In Proceedings of IJCNLP. 2011