Sentiment Analysis in Twitter with Lightweight Discourse Analysis

Subhabrata Mukherjee, Pushpak Bhattacharyya
IBM Research Lab, India
Dept. of Computer Science and Engineering,
Indian Institute of Technology, Bombay

24th International Conference on Computational Linguistics
COLING 2012,
IIT Bombay, Mumbai, Dec 8 - Dec 15, 2012

Discourse

Discourse

- *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).

Discourse

- *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).

- Presence of a discourse marker can alter the overall sentiment of a sentence

Discourse

- *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).

- Presence of a discourse marker can alter the overall sentiment of a sentence

  In most of the bag-of-words models, the discourse markers are ignored as stop words during feature vector creation

Motivation

- *i'm quite excited about Tintin,* ***despite*** *not really liking original comics - probably because Joe Cornish had a hand in*

- *Think i'll stay with the whole 'sci-fi' shit* **but** *this time...a classic movie.*

Motivation Contd…

- Traditional works in *discourse analysis* use parsing of some form like a discourse parser or a dependency parser
- 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

Motivation Contd…

- 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*
- The errors cause natural language processing tools like *parsers* and *taggers* to fail frequently
- Increased processing time adds an overhead to real-time applications

Discourse Relation

- A coherently structured discourse is a collection of sentences having some relation with each other
- A coherent relation reflects how different discourse segments interact
- Discourse segments are non-overlapping spans of text

Discourse Coherent Relations

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*

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

Discourse Coherent Relations Examples

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).*

Discourse Relations and Sentiment Analysis

- Not all discourse relations are significant for sentiment analysis

- Discourse relation essential for Sentiment Analysis
  - That connects segments having contrasting information
    - Violated Expectations
  - That places higher importance to certain discourse segments
    - Inferential Conjunctions
  - That incorporates hypothetical situation in the context
    - Conditionals

- Semantic Operators influencing discourse relations in Sentiment Analysis
  - That incorporates hypothetical situation in the context
    - Modals
  - That negates the information in the discourse segment
    - Negation

Violated Expectations and Contrast

- *Violating expectation* conjunctions oppose or refute the neighboring discourse segment

- We categorize them into *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

Violated Expectations and Contrast

- *Violating expectation* conjunctions oppose or refute the neighboring discourse segment

- We categorize them into *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

(i'm quite excited about Tintin), **despite** *(not really liking original comics.)*
(my daughter is off school very poorly), **but** *(brightened up when we saw you on gmtv today).*

Conclusive or Inferential Conjunctions

- These are the set of conjunctions that tend to draw a conclusion or inference

- Hence, the discourse segment following them should be given more weight

Conclusive or Inferential Conjunctions

- These are the set of conjunctions that tend to draw a conclusion or inference

- Hence, the discourse segment following them should be given more weight

@User *I was not much satisfied with ur so-called good phone and* **subsequently** *decided to reject it.*

Conditionals

- Conditionals introduce 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

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.* Modals depict irrealis events

- We 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*

- In *supervised classifiers*, the modals are marked as features.

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.* Modals depict irrealis events

- We 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*

(Strong Modals): *Unless I missed the announcement their God is now featured on postage stamps, it* **might** *be a hard sell.*
(Weak Modals): *G.E 12* **must** *be the* most *deadly General Election for politicians ever.*

Negation

- The negation operator 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

- **(Negation):** *I do not like Nokia* **but** *I like Samsung*

- 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

Features

Discourse 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*

Algorithm

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Words after them are given more weightage
- Frequency count of those words is incremented by 1

- *The movie looked promising^{+1}, but it failed^{-2} to make an impact in the box-office*

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- Words after them are given more weightage
- Frequency count of those words is incremented by 1

- *The movie looked promising^{+1}, but it failed^{-2} to make an impact in the box-office*

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- Words before them are given more weightage
- Frequency count of those words is incremented by 1

- *India staged a marvelous victory^{+2} down under despite all odds^{-1}.*

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- All sentences containing *if* are marked, in supervised classifiers

> - Words before them are given more weightage
> - Frequency count of those words is incremented by 1
>
> - *India staged a marvelous victory^{+2} down under despite all odds^{-1}.*

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- All sentences containing *if* are marked, in supervised classifiers

- In lexicon-based classifiers, their weights are decreased

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- All sentences containing *if* are marked, in supervised classifiers

- All sentences containing *strong modals* are marked, in supervised classifiers

  - In lexicon-based classifiers, their weights are decreased

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*
- Conj_Prev - *till, until, despite, in spite, though, although*
- All sentences containing *if* are marked, in supervised classifiers
- All sentences containing *strong modals* are marked, in supervised classifiers

- In lexicon-based classifiers, their weights are decreased

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*

- Conj_Prev - *till, until, despite, in spite, though, although*

- All sentences containing *if* are marked, in supervised classifiers

- All sentences containing *strong modals* are marked, in supervised classifiers

  - In lexicon-based classifiers, their weights are decreased

Algorithm

- Conj_Fol, Conj_Infer *: but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence*
- Conj_Prev - *till, until, despite, in spite, though, although*
- All sentences containing *if* are marked, in supervised classifiers
- All sentences containing *strong modals* are marked, in supervised classifiers

> - A window of 5 is considered
> - Polarity of all words in the window are reversed till another violating expectation conjunction is encountered
> - The polarity reversals are specially marked
> - *I do not like*$^{-1}$ *Nokia* **but** *I like*$^{+2}$ *Samsung.*

Algorithm Contd…

- 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
- The weight of a word w_{ij} is adjusted according to the presence of a discourse marker or a semantic operator
- 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.
- **Input**: Review R
- **Output**: w_{ij}, f_{ij} , $flip_{ij}$, hyp_{ij}

Lexical Classification

Bing Liu sentiment lexicon (Hu *et al.*, 2004) is used to find the polarity *pol(wij)* of a word w_{ij}

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

$$\begin{aligned} where\ p(w_{ij}) &= pol(w_{ij})\ \ if\ hyp_{ij} = 0 \\ &= \frac{pol(w_{ij})}{2}\ \ \ if\ hyp_{ij} = 1 \end{aligned}$$

Supervised Classification

- Support Vector Machines are used with the following features:
  - N-grams (N=1,2)
  - Stop Word Removal (except discourse markers)
  - Discourse Weight of Features - f_{ij}
  - Modal and Conditional Indicators - hyp_{ij}
  - Stemming
  - Negation - $flip_{ij}$
  - Emoticons
  - Part-of-Speech Information
  - Feature Space
    - Lexeme - w_{ij}
    - *Sense-Space – Synset-id*(w_{ij})

Datasets

Datasets

- Dataset 1 (Twitter – Manually Annotated)
  - 8507 tweets over 2000 entities from 20 domains
  - Annotated by 4 annotators into positive, negative and objective classes

Datasets

- Dataset 1 (Twitter – Manually Annotated)
  - 8507 tweets over 2000 entities from 20 domains
  - Annotated by 4 annotators into positive, negative and objective classes

- Dataset 2 (Twitter – Auto Annotated)
  - 15,214 tweets collected and annotated based on hashtags
  - Positive hashtags - #positive, #joy, #excited, #happy
  - Negative hashtags - *#negative, #sad, #depressed, #gloomy,#disappointed*

Datasets Contd…

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

Datasets Contd…

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

- Dataset 3 (Travel Domain - Balamurali *et al.*, EMNLP 2011)
  - Each word is manually tagged with its disambiguated WordNet sense
  - Contains 595 polarity tagged documents of each

Dataset Domains

Movie, Restaurant, Television, Politics, Sports, Education, Philosophy, Travel, Books, Technology, Banking & Finance, Business, Music, Environment, Computers, Automobiles, Cosmetics brands, Amusement parks, Eatables, History

Baselines

- Twitter
  - C-Feel-It (Joshi *et al.*, 2011, ACL)

- Travel Reviews
  - Balamurali *et al.*, 2011, EMNLP
  - Iterative Word-Sense Disambiguation Algorithm (Khapra *et al.*, 2010, GWC) is used to auto sense-annotate the words

Features

- Let a user post *R* consist of *m* sentences s_i $(i=1 \ldots m)$, where each s_i consist of n_i words w_{ij} $(i=1 \ldots m, j=1 \ldots n_i)$

- Let f_{ij} be the weight of the word w_{ij} in sentence s_i, initialized to 1

- The weight of a word w_{ij} is adjusted according to the presence of a discourse marker or a semantic operator

- 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.

- **Input**: Review *R*
- **Output**: w_{ij}, f_{ij} , $flip_{ij}$, hyp_{ij}

Classification Results in Twitter (Datasets 1 and 2)

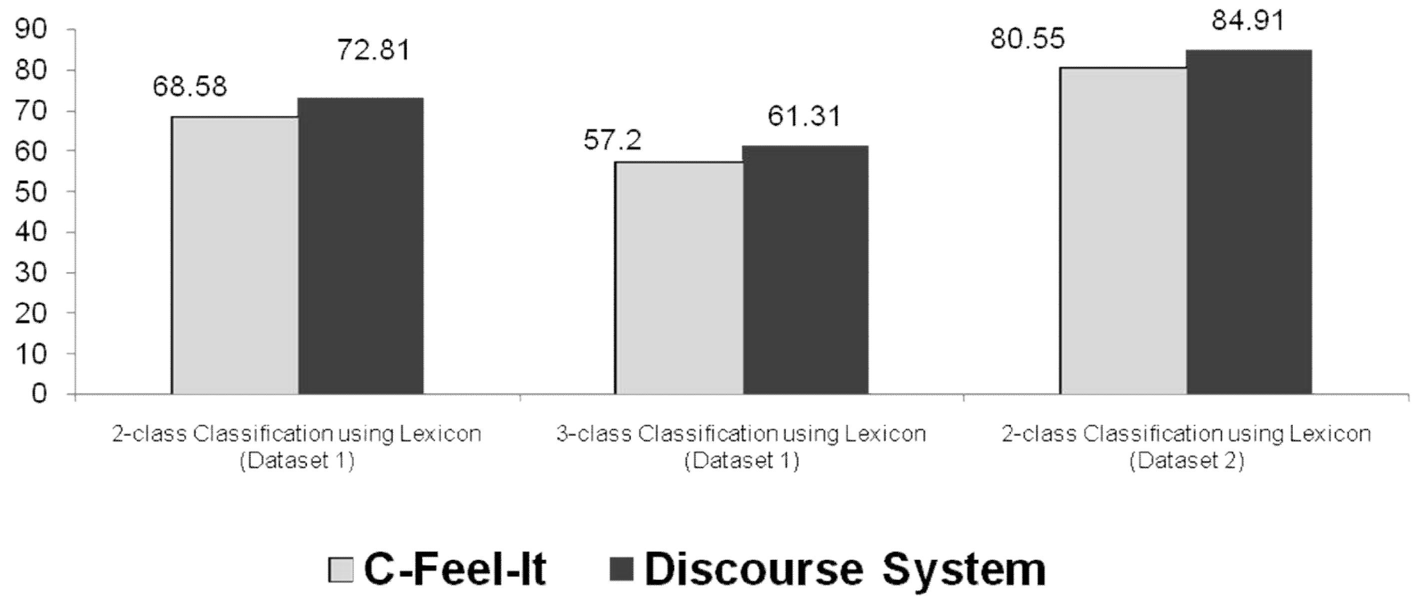
Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

Classification Results in Twitter (Datasets 1 and 2)
Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

Lexicon-based Classification

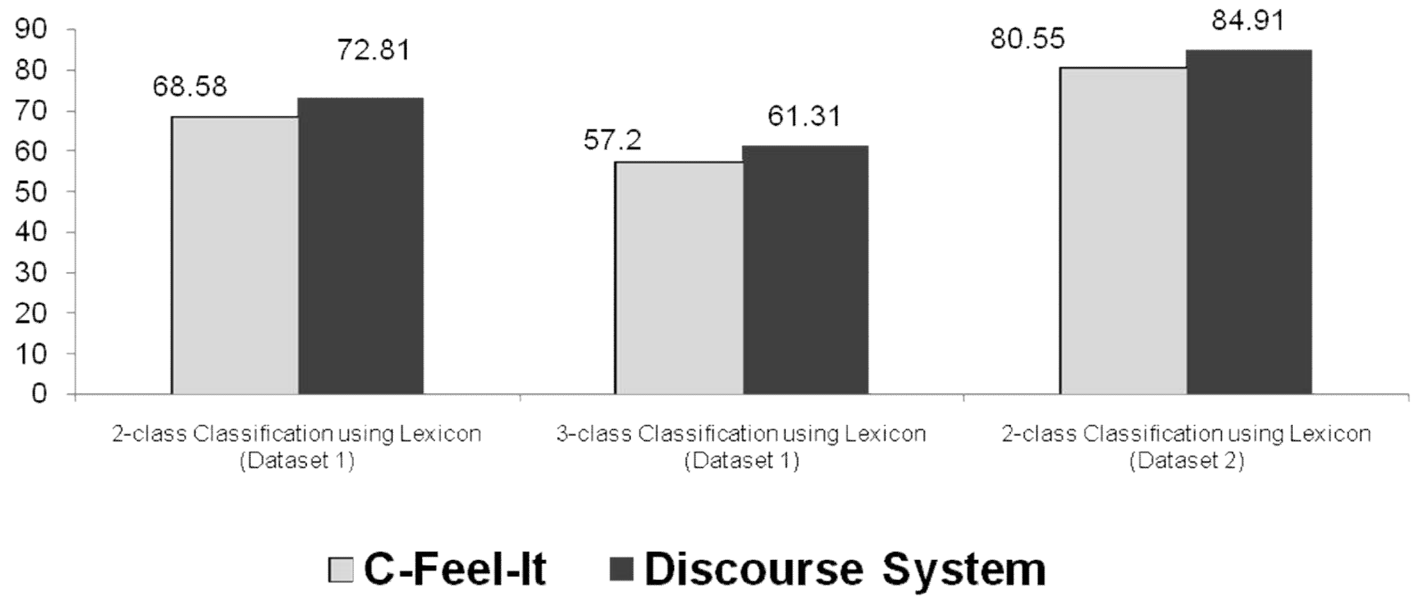
Classification Results in Twitter (Datasets 1 and 2) Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

Lexicon-based Classification



Classification Results in Twitter (Datasets 1 and 2) Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

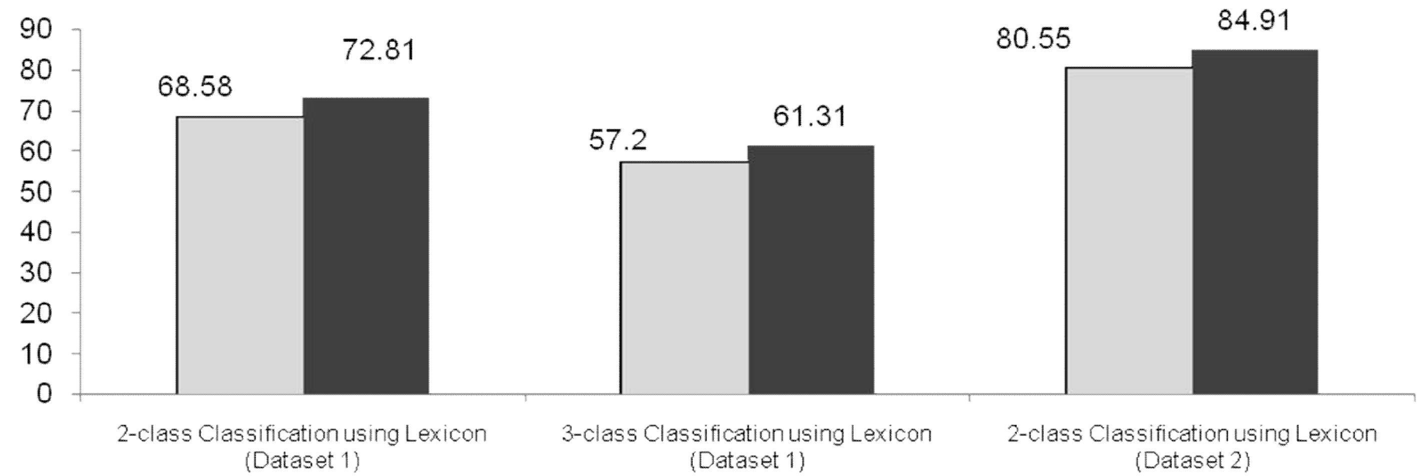
Lexicon-based Classification



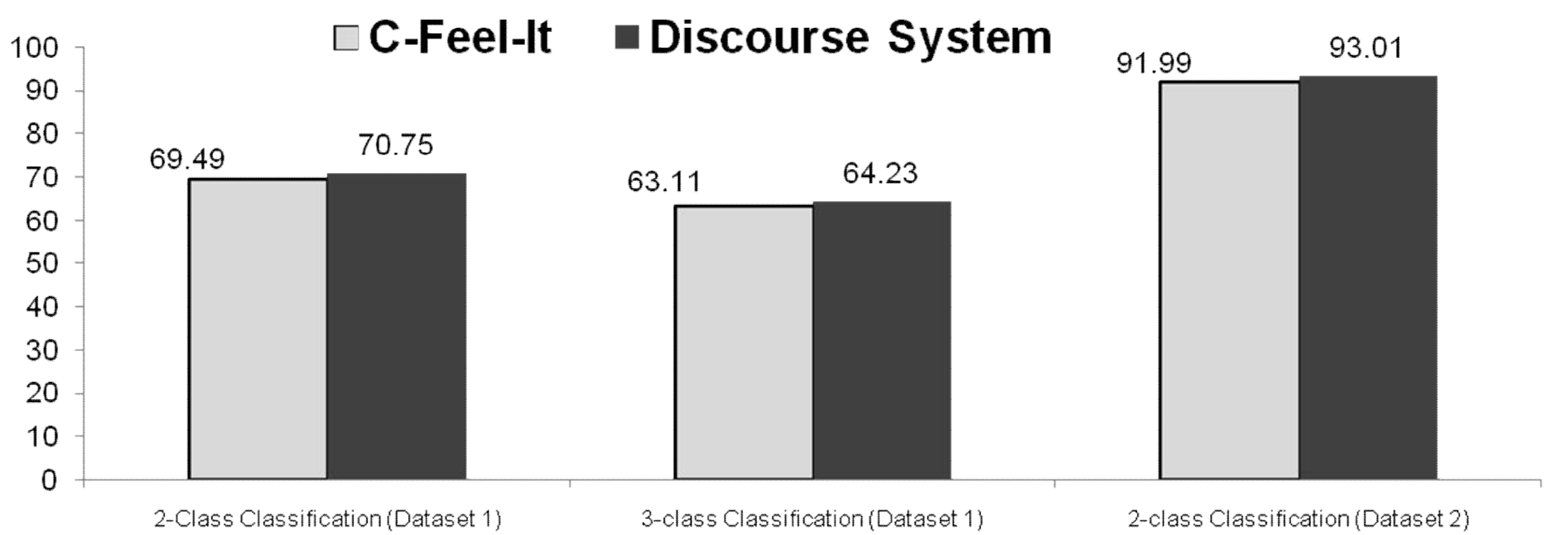
Supervised Classification

Classification Results in Twitter (Datasets 1 and 2)
Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

Lexicon-based Classification



Supervised Classification



Classification Results in Travel Reviews (Dataset 3) Comparison with Balamurali *et al.*, EMNLP 2011)

Classification Results in Travel Reviews (Dataset 3) Comparison with Balamurali *et al.*, EMNLP 2011)

Lexicon-based
Classification

Classification Results in Travel Reviews (Dataset 3) Comparison with Balamurali *et al.*, EMNLP 2011)

Lexicon-based Classification

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

Classification Results in Travel Reviews (Dataset 3) Comparison with Balamurali *et al.*, EMNLP 2011)

Lexicon-based Classification

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

Supervised Classification

Classification Results in Travel Reviews (Dataset 3) Comparison with Balamurali *et al.*, EMNLP 2011)

Lexicon-based Classification

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

Supervised Classification

Systems	Accuracy (%)
Baseline Accuracy (Only Unigrams)	84.90
Balamurali *et al.*, 2011 (Only IWSD Sense of Unigrams)	85.48
Balamurali *et al.*, 2011 (Unigrams+IWSD Sense of Unigrams)	86.08
Unigrams + IWSD Sense of Unigrams+Discourse Features	**88.13**

Drawbacks

- Usage of a generic lexicon in lexeme feature space
- Lexicons do not have entries for interjections like *wow*, *duh etc*. which are strong indicators of sentiment
- Noisy Text (*luv, gr8, spams, …)*
- Sparse feature space (140 chars) for supervised classification
- 70% accuracy of IWSD in sense space for travel review classification

Drawbacks Contd…

- *I wanted^{+2} to follow my dreams and ambitions^{+2} **despite** all the obstacles^{-1}, **but** I did not succeed^{-2}.*

- *want* and *ambition* will get polarity *+2 each*, as they appear before *despite, obstacle* will get polarity *-1* and *not succeed* will get a polarity *-2 as they appear after but*

- Overall polarity is *+1,* whereas the overall sentiment should be *negative*

- We do not consider *positional importance* of a discourse marker in the sentence and consider all markers equally important

- Better give a ranking to the discourse markers based on their *positional* importance

- Thank You
- Questions ?