Adaptation of Sentiment Analysis to New Linguistic Features, Informal Language Form and World Knowledge

Subhabrata Mukherjee

Master's Thesis

Guide: Dr. Pushpak Bhattacharyya

Department of Computer Science and Engineering,

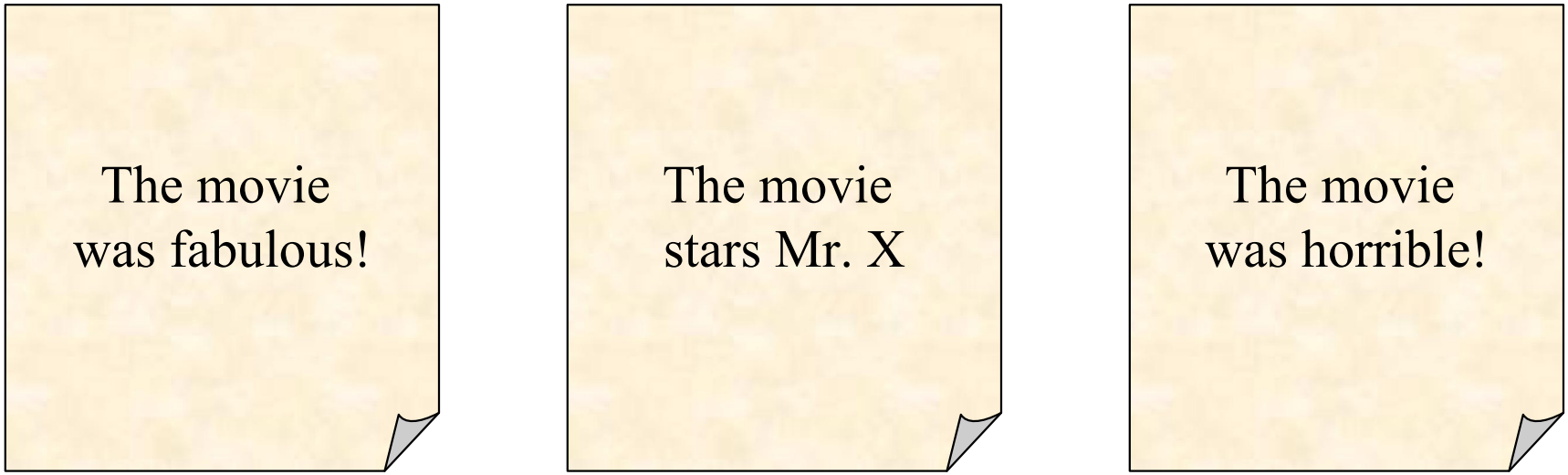
IIT Bombay

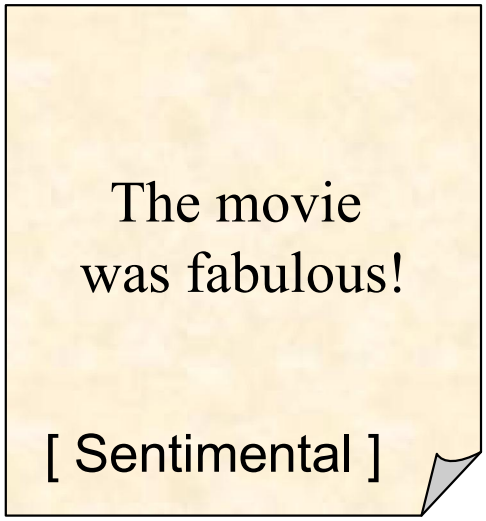
7/23/2013

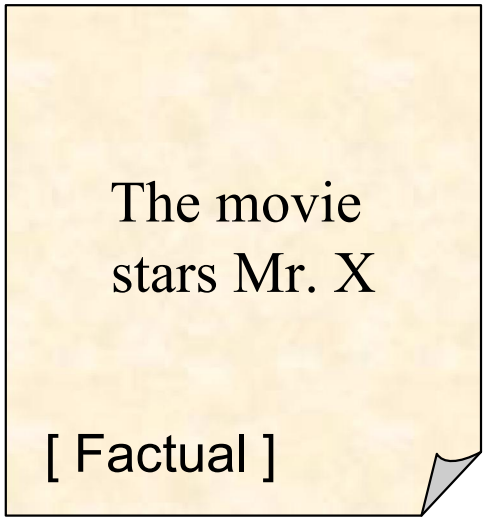
M.TECH STAGE II PRESENTATION

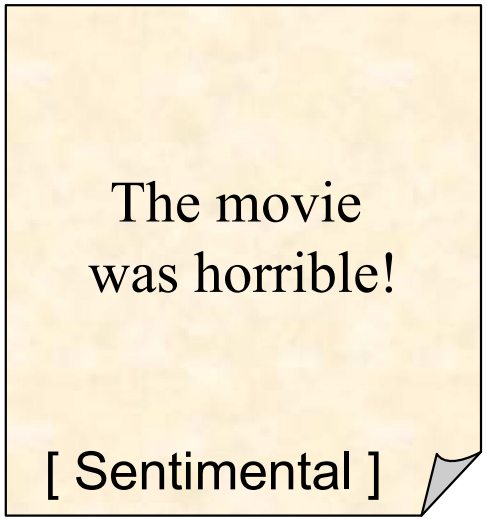
Roadmap

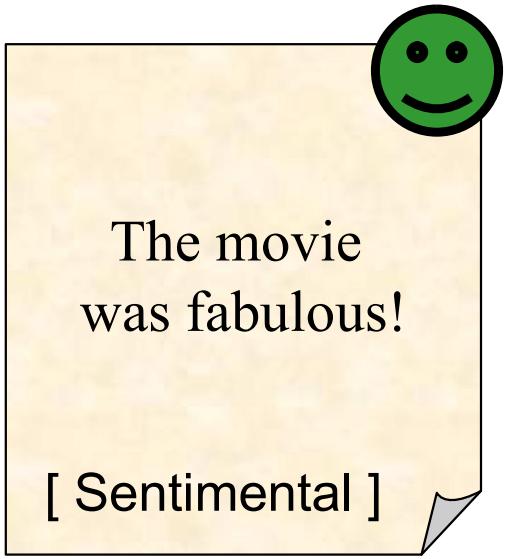
- Motivation
- Role of Feature Specificity in Sentiment Analysis
- Role of Discourse Specificity in Sentiment Analysis
- Role of World Knowledge in Sentiment Analysis
- Role of Social Media Content and Informal Language Form in Sentiment Analysis
- Applications
  - Role of Social Media and World Knowledge in IR
  - Role of Sentiment in Semantic Similarity Measure

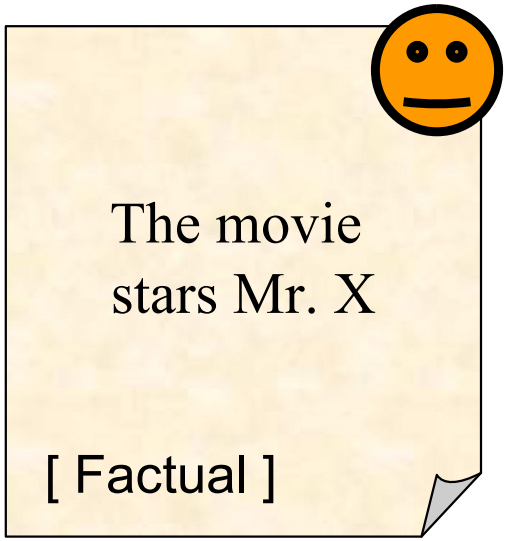

The movie was fabulous!
The movie stars Mr. X
The movie was horrible!

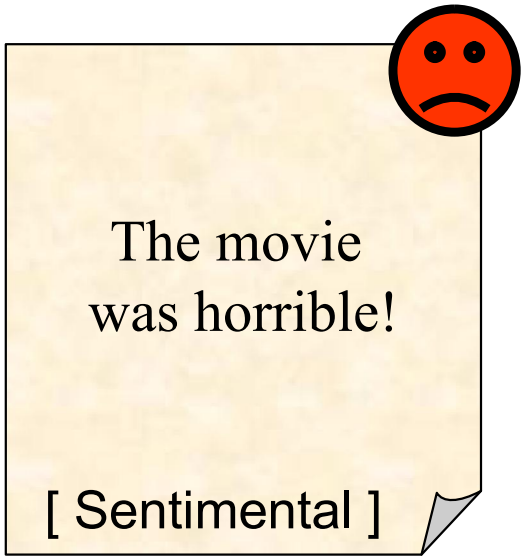

The movie
was fabulous!
[Sentimental]


The movie
stars Mr. X
[Factual]


The movie
was horrible!
[Sentimental]

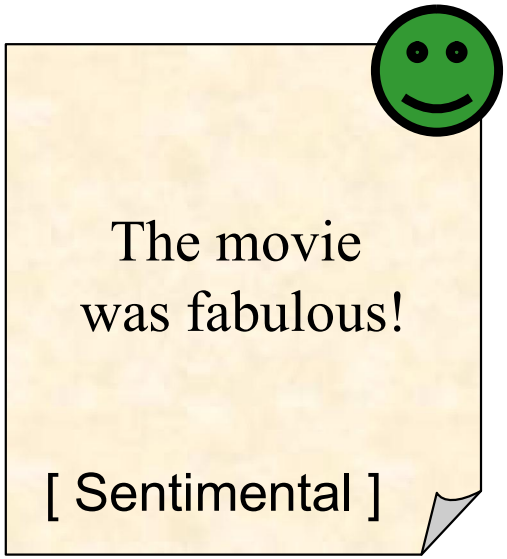

The movie
was fabulous!
[Sentimental]

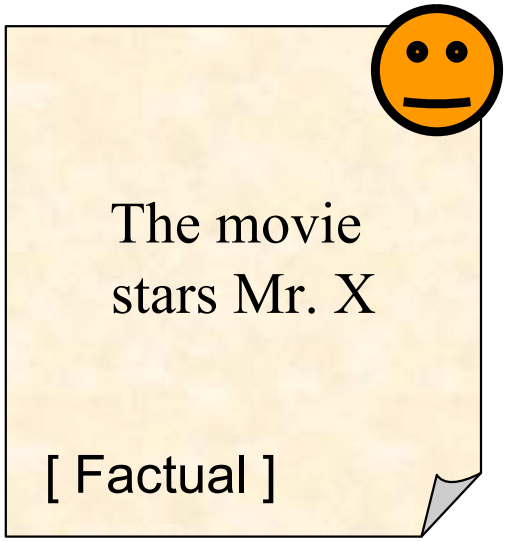

The movie
stars Mr. X
[Factual]

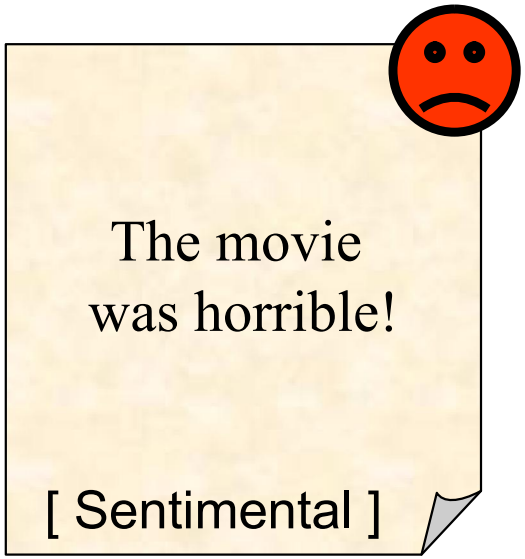

The movie
was horrible!
[Sentimental]

What is Sentiment Analysis

- Identify the orientation of opinion in a piece of text







Can be generalized to a wider set of emotions

Motivation Contd…

Motivation Contd…

- **Discourse Coherency**

Motivation Contd…

- **Discourse Coherency**

He won the election **despite** *a lot of smearing and slandering by the opposition.*

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**

He won the election **despite** *a lot of smearing and slandering by the opposition.*

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**

I like **Samsung's** *multimedia features but that of* **Nokia** *is not that good.*

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**
- **Informal Language Form and Social Media Content**

I like **Samsung's** *multimedia features but that of* **Nokia** *is not that good.*

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**
- **Informal Language Form and Social Media Content**

It really was. :) RT @AlisssonRicardo: Chernobyl lukd dumb but **Men in Black 3** was prety guuud ☺

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**
- **Informal Language Form and Social Media Content**

World Knowledge

It really was. :) RT @AlisssonRicardo: Chernobyl lukd dumb but **Men in Black 3** was prety guuud ☺

Motivation Contd…

- **Discourse Coherency**
- **Feature Specificity**
- **Informal Language Form and Social Media Content**

  World Knowledge

He is behaving like a Frankenstein.
"I am looking forward to spend a nice evening with my **parents***"*
– part of a movie review

- Feature Specific Sentiment Analysis

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*
  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*
  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

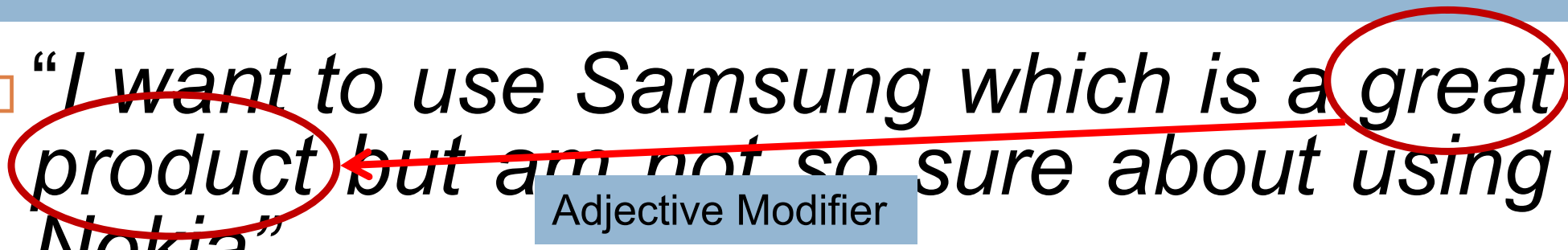
- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*
  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*
  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*



  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*

  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

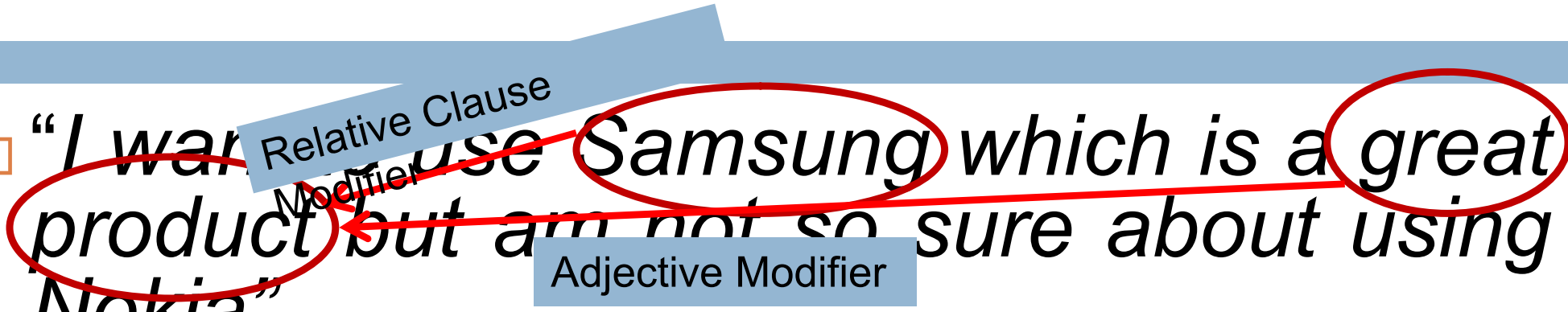
Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*

  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Hypothesis Example

- *"I want to use Samsung which is a great product but am not so sure about using Nokia".*



  - *Here "great" and "product" are related by an adjective modifier relation, "product" and "Samsung" are related by a relative clause modifier relation. Thus "great" and "Samsung" are transitively related.*
  - ***Here "great" and "product" are more related to Samsung than they are to Nokia***
  - *Hence "great" and "product" come together to express an opinion about the entity "Samsung" than about the entity "Nokia"*

Relations

- Direct Neighbor Relation
  - Capture **short range dependencies**
  - Any 2 consecutive words (such that none of them is a StopWord) are directly related
  - Consider a sentence *S* and 2 consecutive words
  - *If* $w_i, w_{i+1} \notin Stopwords$, then they are directly related. $w_i, w_{i+1} \in S$

- Dependency Relation
  - Capture **long range dependencies**
  - Let *Dependency_Relation* be the list of **significant relations**.
- Any 2 words w_i and w_j in *S* are directly related, *if* $\exists D_i$ *s.t.* $D_i(w_i, w_j) \in Dependency\_Relation$

Algorithm

A Graph $G(W, E)$ is constructed such that any $w_i, w_j \in W$ are directly connected by $e_k \in E$, *if* $\exists R_l$ *s.t.* $R_l(w_i, w_j) \in R$.

i. Initialize *n* clusters $C_i \ \forall\, i = 1 \ldots n$

ii. Make each $f_i \in F$ the clusterhead of C_i. The target feature f_t is the clusterhead of C_t. Initially, each cluster consists only of the clusterhead.

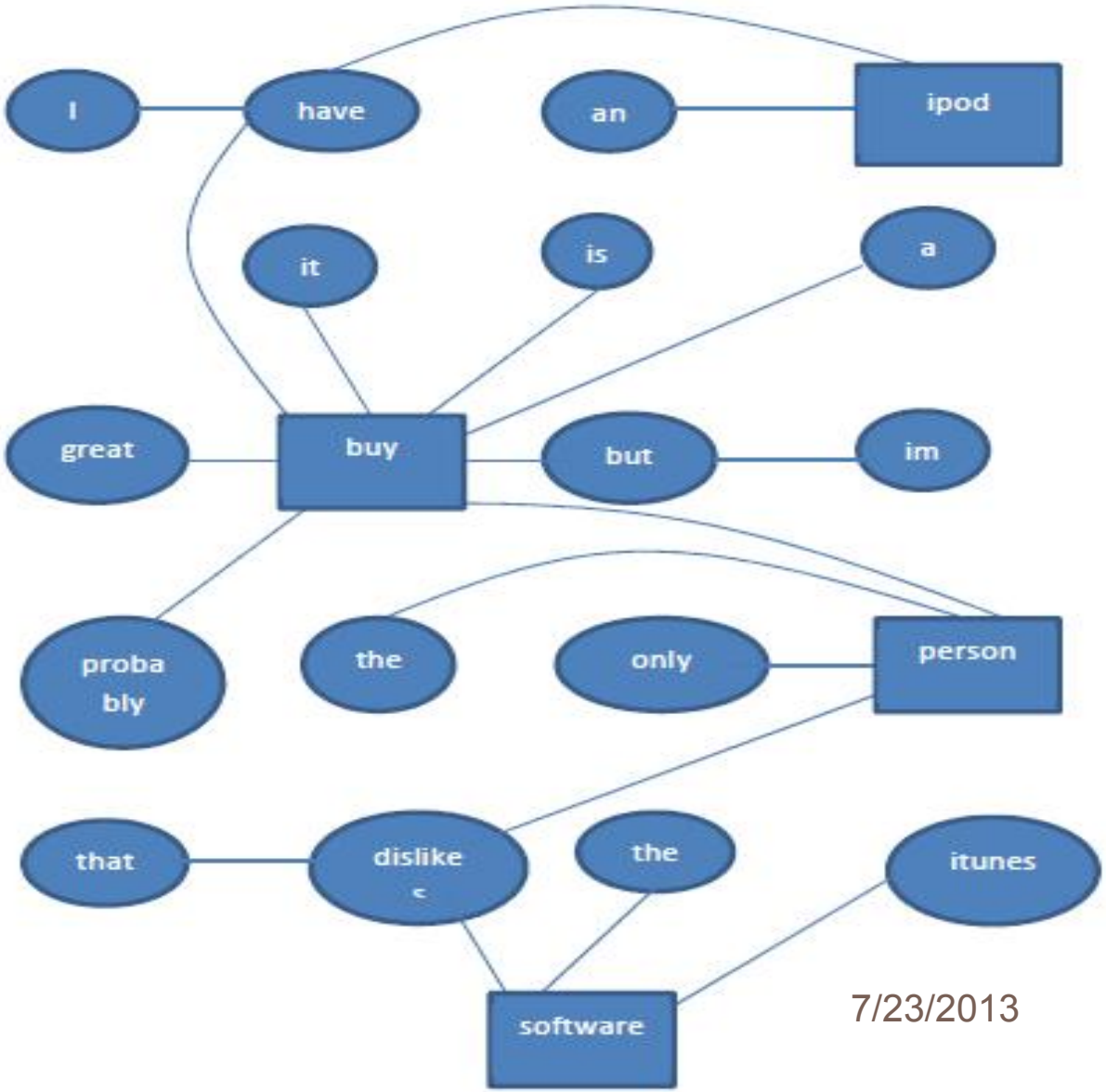
iii. Assign each word $w_j \in S$ to cluster C_k *s.t.,* $k = argmin_{i \in n} dist(w_j, f_i)$, *Where* $dist(w_j, f_i)$ *gives the number of edges, in the shortest path, connecting* w_j *and* f_i *in G.*

iv. Merge any cluster C_i with C_t *if* $dist(w_j, f_i) < \theta$, Where θ is some threshold distance.

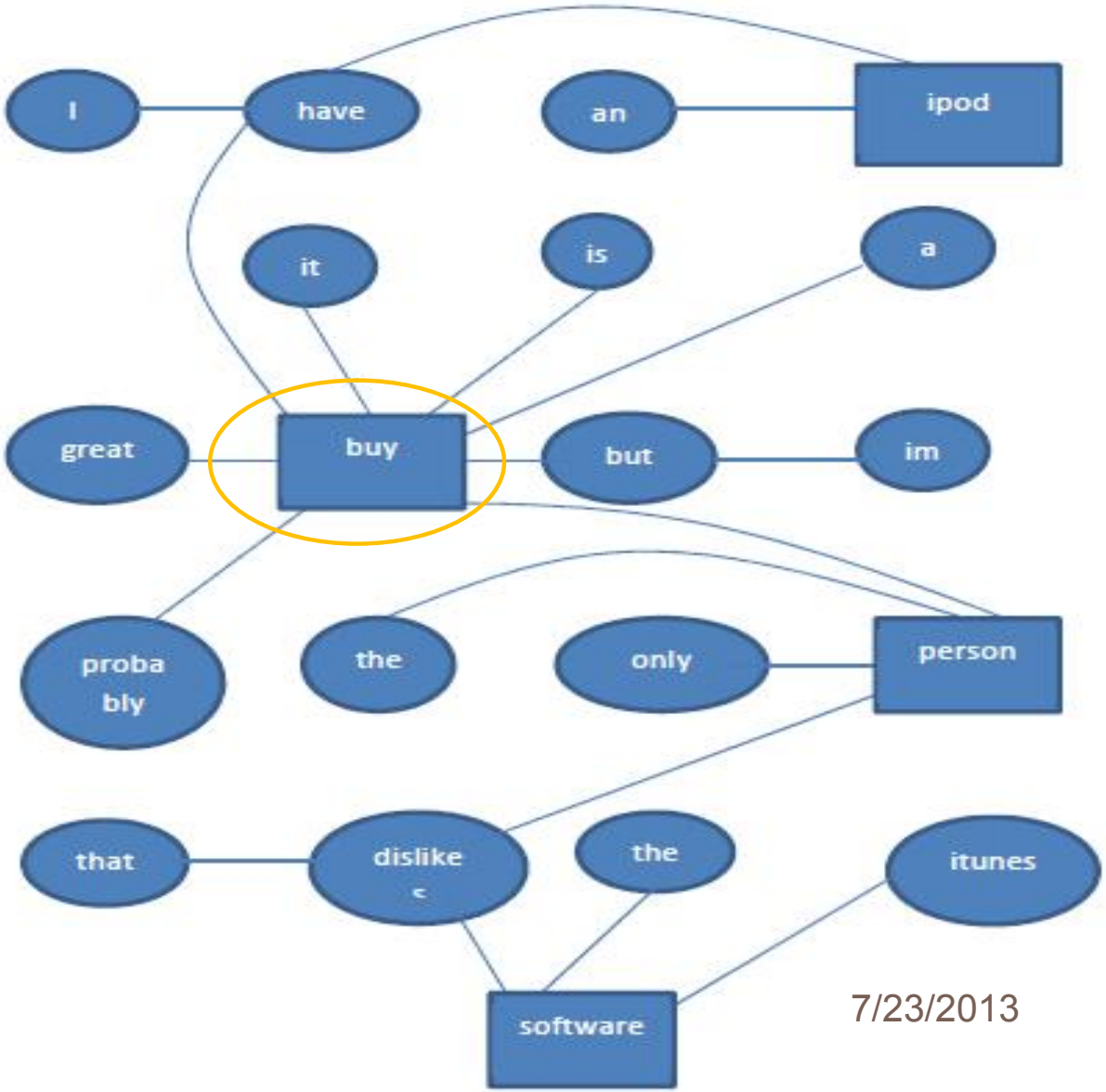
v. Finally the set of words $w_i \in C_t$ gives the opinion expression regarding the target feature f_t.

Graph

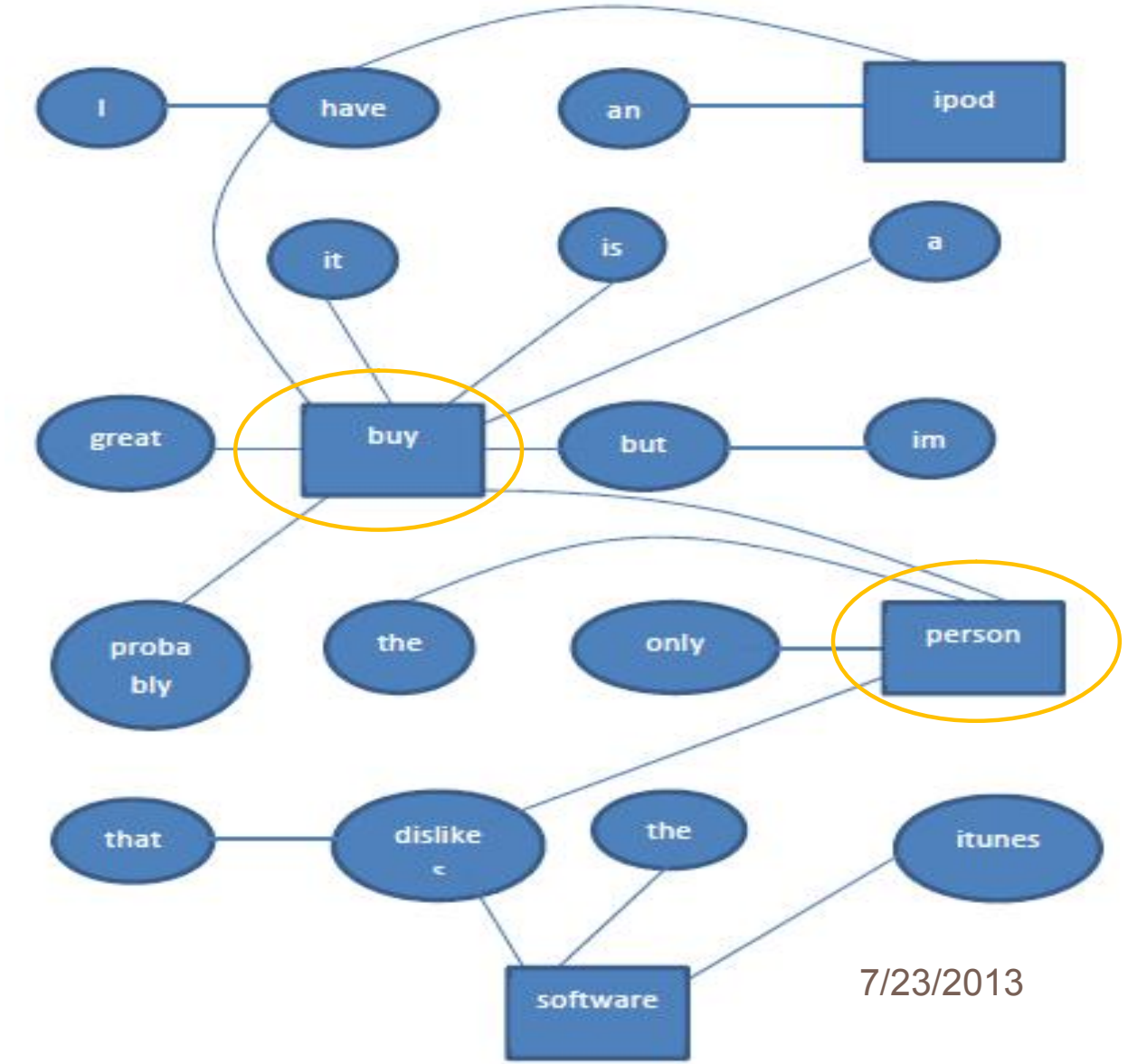
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

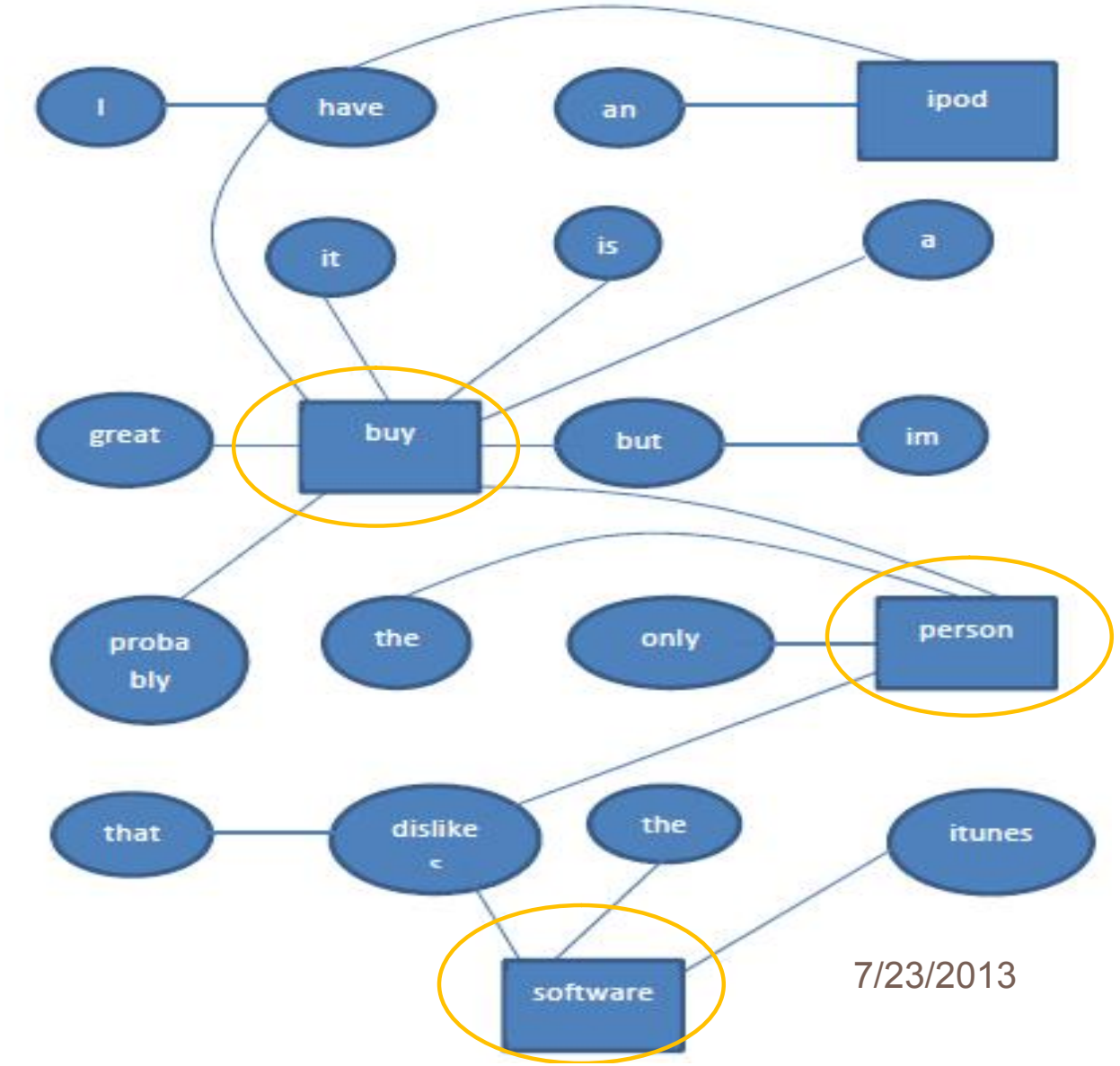
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

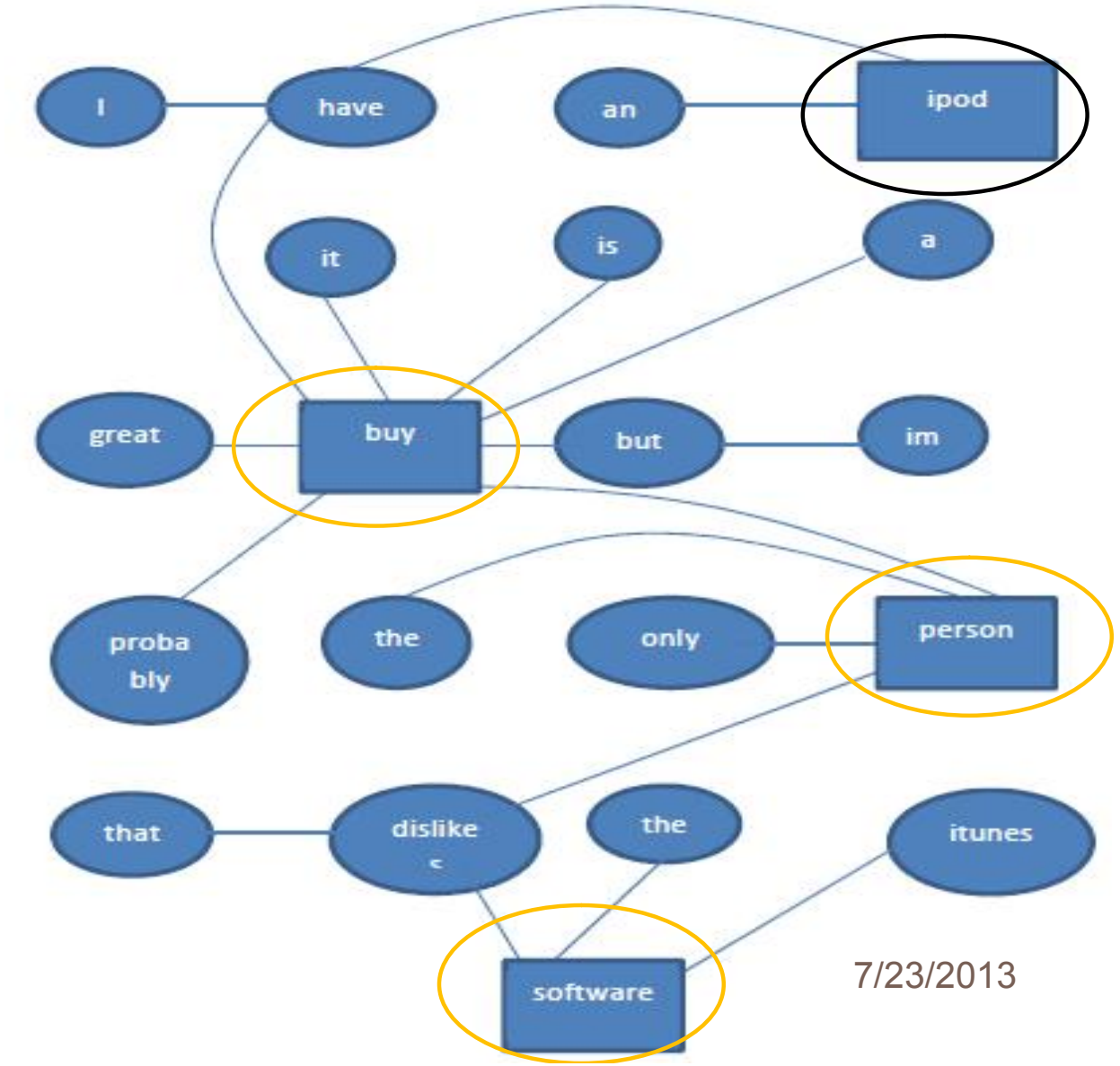
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

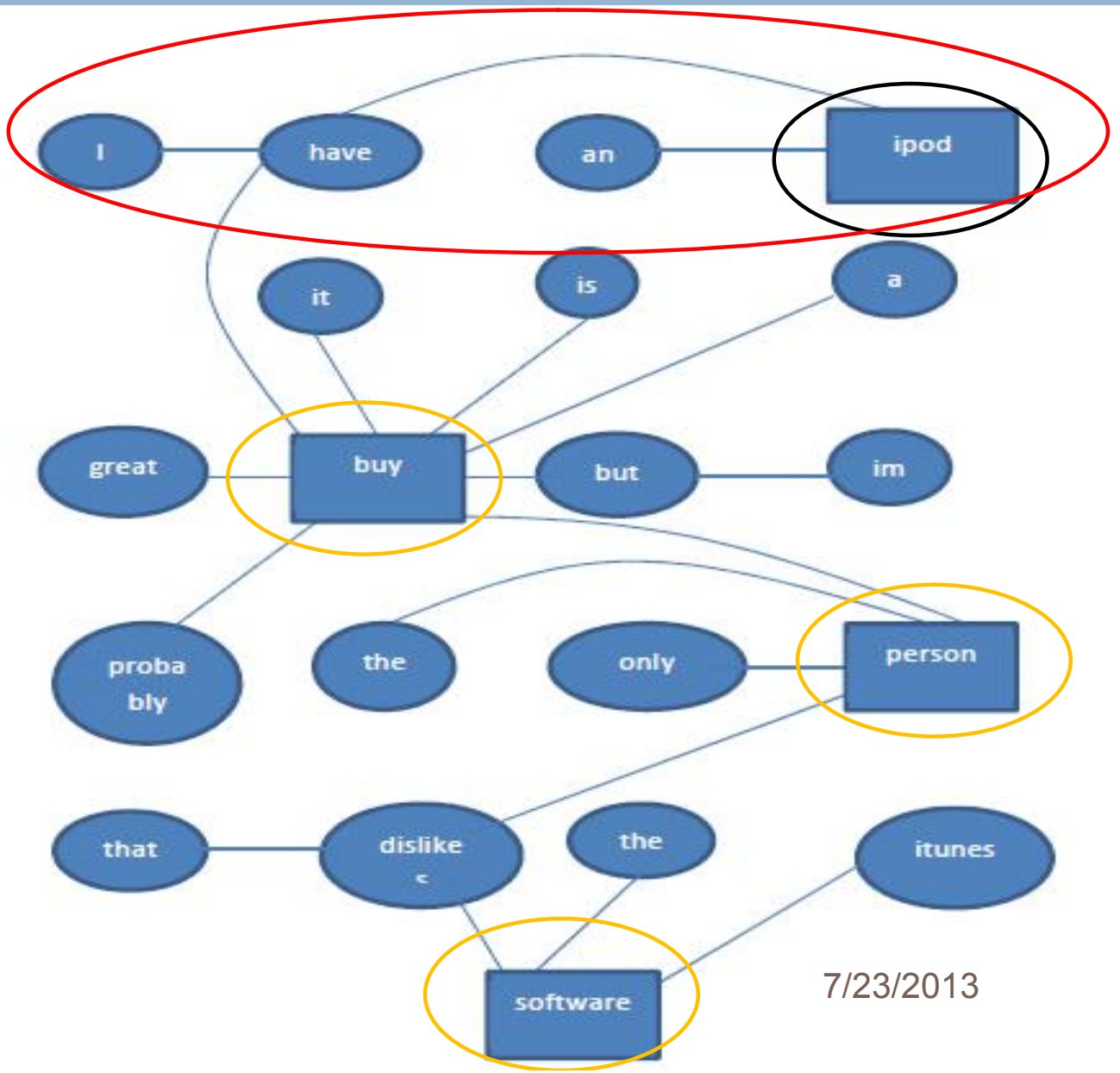
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

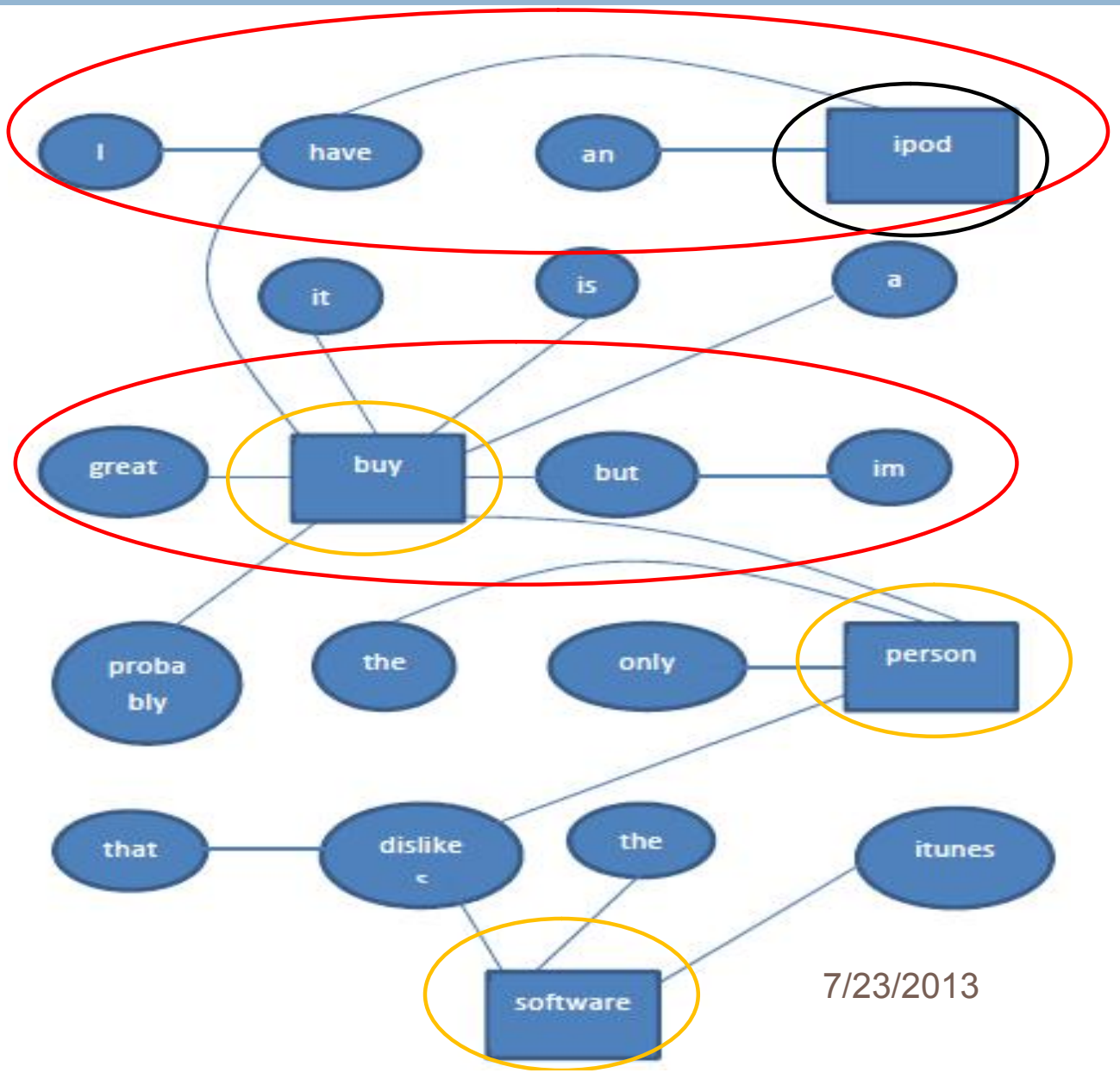
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

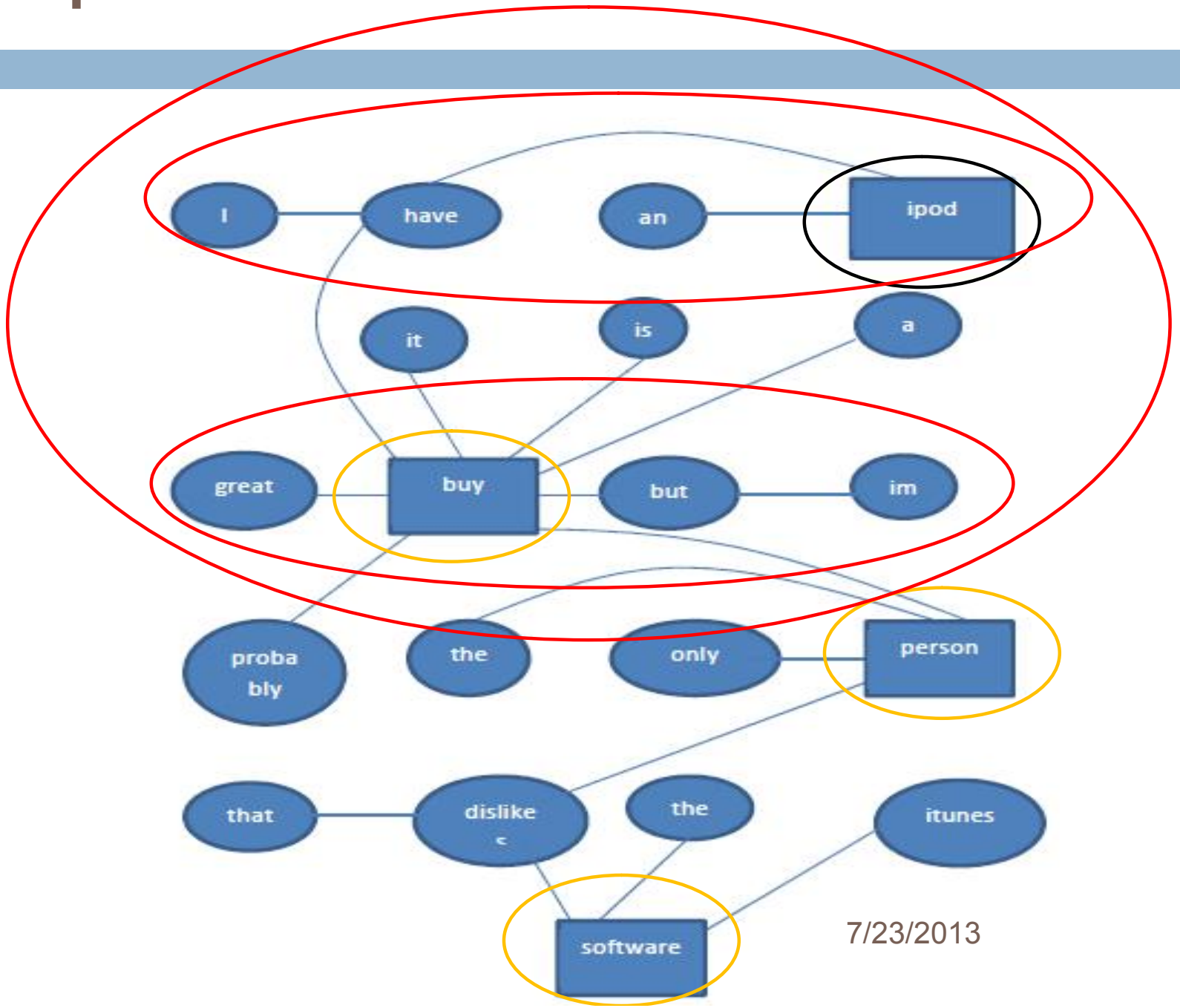
Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

Graph


I
have
an
ipod
it
is
a
great
buy
but
im
proba
bly
the
only
person
that
dislike
s
the
itunes
software

Classification and Datasets

Classification and Datasets

- Lexicon Classification

Classification and Datasets

- Lexicon Classification

Bing Liu Sentiment Lexicon for finding the polarity of the words in the target cluster (majority voting)

Classification and Datasets

- Lexicon Classification
- Supervised Classification

Bing Liu Sentiment Lexicon for finding the polarity of the words in the target cluster (majority voting)

Classification and Datasets

- Lexicon Classification
- Supervised Classification

Words in the target cluster as features (unigram bag-of-words)
SVM's as classifier

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)

Words in the target cluster as features (unigram bag-of-words)
SVM's as classifier

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)

Domains - camera, laptop, mobile, printer

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)

Domains - camera, laptop, mobile, printer

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)

Domains - antivirus, camera, dvd, ipod, music player, router, mobile

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)
- Baseline 1

Domains - antivirus, camera, dvd, ipod, music player, router, mobile

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)
- Baseline 1

Majority Voting on the number of positive and negative opinion words

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)
- Baseline 1
- Baseline 2

Majority Voting on the number of positive and negative opinion words

Classification and Datasets

- Lexicon Classification
- Supervised Classification
- Dataset 1 - Lakkaraju *et al*. (SDM 2011)
- Dataset 2 - Hu and Liu *et al.* (SIGKDD 2004)
- Baseline 1
- Baseline 2

Opinion word nearest to the target entity decides the polarity

Parameter Learning

Parameter Learning

- Dependency Parsing uses approx. 40 relations.

Parameter Learning

- Dependency Parsing uses approx. 40 relations.
  - Relation Space – $(2^{40} - 1)$

Parameter Learning

- Dependency Parsing uses approx. 40 relations.
  - Relation Space – $(2^{40} -1)$
- *Fix relations certain to be significant*
  - *nsubj, nsubjpass, dobj, amod, advmod, nn, neg*

Parameter Learning

- Dependency Parsing uses approx. 40 relations.
  - Relation Space – $(2^{40} - 1)$
- *Fix relations certain to be significant*
  - *nsubj, nsubjpass, dobj, amod, advmod, nn, neg*
- *Reject relations certain to be non-significant*
  - *copula, det, predet etc.*
- Compute Leave-One-Relation out accuracy over the remaining 21 relations over a development set
  - Find the relations for which there is ***significant accuracy change***
- Find inter-clusters distance threshold using development set

Ablation test

Ablation test

Relations	Accuracy (%)
All	63.5
Dep	**67.3**
Rcmod	**65.4**
xcomp, conj_and ccomp, iobj	61.5
advcl , appos, csubj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj	63.5

Ablation test

Relations	Accuracy (%)
All	63.5
Dep	**67.3**
Rcmod	**65.4**
xcomp, conj_and ccomp, iobj	61.5
advcl , appos, csubj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj	63.5

Relation Set	Accuracy
With Dep+Rcmod	66
Without Dep	**69**
Without Rcmod	67
Without Dep+Rcmod	68

Ablation test

Relations	Accuracy (%)
All	63.5
Dep	**67.3**
Rcmod	**65.4**
xcomp, conj_and ccomp, iobj	61.5
advcl , appos, csubj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj	63.5

Relation Set	Accuracy
With Dep+Rcmod	66
Without Dep	**69**
Without Rcmod	67
Without Dep+Rcmod	68

θ	Accuracy (%)
2	67.85
3	**69.28**
4	68.21
5	67.40

Lexicon based classification (Dataset 2)

Domain	Baseline 1 (%)	Baseline 2 (%)	Proposed System (%)
Antivirus	50.00	56.82	**63.63**
Camera 1	50.00	61.67	**78.33**
Camera 2	50.00	61.76	**70.58**
Camera 3	51.67	53.33	**60.00**
Camera 4	52.38	57.14	**78.57**
Diaper	50.00	63.63	**57.57**
DVD	52.21	63.23	**66.18**
IPOD	50.00	57.69	**67.30**
Mobile 1	51.16	61.63	**66.28**
Mobile 2	50.81	65.32	**70.96**
Music Player 1	50.30	57.62	**64.37**
Music Player 2	50.00	60.60	**67.02**
Router 1	50.00	58.33	**61.67**
Router 2	50.00	59.72	**70.83**

Overall Classification Accuracy (Dataset 2)

Lexicon-based Classification

Supervised Classification

Overall Classification Accuracy (Dataset 2)

Lexicon-based Classification

System	Accuracy (%)
$Baseline_1$	50.35
$Baseline_2$	58.93
Proposed System	**70.00**

Supervised Classification

Overall Classification Accuracy (Dataset 2)

Lexicon-based Classification

System	Accuracy (%)
$Baseline_1$	50.35
$Baseline_2$	58.93
Proposed System	**70.00**

Supervised Classification

Domain	Baseline 1 (%) Accuracy (Precision/Recall)	Proposed System (%) Accuracy (Precision/Recall)
Mobile	51.42 (50.72/99.29)	**83.82 (83.82/83.82)**
Camera	50	**86.99 (84.73/90.24)**

Lexicon Classification Results (Dataset 1, Lakkaraju *et al.*, SDM 2011)

System	Sentiment Evaluation Accuracy (%)
$Baseline_1$	68.75
$Baseline_2$	61.10
CFACTS-R	80.54
CFACTS	**81.28**
FACTS-R	72.25
FACTS	75.72
JST	76.18
Proposed System	**80.98**

Drawbacks

- Features should be explicitly present in review
  - *Mobile is heavy. (implicit feature - weight)*
- Cannot capture domain-specific implicit sentiment
  - *read* in *movie* domain and *book* domain

- Discourse Specific Sentiment Analysis
  - Sentiment Analysis in Twitter with Lightweight Discourse Analysis

Discourse Specific Sentiment Analysis

Slide Courtesy, Akshat Malu

Discourse Specific Sentiment Analysis

- Presence of words, at times, weighs more than the frequencies

Slide Courtesy, Akshat Malu

Discourse Specific Sentiment Analysis

- Presence of words, at times, weighs more than the frequencies
- Presence of a discourse marker can alter the overall sentiment of a sentence

*"The actors have done an okay job, not too brilliant, the script goes off track in between. Songs are mediocre. The direction could have been better. **But** overall, I like the movie"*

Slide Courtesy, Akshat Malu

Discourse Specific Sentiment Analysis

- Presence of words, at times, weighs more than the frequencies
- 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

*"The actors have done an okay job, not too brilliant, the script goes off track in between. Songs are mediocre. The direction could have been better. **But** overall, I like the movie"*

Slide Courtesy, Akshat Malu

Motivation Contd…

- Requirement of a **lightweight** method of discourse analysis for social media applications

- Use of heavy linguistic resources like parsing is not preferred
  - Increased processing time which slows down inter-active applications
  - Parsing does not work well in presence of noisy text

"*@user share 'em! i'm* **quite excited** *bout Tintin, despite* **not realy liking** *original comics. Probably because Joe Cornish had a hand in.*

Discourse Markers

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***

The direction was not that great, **but** *still we loved the movie.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***

The direction was not that great, **but** *still we loved the movie.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***

India managed to win **despite** *the initial setback.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***

India managed to win **despite** *the initial setback.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***

We were not much satisfied with the greatly acclaimed brand X and **subsequently** *decided to reject it.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals** ***(might, could, can, would, may)***
  - These express a higher degree of uncertainty

We were not much satisfied with the greatly acclaimed brand X and **subsequently** *decided to reject it.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty

That film might be good.

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly

That film might be good.

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly

I heard the movie is good, so you **must** *go to watch that movie.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly
- **Conditionals (*If)***

I heard the movie is good, so you **must** *go to watch that movie.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly
- **Conditionals (*If)***

If *I had studied well before the exams, I would have done well.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly
- **Conditionals (*If)***

If *I had studied well before the exams, I would have done well.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly
- **Conditionals (*If)***

I do **not** *like Nokia but I like Samsung.*

Discourse Markers

- Conjunctions that give more importance to the following discourse segment
  - ***(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless)***
- Conjunctions that give more importance to the previous discourse segment
  - **(*till, until, despite, in spite, though, although)***
- Conjunctions that tend to draw a conclusion or inference
  - ***(therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)***
- **Strong Modals *(might, could, can, would, may)***
  - These express a higher degree of uncertainty
- **Weak Modals (should, ought to, need not, shall, will, must)**
  - These express lesser degree of uncertainly
- **Conditionals (*If)***

I do **not** *like Nokia but I like Samsung.*

Algorithm

Algorithm

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

Algorithm

- *(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 are incremented by 1

- *The movie looked promising, but it failed to make an impact in the box-office*

Algorithm

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

- (*till, until, despite, in spite, though, although)*

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

- *The movie looked promising, but it failed to make an impact in the box-office*

Algorithm

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

- (*till, until, despite, in spite, though, although)*

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

- *India staged a marvelous victory down under despite all odds.*

Algorithm

- *(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)*
- (*till, until, despite, in spite, though, although)*
- All sentences containing *if* are marked

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

- *India staged a marvelous victory down under despite all odds.*

Algorithm

- *(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)*
- (*till, until, despite, in spite, though, although)*
- All sentences containing *if* are marked

- In supervised classifiers, their weights are decreased

Algorithm

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

- In supervised classifiers, their weights are decreased

Algorithm

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

- In supervised classifiers, their weights are decreased

Algorithm

- *(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)*
- (*till, until, despite, in spite, though, although)*
- All sentences containing *if* are marked
- All sentences containing *strong modals* are marked
- Negation

  - In supervised classifiers, their weights are decreased

Algorithm

- *(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)*
- (*till, until, despite, in spite, though, although)*
- All sentences containing *if* are marked
- All sentences containing *strong modals* are marked
- Negation

  - 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 Nokia* **but** *I like Samsung.*

Feature Space

- Lexeme Feature Space

Feature Space

- Lexeme Feature Space
  - Bag-of-words features

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i
  - Initially all w_i's are initialized to 1

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i
  - Initially all w_i's are initialized to 1
  - c_i is incremented with occurence of w_i in the doc

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i
  - Initially all w_i's are initialized to 1
  - c_i is incremented with occurence of w_i in the doc
  - c_i is also incremented according to w_i 's importance

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i
  - Initially all w_i's are initialized to 1
  - c_i is incremented with occurence of w_i in the doc
  - c_i is also incremented according to w_i 's importance
  - $\{w_i, c_i\}$ where c_i represents the weight of the term w_i.

Feature Space

- Lexeme Feature Space
  - Bag-of-words features
  - Maintain a count c_i of each word w_i
  - Initially all w_i's are initialized to 1
  - c_i is incremented with occurence of w_i in the doc
  - c_i is also incremented according to w_i 's importance
  - $\{w_i, c_i\}$ where c_i represents the weight of the term w_i.

- Sense Feature Space

  s_i instead of w_i

Classification of features

- Rule based System

Classification of features

- Rule based System
  - Bing Liu Sentiment Lexicon (Hu *et al.*, 2004)

Classification of features

- Rule based System
  - Bing Liu Sentiment Lexicon (Hu *et al.*, 2004)

- Supervised Classification
  - SVM on feature vectors $\{w_i , c_i\}$ or $\{s_i , c_i\}$
  - Features
    - N-grams (N=1,2)
    - Stop Word Removal (except discourse markers)
    - Discourse Weight of Features
    - Modal and Conditional Indicators
    - Stemming
    - Negation
    - Emoticons
    - Part-of-Speech Information
    - Feature Space (Lexeme or Synset)

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

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

- Negative hashtags - *#negative, #sad, #depressed, #gloomy,#disappointed*

Datasets

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

- Negative hashtags - *#negative, #sad, #depressed, #gloomy,#disappointed*

- 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 class (positive and negative)

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

Classification Results in Twitter (Datasets 1 and 2)

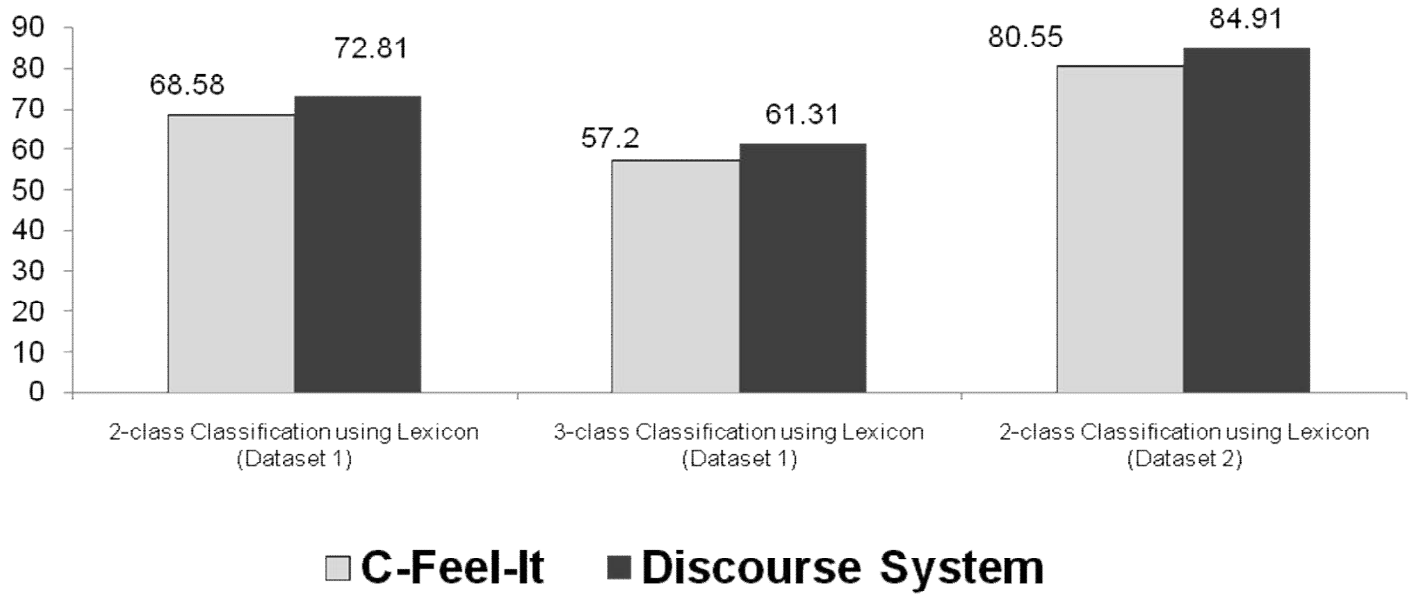
Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

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

Lexicon-based Classification

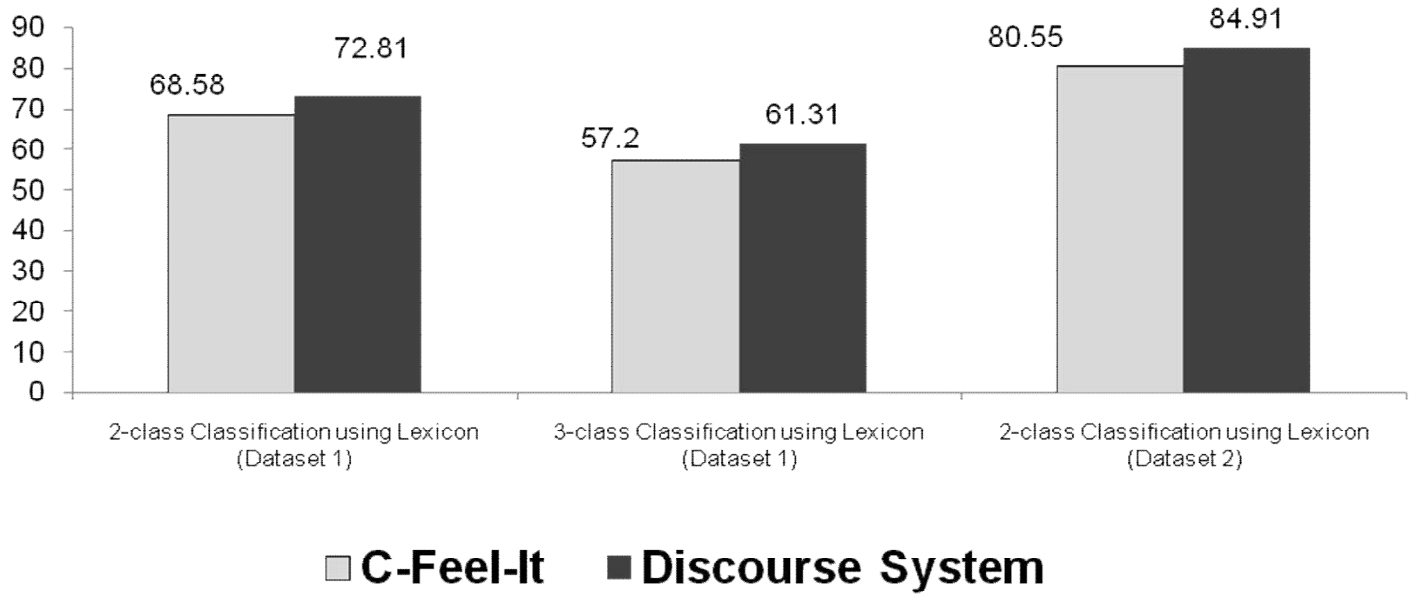
Classification Results in Twitter (Datasets 1 and 2)
Comparison with C-Feel-It (Joshi *et al.*, ACL 2011)

Lexicon-based Classification



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

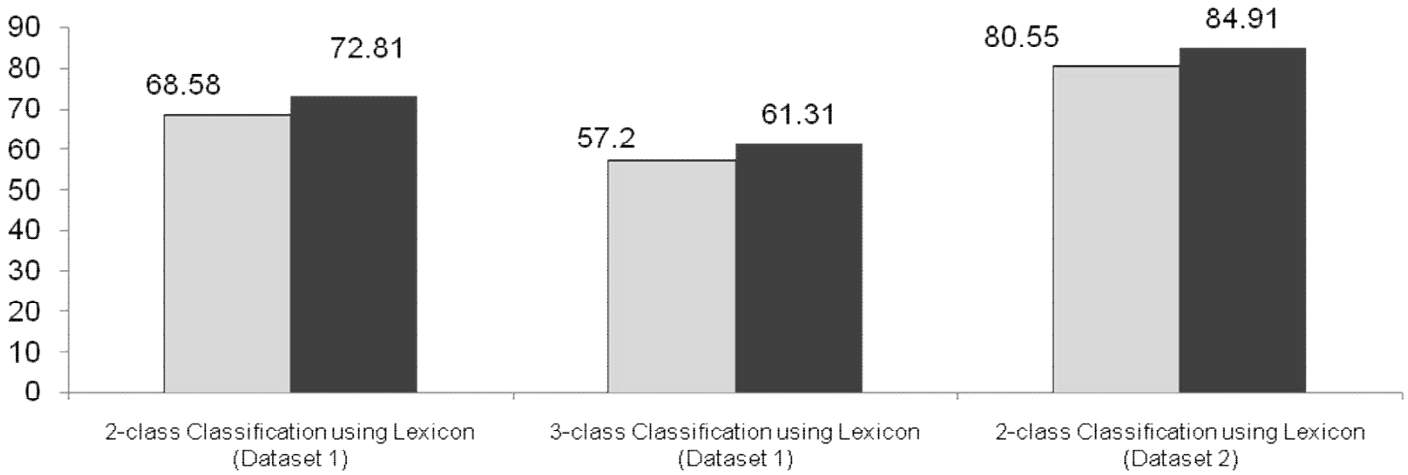
Lexicon-based Classification



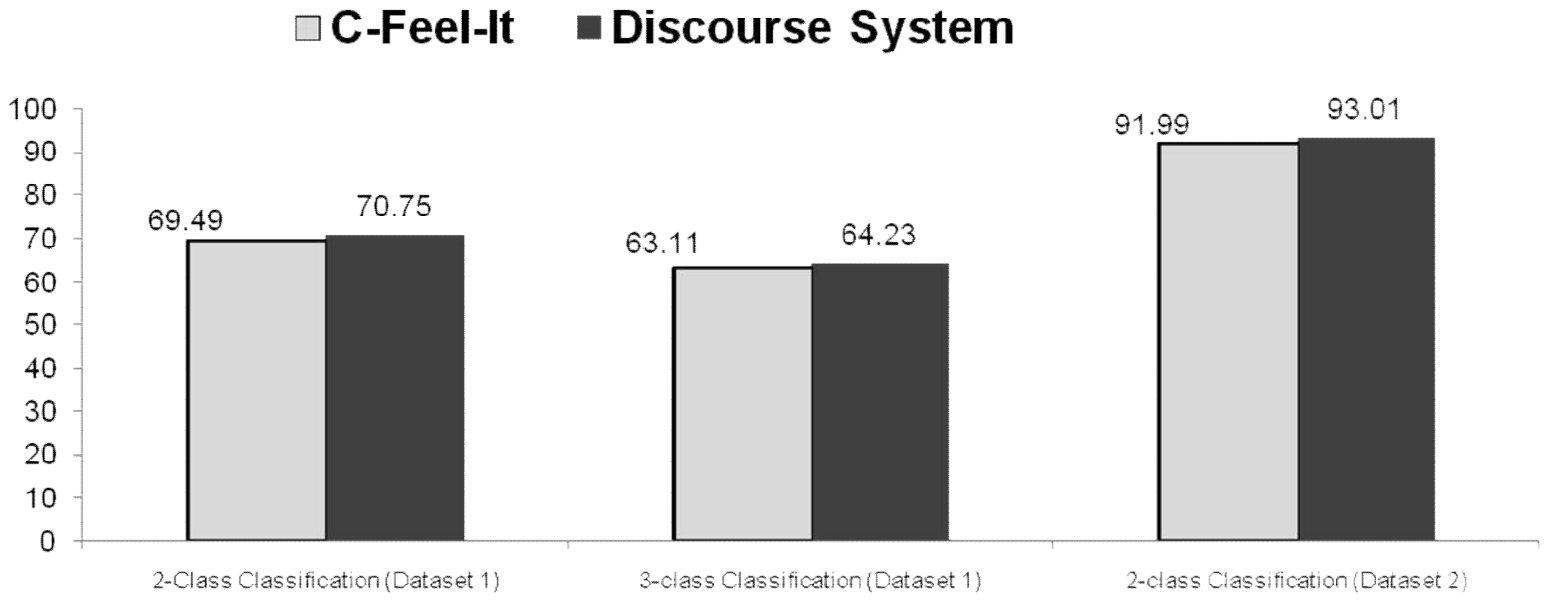
Supervised Classification

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)

Supervised Classification

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

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
- Scope of discourse markers

Positional Importance: Back-up Slide

- "*I wanted to follow my dreams and ambitions despite all the obstacles, but I did not succeed.*"
- *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*
- Overall polarity *+1*, whereas the overall sentiment should be *negative*
- Reason:
  - 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* and *pragmatic* importance.

Role of World Knowledge in Sentiment Analysis

- WikiSent: Unsupervised Sentiment Analysis System Using Extractive Summarization With Wikipedia

World Knowledge

- Extensive world knowledge required to perform analysis of reviews
- Distinguish between **What's** and **About's** of the review
- Filter out concepts irrelevant to the reviewer opinion about the movie
- Retain only the objects of interest and corresponding opinions

A Negative Review

- **best remembered** for his understated **performance** as dr. hannibal lecter in michael mann's forensics thriller , manhunter , scottish character actor brian cox brings something **special** to every movie he works on .
- usually playing a bit role in some studio schlock (he dies halfway through the long kiss goodnight) , he's only occasionally given something **meaty** and **substantial** to do .
- if you want to see some **brilliant** acting , check out his work as a dogged police inspector opposite frances mcdormand in ken loach's hidden agenda .
- cox plays the role of big john harrigan in the disturbing new indie flick l . i . e . , which lot 47 picked up at sundance when other distributors were scared to budge
- big john feels the **love** that dares not speak its name , but he expresses it through seeking out adolescents and bringing them back to his pad .
- what bothered some audience members was the presentation of big john in an oddly empathetic light .
- he's an **even-tempered , funny , robust** old man who actually listens to the kids' problems (as opposed to their parents and friends , both caught up in the high-wire act of their own confused lives .)
- he'll have sex-for-pay with them only after an elaborate courtship , **charming** them with temptations from the grown-up world .

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *John Travolta is considered by many to be a has-been, or a one-hit wonder …*
- *Leornardo DeCaprio is an awesome actor.*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *Born into a family of thespians -- parents Roger Winslet and Sally Bridges-Winslet were both stage actors, maternal grandparents Oliver and Linda*
- *Bridges ran the Reading Repertory Theatre, and uncle Robert Bridges was a fixture in London's West End theatre district, Kate Winslet came into her talent at an early age.*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *The role that transformed Winslet from art house attraction to international star was Rose DeWitt Bukater, the passionate, rosy-cheeked aristocrat in James Cameron's Titanic (1997).*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *I cancelled the date with my girlfriend just to watch my favorite star featuring in this movie.*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *L.I.E. stands for Long Island Expressway, which slices through the strip malls and middle-class homes of suburbia. Filmmaker Michael Cuesta uses it as a (pretty transparent) metaphor of dangerous escape for his 15-year old protagonist, Howie (Paul Franklin Dano).*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *He's an even-tempered, funny, robust old man who actually listens to the kids' problems (as opposed to their parents and friends, both caught up in the high-wire act of their own confused lives.).*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *Horror movies are supposed to be scary.*
- *There is an axiom that directors who have a big hit with their debut have a big bomb with their second film.*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *While the movie is brutal, the violence is neither very graphic nor gratuitous. It may scare the little ones, but the teen-age audience for which it is aimed will appreciate the man-eating chomping that runs through the film.*

Facets of a Movie Review

- General Perception about the Crew
- Objective Facts about the Crew and Movies
- Past Performance of the Crew and Movies
- Expectations from the Movie or Crew
- Movie Plot
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- **Opinion about the Movie and Crew**
- Unrelated Category

- *So my grandson gives me passes to this new picture One Night at McCool's because the free screening is the same night as that horrible show with those poor prisoners trapped on the island who eat the bugs. "Go," he says, "it's just like Rush-o-Man."*

Wikipedia

- Extensive World Knowledge required to perform this analysis
- Wikipedia is used to create a *topic-specific, extractive summary* of a review
- The extract is classified with a Lexicon, instead of the entire review

Feature Extraction from Wikipedia

Feature Extraction from Wikipedia

METADATA

Harry Potter and the Deathly Hallows – Part 1 is a 2010 fantasy film[5] directed by David Yates and the first of two films based on the novel Harry Potter and the Deathly Hallows by J. K. Rowling. It is the seventh instalment in the Harry Potter film series, written by Steve Kloves and produced by David Heyman, David Barron and Rowling. The story follows Harry Potter on a quest to find and destroy Lord Voldemort's secret to immortality – the Horcruxes. The film stars Daniel Radcliffe as Harry Potter, alongside Rupert Grint and Emma Watson as Harry's best friends Ron Weasley and Hermione Granger. It is the sequel to Harry Potter and the Half-Blood Prince and is followed by the concluding film, Harry Potter and the Deathly Hallows – Part 2.

Principal photography began on 19 February 2009 and was completed on 12 June 2010.[6] Part 1 was released in 2D cinemas and IMAX formats worldwide on 19 November 2010.

Feature Extraction from Wikipedia

PLOT

Further information: Harry Potter and the Deathly Hallows Novel Plot

Minister Rufus Scrimgeour addresses the wizarding media stating that the Ministry of Magic will remain strong as Lord Voldemort gains power throughout the wizarding and Muggle worlds. Severus Snape arrives at Malfoy Manor to inform Lord Voldemort and his Death Eaters of Harry's departure from No. 4 Privet Drive. Voldemort commandeers Lucius Malfoy's wand, as Voldemort's own wand cannot be used to kill Harry; their wands are "twins".

Meanwhile, the Order of the Phoenix arrive at Privet Drive and escort Harry to safety using Polyjuice Potion to create six decoy Harrys. During their flight to the Burrow, they are ambushed by Death Eaters, who kill Mad-Eye Moody and Hedwig, and injure George Weasley. They arrive at the Burrow, where Harry has a vision of Ollivander being tormented by Voldemort, who claims that the wand-maker had lied to him by informing him of the only way to kill Harry: obtaining another's wand.

Feature Extraction from Wikipedia

CREW

Directed by : David Yates
Produced by : David Heyman,David Barro, J. K. Rowling
Screenplay by : Steve Kloves Based on Harry Potter and the Deathly Hallows by J. K. Rowling
Starring : Daniel Radcliffe, Rupert Grint, Emma Watson
Music by : Alexandre Desplat
Themes: John Williams
Cinematography : Eduardo Serra
Editing by: Mark Day
Studio :Heyday Films
Distributed by: Warner Bros. Pictures

that the wand-maker had lied to him by informing him of the only way to kill Harry: obtaining another's wand.

Feature Extraction from Wikipedia

CREW

Directed by : David Yates
Produced by : David Heyman,David Barro, J. K. Rowling
Screenplay by : Steve Kloves Based on Harry Potter and the Deathly Hallows by J. K. Rowling
Starring : Daniel Radcliffe, Rupert Grint, Emma Watson
Music by : Alexandre Desplat
Themes: John Williams
Cinematography : Eduardo Serra
Editing by: Mark Day
Studio :Heyday Films
Distributed by: Warner Bros. Pictures

Narcissa Malfoy.
Ralph Fiennes as Lord Voldemort, the film's main antagonist

Feature Extraction from Wikipedia

DOMAIN SPECIFIC FEATURE LIST

Movie, Staffing, casting, Writing, Theory, Writing, Rewriting, Screenplay, Format, Treatments, Scriptments, Synopsis, Logline, Pitching, Certification,scripts, Budget, Ideas, Funding, budgeting, Funding, Plans, Grants, Pitching, Tax, Contracts, law, Copyright, Pre-production, Budgeting, Scheduling, Pre-production, film , stock, Story, boarding, plot, Casting , Directors, Location, Scouting, …..

Ralph Fiennes as Lord Voldemort, the film's main antagonist

Feature List Creation

Feature List Creation

- Metadata and Plot sentences are POS-tagged
  - Nouns retrieved
  - Stemmed
  - Added to the **Plot** list
  - Character information also added to **Plot** list

Feature List Creation

- Metadata and Plot sentences are POS-tagged
  - Nouns retrieved
  - Stemmed
  - Added to the **Plot** list
  - Character information also added to **Plot** list

- Domain specific feature list extracted from Wiki articles on films and movies
  - Added to **MovieFeature** list

Feature List Creation

- Metadata and Plot sentences are POS-tagged
  - Nouns retrieved
  - Stemmed
  - Added to the **Plot** list
  - Character information also added to **Plot** list
- Domain specific feature list extracted from Wiki articles on films and movies
  - Added to **MovieFeature** list
- Crew information added to **Crew** list

Feature List Creation

- Metadata and Plot sentences are POS-tagged
  - Nouns retrieved
  - Stemmed
  - Added to the **Plot** list
  - Character information also added to **Plot** list
- Domain specific feature list extracted from Wiki articles on films and movies
  - Added to **MovieFeature** list
- Crew information added to **Crew** list
- Laptop and Printer domain data used to filter frequent occurring concepts
  - Non-overlapping domains
  - Frequently occurring terms in all the domains added t**o FreqWords** list
    The **FreqWords** are pruned from the other feature lists

Why only Nouns?

- To restrict genre-specific concepts to be entities
- *Harry acted as if nothing has happened* vs. *Kate Winslet acted awesome in the movie.*
- Here *act* is present as a Verb in both the sentences
  - First sentence belongs to the *plot* (*Category 5*)
  - Second sentence depicts the reviewer opinion (*Category 8*)
- Difference lies in the presence of different subjects of interest with the Verb
- Our focus is to capture the *subjects* and *objects* in the sentence which give direct clues about the *category* of the reviewer statement, so that feature lists are as pure as possible

Extractive Summary

- Given a review R with n sentences S_i, determine if each sentence S_i is to be *accepted* or *rejected* based on a *relevancy factor* in judging this movie

- $$Rel_{factor_i} = 2\sum_j 1_{w_{ij} \in Crew\ or\ MovieTitle} + \sum_j 1_{w_{ij} \in MovieFeature} - \sum_j 1_{w_{ij} \in Plot, w_{ij} \notin Crew, w_{ij} \notin MovieTitle}$$

- $$Acc_{factor_i} = 1\ \boldsymbol{if}\ Rel_{factor_i} \geq 0\ and\ \exists w_{ij} \in S_i\ \boldsymbol{s.t.}\ w_{ij} \in Crew\ or\ MovieFeature\ or\ Movie\ Title$$
$$= 0\ otherwise$$

Semi-Supervised Learning of Parameters

- *Equations 2* can be re-written as:
- $Rel_{factor_i} = \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3}$
- $Acc_{factor_i} = Rel_{factor_i} - \theta$
- $= \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3} - \theta$
- $= \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3} - \theta \times X_{i,4}$ (where $X_{i,4} = 1$)
- Let Y_i be the binary label information corresponding to each sentence in the development set, where $Y_i=1$ if $Acc_{factor_i} \geq 0$ and -1 otherwise.
- $Y_i = \boldsymbol{W}.X_i$ where, $\boldsymbol{W} = [\alpha\ \beta\ -\gamma\ -\theta]^T$ *and* $X_i = [X_{i,1}\ X_{i,2}\ X_{i,3}\ X_{i,4}]$
- $or, \boldsymbol{Y} = \boldsymbol{W^T}.\boldsymbol{X}$

Algorithm

Algorithm

Input : Review R

Output: OpinionSummary

Step 1: Extract the Crew list from Wikipedia

Step 2: Extract the Plot list from Wikipedia

Step 3: Extract the MovieFeature list from Wikipedia

Step 4: Extract the FreqWords list as the common frequently occurring concepts in Mobile Phone, Printer and Movie domains.

Let $OpinionSummary = \emptyset$

for i=1..n

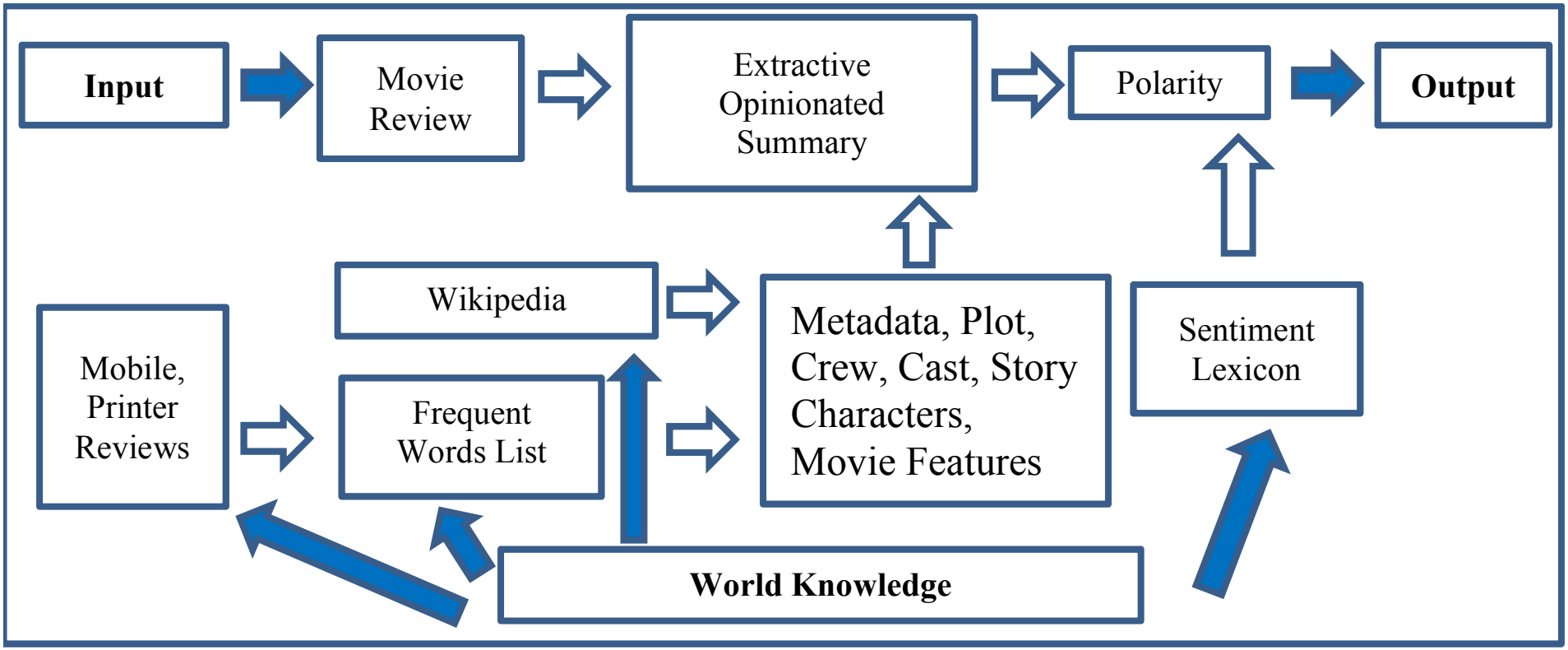
  if $Acc_{factor_i} == 1$

    add S_i *to* $OpinionSummary$

  end if

end for

WikiSent


Input
Movie Review
Extractive Opinionated Summary
Polarity
Output
Wikipedia
Metadata, Plot, Crew, Cast, Story Characters, Movie Features
Sentiment Lexicon
Mobile, Printer Reviews
Frequent Words List
World Knowledge

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

*Best remembered for his understated performance as Dr. Hannibal Lecter in Michael Mann's forensics thriller, Manhunter, Scottish character actor **Brian Cox** brings something special to every movie he works on.*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

Usually playing a bit role in some studio schlock (he dies halfway through The Long Kiss Goodnight), he's only occasionally given something meaty and substantial to do.

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

*If you want to see some brilliant **acting**, check out his work as a dogged police inspector opposite Frances McDormand in Ken Loach's Hidden Agenda.*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

***Cox** plays the role of Big **John Harrigan** in the disturbing new indie flick **L.I.E.**, which Lot 47 picked up at Sundance when other **distributors** were scared to budge.*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

Big John *feels the love that dares not speak its name, but he expresses it through seeking out adolescents and bringing them back to his pad.*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

*What bothered some **audience** members was the presentation of **Big John** in an oddly empathetic light.*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

*He's an even-**tempered**, funny, robust old man who actually listens to the kids' problems (as opposed to their parents and **friends**, both caught up in the high-wire act of their own confused lives.).*

Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the *Cast* list. Rel_{factor_i}=2*1+1*0-0*0=2 > -1 and the sentence is accepted.
- In [2], there is no keyword from the lists and it is rejected.
- [3] has the keyword **acting** from MovieFeature and is accepted.
- [4] has the keywords **Cox**, **L.I.E** from *Cast*, *MovieTitle*, **John Harrigan** from *Character* list and **distributor** from *MovieFeature* list. Rel_{factor_i}=2*2+1-1=4 > -1 and is accepted.
- [5] has only the keyword **Big John** from *Character* and Rel_{factor_i}=0+0-1=-1 and is rejected.
- [6] has the keyword **audience** from *MovieFeature* and **Big John** from *Character* and its Rel_{factor_i}=0+1-1=0 > -1 and is accepted.
- [7] has the keywords **temper**, **friend** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.
- [8] has the keywords **sex**, **charm** from *Plot* and Rel_{factor_i}=0+0-2=-2 and is rejected.

*He'll have **sex**-for-pay with them only after an elaborate courtship, **charming** them with temptations from the grown-up world"*

Lexicons and Datasets

Lexicons and Datasets

- Lexicons
  - SentiWordNet (Esuli *et al.*, 2006)
  - Subjectivity Lexicon (Wilson *et al.*, 2005)
  - General Inquirer (Stone *et al.*, 1966)

Lexicons and Datasets

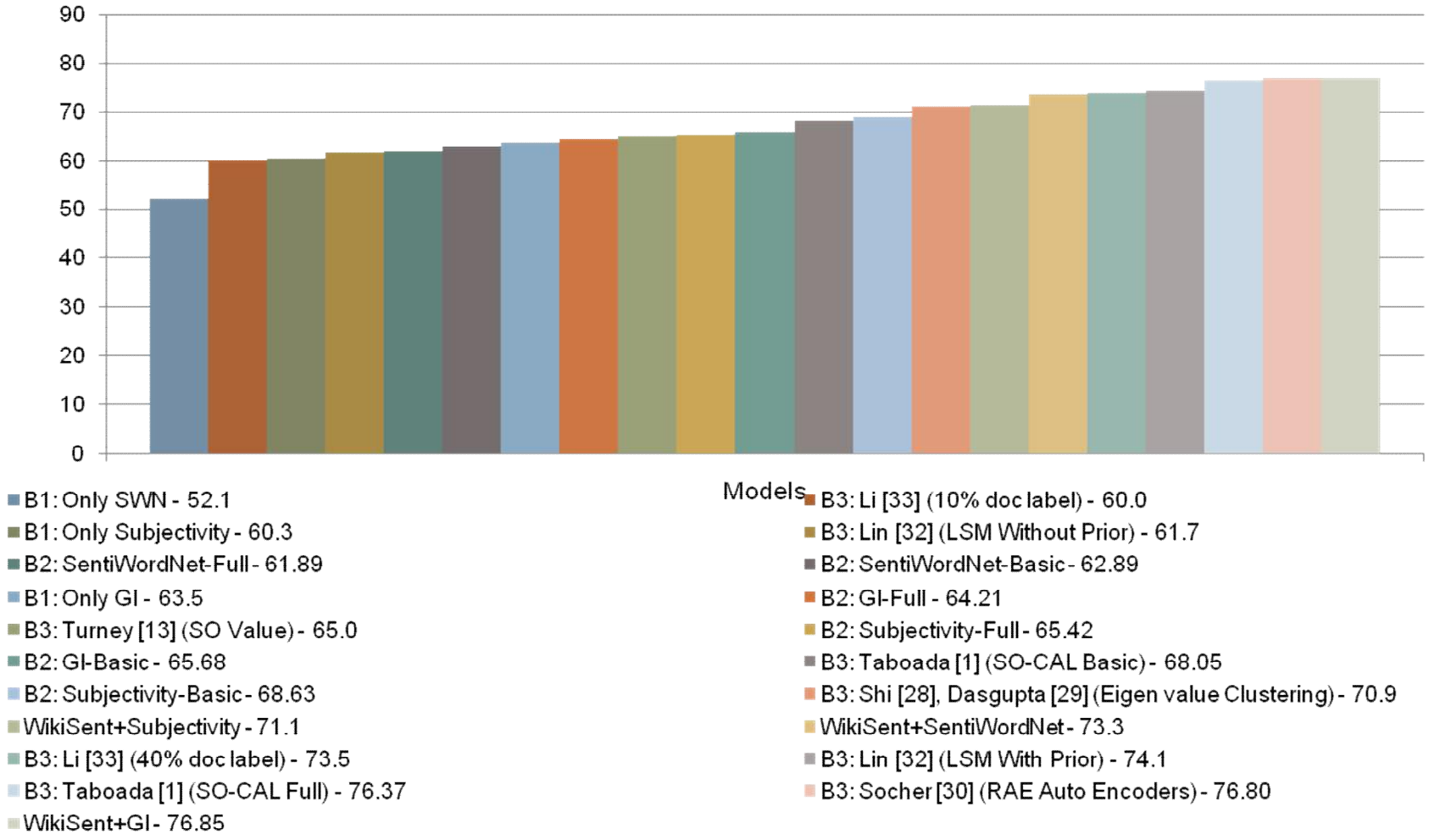
- Lexicons
  - SentiWordNet (Esuli *et al.*, 2006)
  - Subjectivity Lexicon (Wilson *et al.*, 2005)
  - General Inquirer (Stone *et al.*, 1966)

- Baseline 1
  - Bag-of-Words based on the 3 lexicons
- Baseline 2
  - SO-CAL (Taboada *et al.*, 2011)
- Baseline 3
  - All the semi-supervised and unsupervised systems in the domain

Lexicons and Datasets

- Lexicons
  - SentiWordNet (Esuli *et al.*, 2006)
  - Subjectivity Lexicon (Wilson *et al.*, 2005)
  - General Inquirer (Stone *et al.*, 1966)
- Baseline 1
  - Bag-of-Words based on the 3 lexicons
- Baseline 2
  - SO-CAL (Taboada *et al.*, 2011)
- Baseline 3
  - All the semi-supervised and unsupervised systems in the domain
- Movie Review Dataset (Pang *et al.*, 2002)
  - 1000 positive and 1000 negative reviews (labeled at document level)
    27,000 untagged reviews

Accuracy Comparison with All the Semi-supervised and Unsupervised Systems in the Same Dataset



Accuracy Comparison with Different Baselines

Accuracy Comparison with Different Baselines

Baseline 1 : Simple Bag-of-Words	
Only SentiWordNet	52.1
Only Subjectivity	60.3
Only GI	63.5
Baseline 2 : Worse, Median and Best Performing Lexicons with SO-CAL	
SentiWordNet-Full	61.89
SentiWordNet-Basic	62.89
GI-Full	64.21
GI-Basic	65.68
Subjectivity-Full	65.42
Subjectivity-Basic	68.63
WikiSent with Different Lexicons	
WikiSent+Subjectivity	71.1
WikiSent+SentiWordNet	73.3
WikiSent+GI	**76.85**

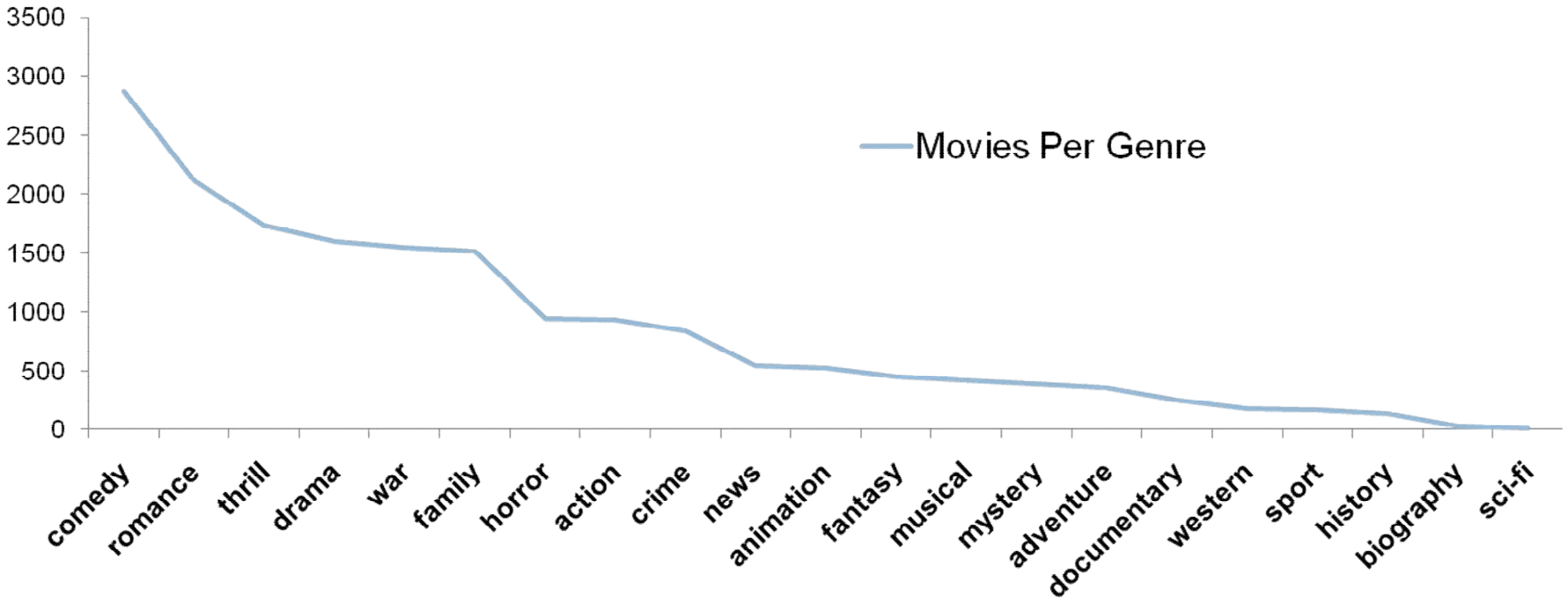
Accuracy Comparison with Different Baselines

Systems	Classification Method	Accuracy
Turney SO value	Unsupervised (PMI)	65
Taboada SO-CAL Basic [1]	Lexicon Generation	68.05
Taboada SO-CAL Full [1]	Lexicon Generation	76.37
Shi [28], Dasgupta [29]	Unsupervised Eigen Vector Clustering	70.9
Socher [30] RAE	Semi Supervised Auto Encoders	76.8
Lin [32] LSM	Unsupervised without prior info	61.7
Lin [32] LSM	Weakly Supervised with prior info	74.1
Li [33]	Semi Supervised 10% doc. Label	60
Li [33]	Semi Supervised 40% doc. Label	60
WikiSent	**Wikipedia+GI Lexicon**	**76.85**

Trend Analysis

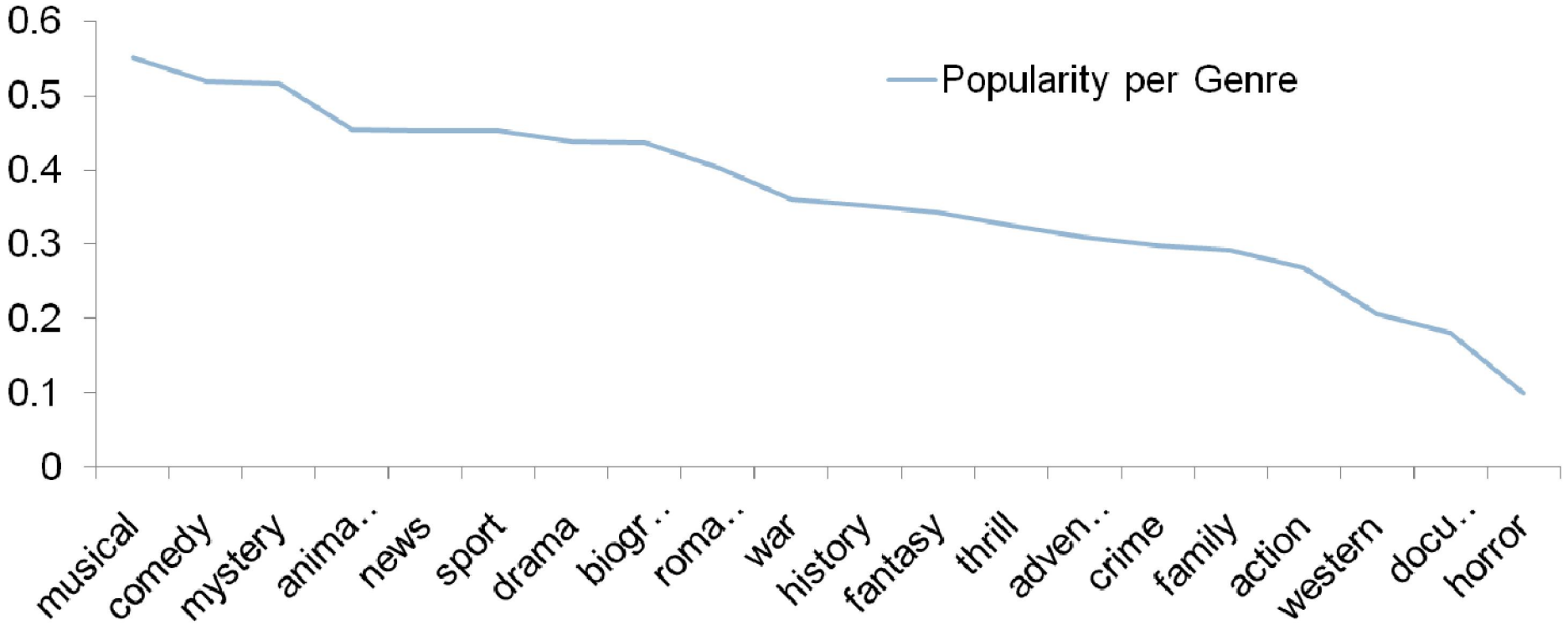
$$Genre\ Popularity = \frac{Positive\ Movie\ Reviews\ per\ Genre}{Total\ Movie\ Reviews\ per\ Genre}$$

Trend Analysis



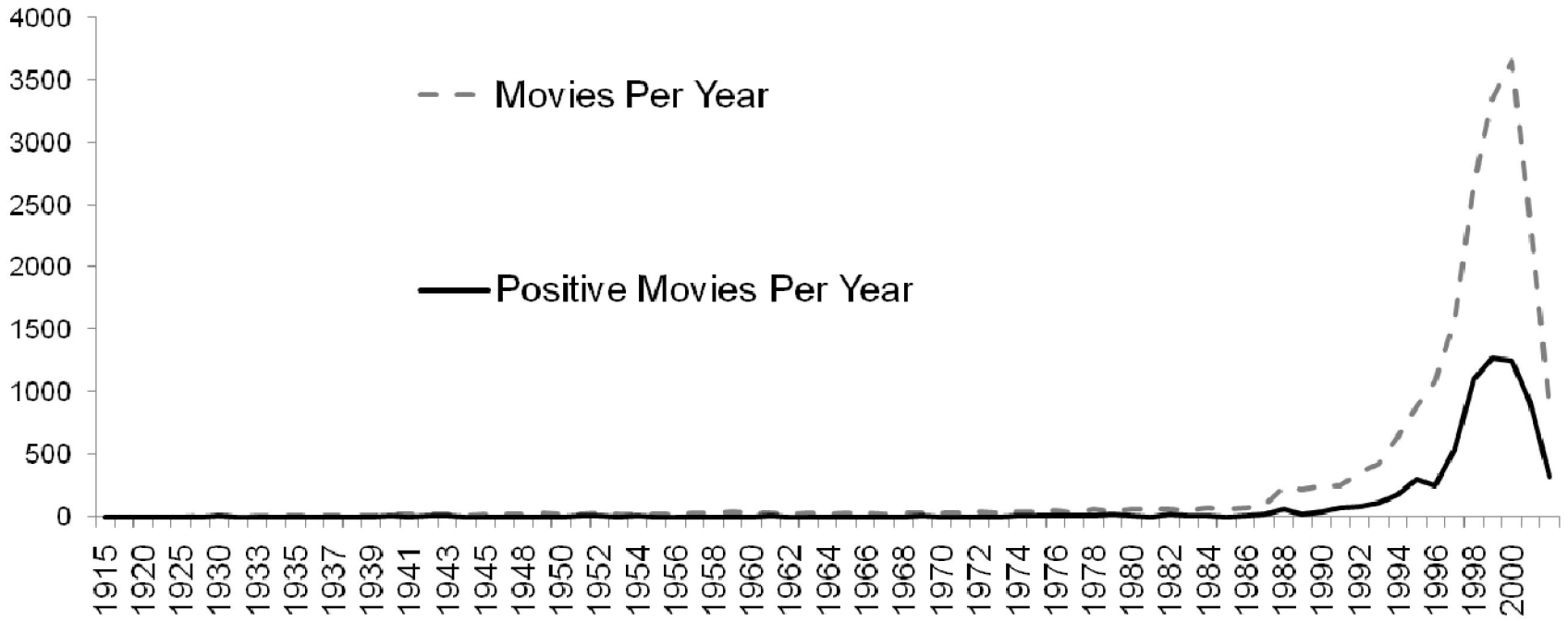
$$Genre\ Popularity = \frac{Positive\ Movie\ Reviews\ per\ Genre}{Total\ Movie\ Reviews\ per\ Genre}$$

Trend Analysis



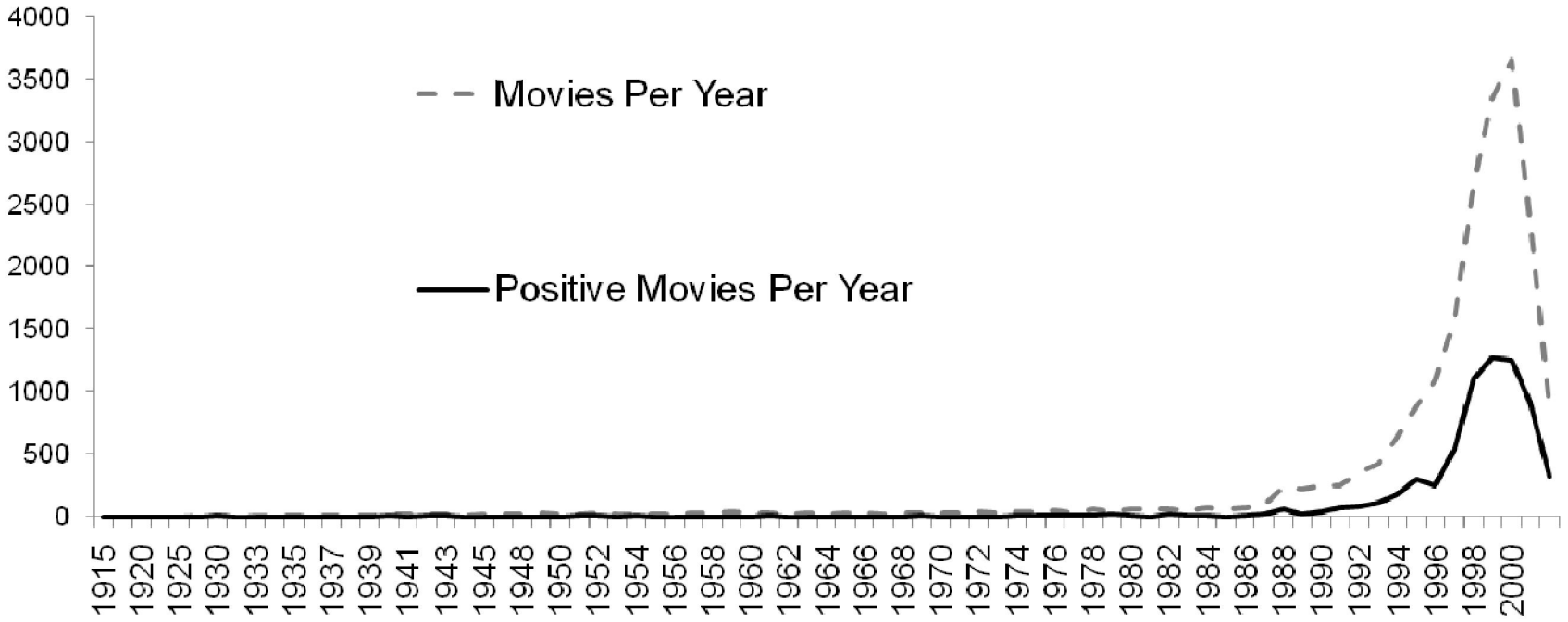
$$Genre\ Popularity = \frac{Positive\ Movie\ Reviews\ per\ Genre}{Total\ Movie\ Reviews\ per\ Genre}$$

Trend Analysis



$$Genre\ Popularity = \frac{Positive\ Movie\ Reviews\ per\ Genre}{Total\ Movie\ Reviews\ per\ Genre}$$

Trend Analysis



	WikiSent	Bag-of-Words Baseline 1
Positive Reviews (%)	48.95	81.2
Negative Reviews (%)	51.05	18.79

Drawbacks

- Absence of a co-reference resolution module
  - *false negative*
- Synonymous concepts not handled
  - Does not matter much as genre-specific concepts like *acting*, *direction*, *story-writer* occur in same lexical form
- Reviewer opinion bias can affect the system
  - *false positive*
- Absence of WSD module affect lexicon-based classification

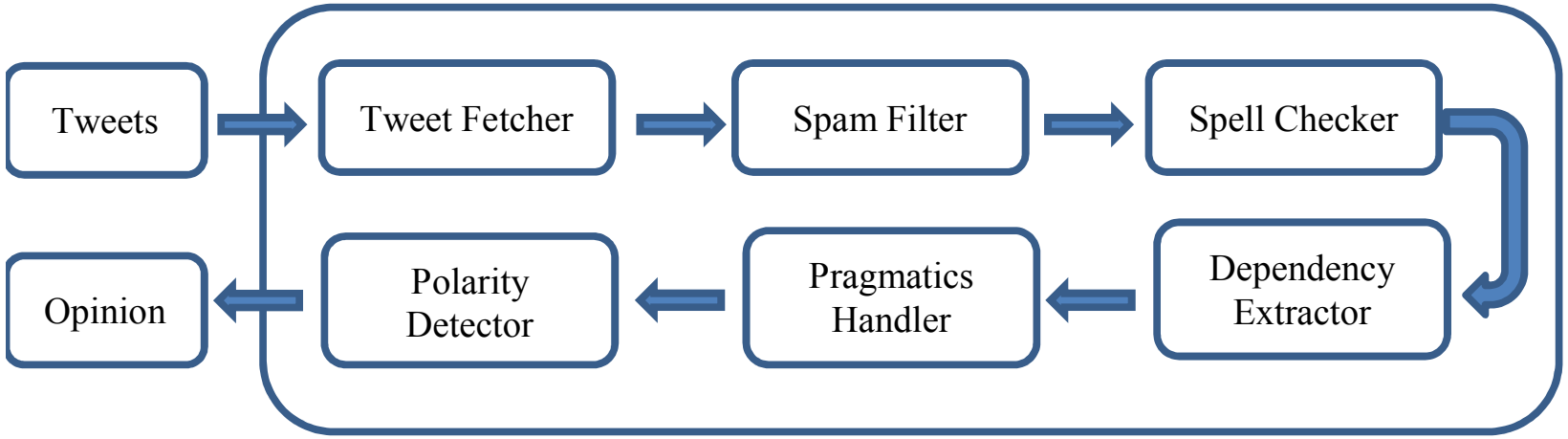
Role of Social Media Content and Informal Language Form in Sentiment Analysis

Social Media Analysis

- Social media sites, like Twitter, generate around 250 million tweets daily
- This information content could be leveraged to create applications that have a social as well as an economic value
- Text limit of 140 characters per tweet makes Twitter a noisy medium
  - Tweets have a poor syntactic and semantic structure
  - Problems like *slangs, ellipses, nonstandard vocabulary etc.*
- Problem is compounded by increasing number of *spams* in Twitter
  - P*romotional* tweets, *bot-generated* tweets, *random links* to websites etc.
  - In fact Twitter contains around 40% tweets as pointless babble

Had Hella fun today with the team. Y'all are hilarious! &Yes, i do need more black homies.....

TwiSent: Multi-Stage System Architecture



Spam Categorization and Features

- *Re-tweets*
- *Promotional tweets for some entity*
- *Tweets containing links to some other websites*
- *Tweets in languages other than English*
- *Tweets with incomplete text*
- *Automatically generated tweets by bots*
- *Tweets built primarily for search engines or tweets with excessive off-topic keywords*
- *Multiple tweets offering substantially the same content*

1. *Number of Words per Tweet*	**9.** *Frequency of Foreign Words*
2. *Average Word Length*	**10.** *Validity of First Word*
3. *Frequency of "?" and "!"*	**11.** *Presence / Absence of links*
4. *Frequency of Numeral Characters*	**12.** *Frequency of POS Tags*
5. *Frequency of hashtags*	**13.** *Strength of Character Elongation*
6. *Frequency of @users*	**14.** *Frequency of Slang Words*
7. *Extent of Capitalization*	**15.** *Average Positive and Negative Sentiment of Tweets*
8. *Frequency of the First POS Tag*	

Algorithm for Spam Filter

Input: Build an initial naive bayes classifier NB- C, using the tweet sets *M* (mixed unlabeled set containing spams and non-spams) and *P* (labeled non-spam set)

1: Loop while classifier parameters change

2: for each tweet $t_i \in M$ do

3: Compute $Pr[c_1 | t_i]$, $Pr[c_2 | t_i]$ using the current NB //c_1 - non-spam class , c_2 - spam class

4: $Pr[c_2 | t_i] = 1 - Pr[c_1 | t_i]$

5: Update $Pr[f_{i,k}|c_1]$ and $Pr[c_1]$ given the probabilistically assigned class for all t_i ($Pr[c_1|t_i]$). (a new NB-C is being built in the process)

6: end for

7: end loop

$$\Pr[c_j \mid t_i] = \frac{\Pr[c_j] \prod_k \Pr[f_{i,k}|c_j]}{\sum_r \Pr[c_r] \prod_k P(f_{i,k}|c_r)}$$

Categorization of Noisy Text

- Dropping of Vowels - *btfl (beautiful), lvng (loving)*
- Vowel Exchange - *good vs. gud (o,u)*
- Mis-spelt words - *redicule (ridicule), magnificant (magnificent)*
- Text Compression - *shok (shock), terorism (terrorism)*
- Phonetic Transformation - *be8r (better), gud (good), fy9 (fine), gr8 (great)*
- Normalization and Pragmatics - *hapyyyyyy (happy), guuuuud (good)*
- Segmentation with Punctuation - *beautiful, (beautiful)*
- Segmentation with Compound Words - *breathtaking (breath-taking), eyecatching (eye-catching), good-looking (good looking)*
- Hashtags and Segmentation - *#notevenkidding, #worthawatch*
- Combination of all - *#awsummm (awesome), gr88888 (great), amzng,btfl (amazing, beautiful).*

Spell-Checker Algorithm

- Heuristically driven to resolve the *identified errors* with a *minimum edit distance based spell checker*
- A *normalize* function takes care of *Pragmatics* and *Number Homophones*
  - Replaces *happpyyyy* with *hapy*, *‘2’ with ‘to’, ‘8’ with ‘eat’, ‘9’ with ‘ine’*
- A *vowel_dropped* function takes care of the *vowel dropping* phenomenon
- The parameters *offset* and *adv* are determined empirically
- Words are marked during normalization, to preserve their pragmatics
  - *happppyyyyy,* normalized to *hapy* and thereafter spell-corrected to *happy,* is marked so as to not lose its pragmatic content

Spell-Checker Algorithm

- **Input:** For string s, let *S* be the set of words in the lexicon starting with the initial letter of *s*.
- /* Module Spell Checker */
- **for** each word *w* ∈ *S* **do**
- *w'*=vowel_dropped(*w*)
- *s'*=normalize(*s*)
- */*diff(s,w) gives difference of length between s and w*/*
- **if** diff(*s'* , *w'*) < offset **then**
- score[*w*]=min(edit_distance(*s,w*),edit_distance(*s,*
- *w'*), edit_distance(*s'* , *w*))
- else
- score[*w*]=max_centinel
- **end if**
- **end for**

Spell-Checker Algorithm Contd..

- Sort score of each *w* in the Lexicon and retain the top m entries in suggestions(*s*) for the original string *s*
- **for** each *t* in suggestions(*s*) **do**
- $edit_1$=edit_distance(*s' , s*)
- */*t.replace(char1,char2) replaces all occurrences of char1 in the string t with char2*/*
- $edit_2$=edit_distance(t.replace(*a , e*), *s'*)
- $edit_3$=edit_distance(t.replace(*e , a*), *s'*)
- $edit_4$=edit_distance(t.replace(*o , u*), *s'*)
- $edit_5$=edit_distance(t.replace(*u , o*), *s'*)
- $edit_6$=edit_distance(t.replace(*i , e*), *s'*)
- $edit_7$=edit_distance(t.replace(*e, i*), *s'*)
- count=overlapping_characters(*t, s'*)
- min_edit=
- min($edit_1$,$edit_2$,$edit_3$,$edit_4$,$edit_5$,$edit_6$,$edit_7$)
- **if** (min_edit ==0 or score[*s'*] == 0) **then**
- adv=-2 */* for exact match assign advantage score */*
- else
- adv=0
- **end if**
- final_score[t]=min_edit+adv+score[*w*]-count;
- **end for**
- return *t* with minimum final_score;

Pragmatics

Pragmatics

- *Elongation of a word, repeating alphabets multiple times* - Example: *happppyyyyyy*, *goooooooood*. More weightage is given by repeating them twice

Pragmatics

- *Elongation of a word, repeating alphabets multiple times* - Example: *happppyyyyyy*, *goooooooood*. More weightage is given by repeating them twice

- *Use of Hashtags* - *#overrated, #worthawatch.* More weightage is given by repeating them thrice

Pragmatics

- *Elongation of a word, repeating alphabets multiple times* - Example: *happppyyyyyy*, *goooooood*. More weightage is given by repeating them twice
- *Use of Hashtags* - *#overrated, #worthawatch.* More weightage is given by repeating them thrice
- *Use of Emoticons* - ☺ *(happy)*, ☹ *(sad)*

Pragmatics

- *Elongation of a word, repeating alphabets multiple times* - Example: *happppyyyyyy*, *gooooooood*. More weightage is given by repeating them twice
- *Use of Hashtags* - *#overrated, #worthawatch.* More weightage is given by repeating them thrice
- *Use of Emoticons* - ☺ *(happy)*, ☹ *(sad)*
- *Use of Capitalization* - where words are written in capital letters to express intensity of user sentiments
  - *Full Caps* - Example: *I HATED that movie.* More weightage is given by repeating them thrice
  - *Partial Caps*- Example: *She is a Loving mom*. More weightage is given by repeating them twice

Spam Filter Evaluation

2-Class Classification

Tweets	Total Tweets	Correctly Classified	Misclassified	Precision (%)	Recall (%)
All	7007	3815	3192	54.45	55.24
Only spam	**1993**	**1838**	**155**	**92.22**	**92.22**
Only non-spam	5014	2259	2755	45.05	-

4-Class Classification

Tweets	Total Tweets	Correctly Classified	Misclassified	Precision (%)	Recall (%)
All	7007	5010	1997	**71.50**	**54.29**
Only spam	1993	1604	389	80.48	80.48
Only non-spam	5014	4227	787	84.30	-

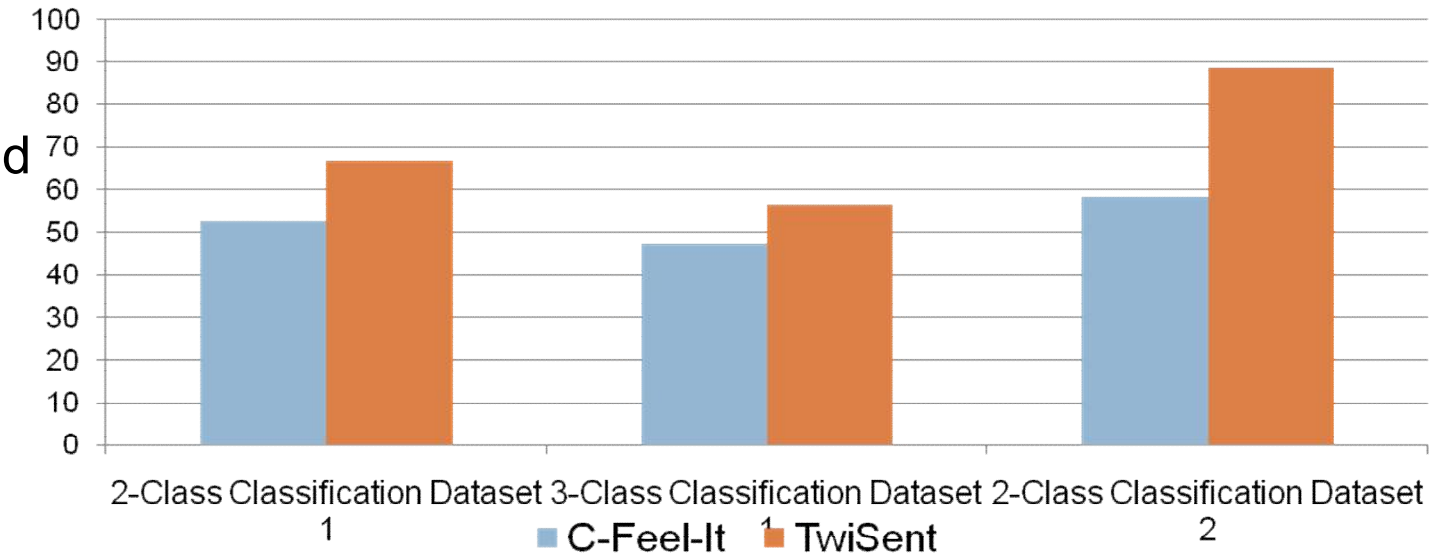
TwiSent Evaluation

TwiSent Evaluation

Lexicon-based
Classification

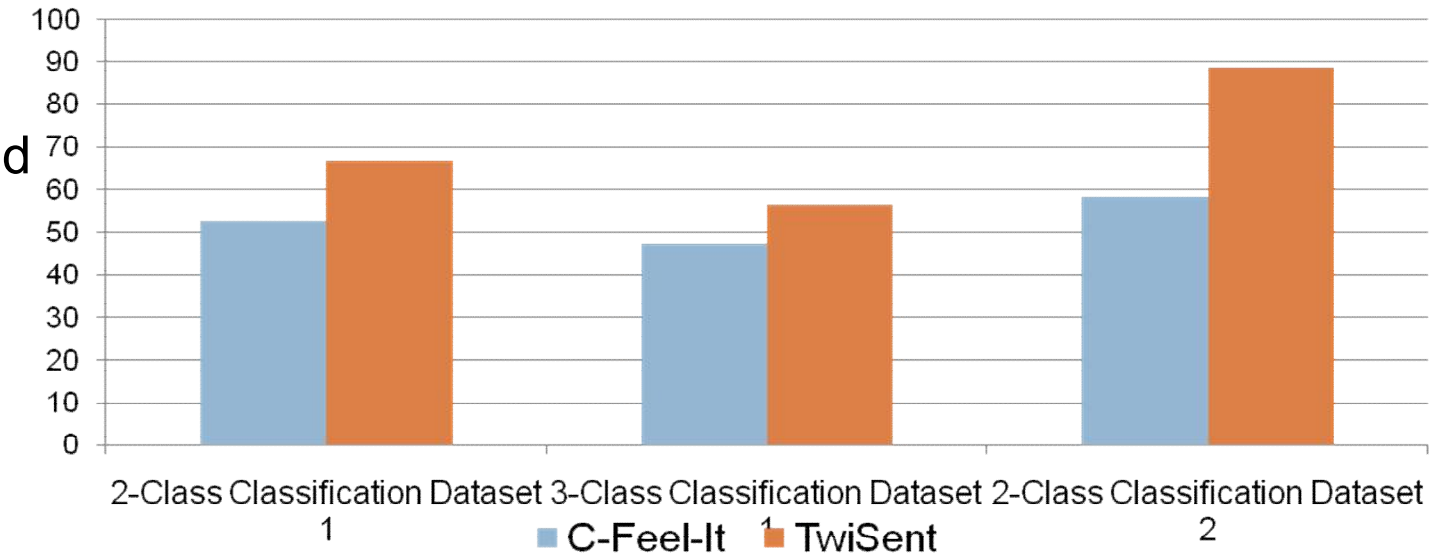
TwiSent Evaluation

Lexicon-based Classification



TwiSent Evaluation

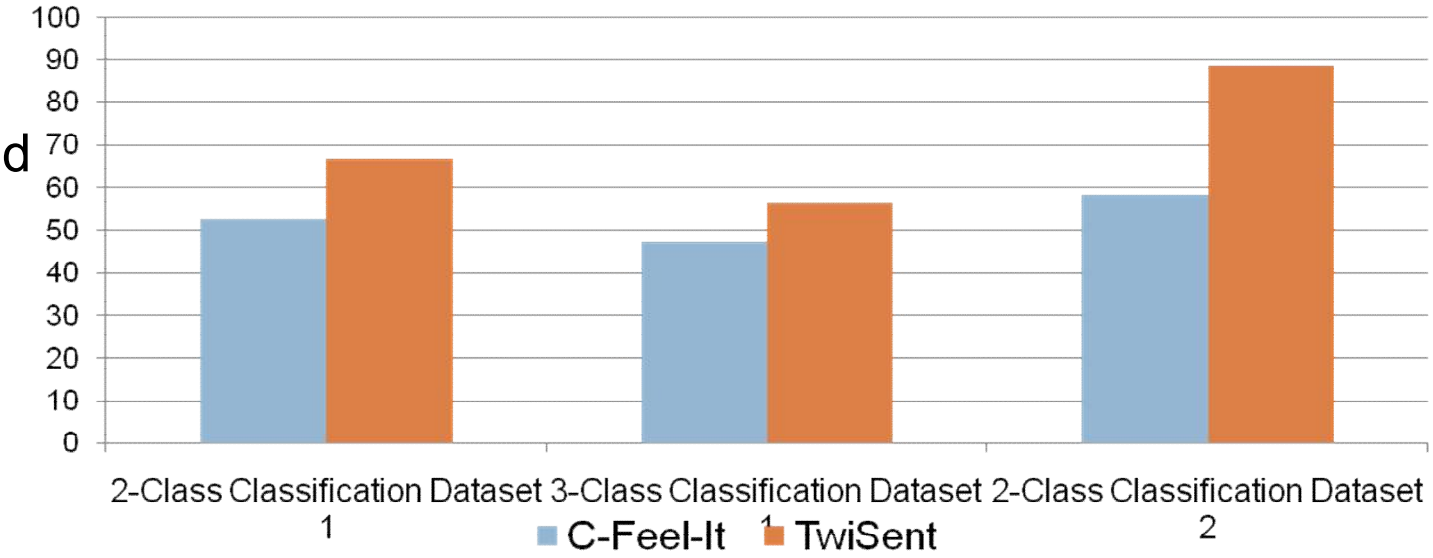
Lexicon-based Classification



Supervised Classification

TwiSent Evaluation

Lexicon-based Classification

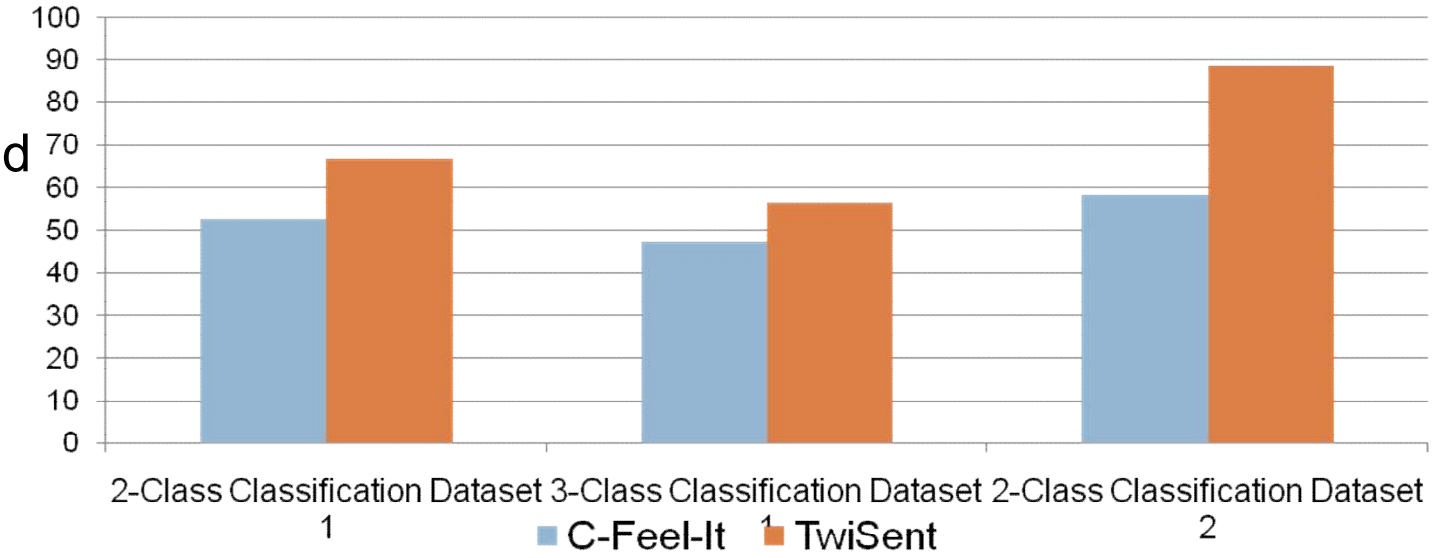


Supervised Classification

System	2-class Accuracy	Precision/Recall
C-Feel-It	50.8	53.16/72.96
TwiSent	**68.19**	**64.92/69.37**

TwiSent Evaluation

Lexicon-based Classification

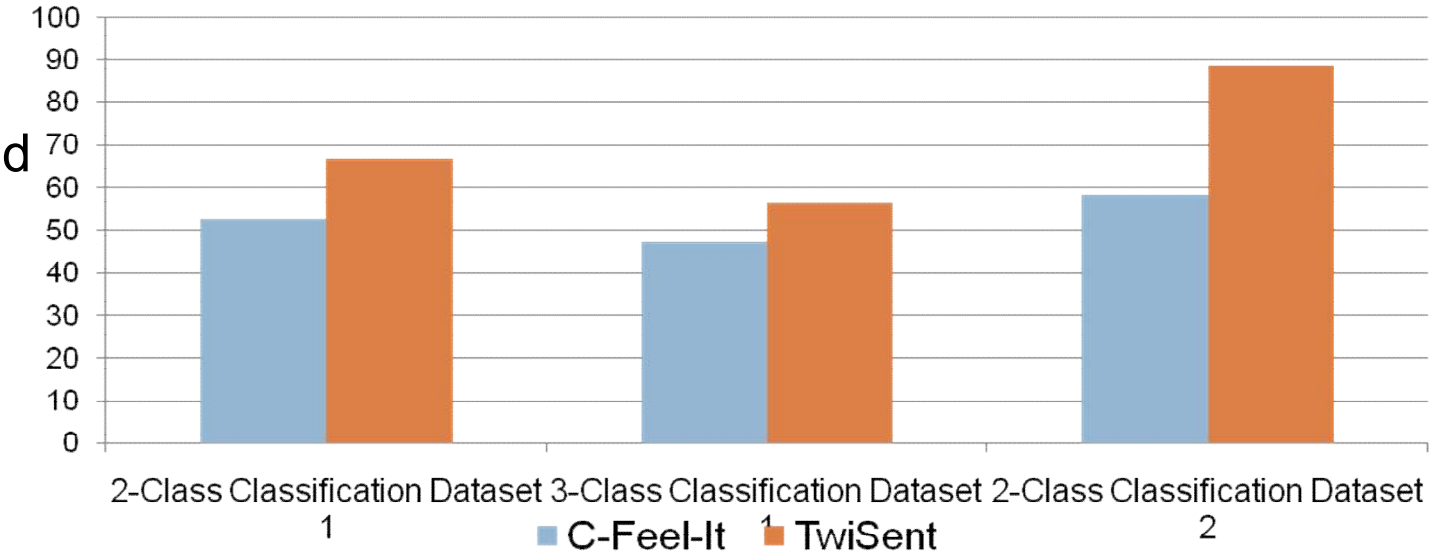


Supervised Classification

System	2-class Accuracy	Precision/Recall
C-Feel-It	50.8	53.16/72.96
TwiSent	**68.19**	**64.92/69.37**

TwiSent Evaluation

Lexicon-based Classification

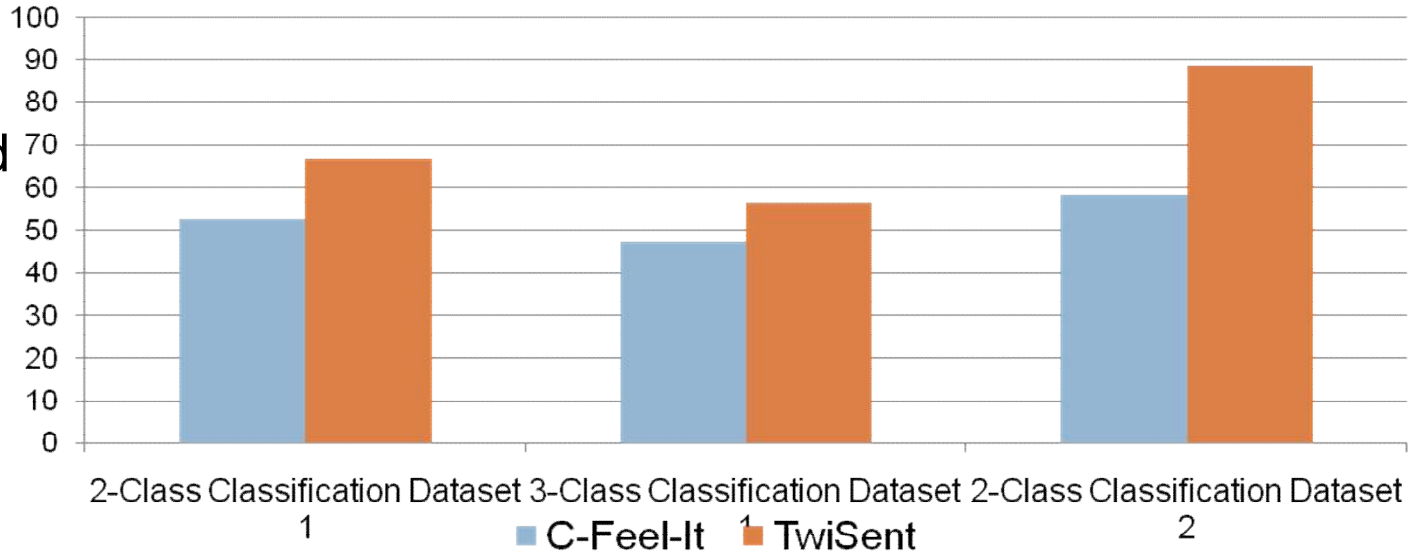


Ablation Test

System	2-class Accuracy	Precision/Recall
C-Feel-It	50.8	53.16/72.96
TwiSent	**68.19**	**64.92/69.37**

TwiSent Evaluation

Lexicon-based Classification



Ablation Test

Module Removed	Accuracy	Statistical Significance Confidence (%)
Entity-Specificity	65.14	95
Spell-Checker	64.2	99
Pragmatics Handler	63.51	99
Complete System	**66.69**	-

Role of Social Media Content and World Knowledge in Information Retrieval

- YouCat: Unsupervised System for Youtube Video Categorization from User Comments and Meta Data using WordNet and Wikipedia

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres

For Example: A *Tom and Jerry Show* is tagged by the genres **Comedy** and **Animation**

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video

For Example: A *Tom and Jerry Show* is tagged by the genres **Comedy** and **Animation**

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video

Meta-Description of a video : It was an awesome slam dunk in the NBA finals by Michael Jordan.
Comments : He is the greatest basketball player of all times. ….

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words

Meta-Description of a video : It was an awesome slam dunk in the NBA finals by Michael Jordan.
Comments : He is the greatest basketball player of all times. ….

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words

Funny and *Laugh* are the keywords that define the **Comedy** genre

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words
- The words and synonymous concepts extracted using a Thesaurus are compiled into a Seed List

Funny and *Laugh* are the keywords that define the **Comedy** genre

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words
- The words and synonymous concepts extracted using a Thesaurus are compiled into a Seed List

Seed List for **Comedy** genre: funny, humor, hilarious, joke, comedy, roflmao, laugh, lol, rofl, roflmao, …

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words
- The words and synonymous concepts extracted using a Thesaurus are compiled into a Seed List
- A Concept List is created using WordNet and Named Entities from Wikipedia based on related concepts from Seed List using an **overlap-based** classification

Seed List for **Comedy** genre: funny, humor, hilarious, joke, comedy, roflmao, laugh, lol, rofl, roflmao, …

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words
- The words and synonymous concepts extracted using a Thesaurus are compiled into a Seed List
- A Concept List is created using WordNet and Named Entities from Wikipedia based on related concepts from Seed List using an **overlap-based** classification

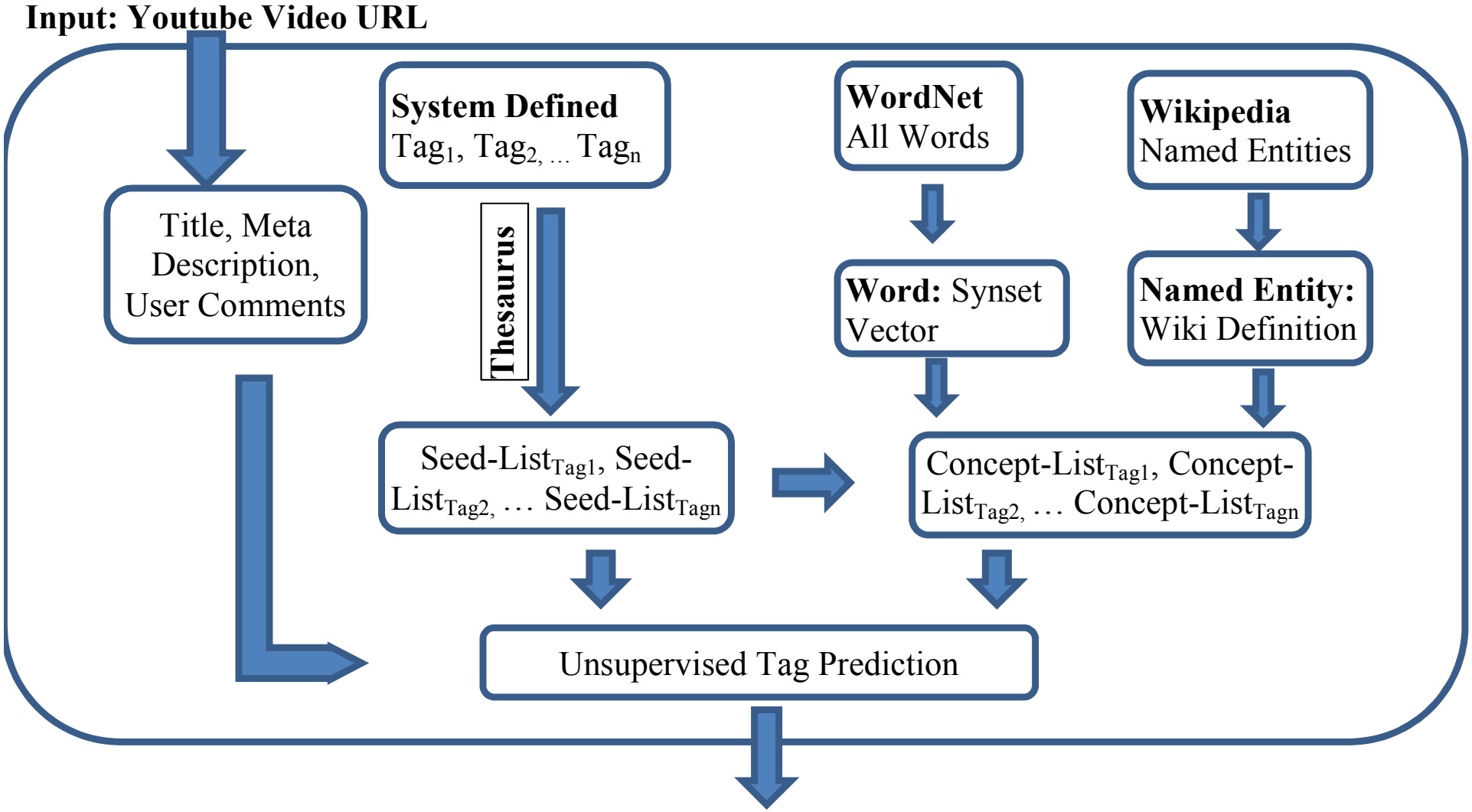
- *dunk - {dunk, dunk **shot**, stuff **shot**; dunk, dip, souse, plunge, douse; dunk; dunk, dip}* is classified to Sports with the help of **Sports Seed List**
- Michael Jordon is also Classified to Sports based on the overlap of its Wikipedia definition with the **Sports Seed List**

YouCat: Youtube Video Categorization from User Comments and Meta-Data using WordNet and Wiki

- Given a Youtube video, predict its genres
- Extract features from User Comments and Meta-Description of the video
- A genre is pre-defined based on a few seed words
- The words and synonymous concepts extracted using a Thesaurus are compiled into a Seed List
- A Concept List is created using WordNet and Named Entities from Wikipedia based on related concepts from Seed List using an **overlap-based** classification
- The final classification is based on an **overlap approach** based on the *Seed List* and *Concept List* for each genre – some parameters learnt from data

  - *dunk - {dunk, dunk **shot**, stuff **shot**; dunk, dip, souse, plunge, douse; dunk; dunk, dip}* is classified to Sports with the help of **Sports Seed List**
  - Michael Jordon is also Classified to Sports based on the overlap of its Wikipedia definition with the **Sports Seed List**

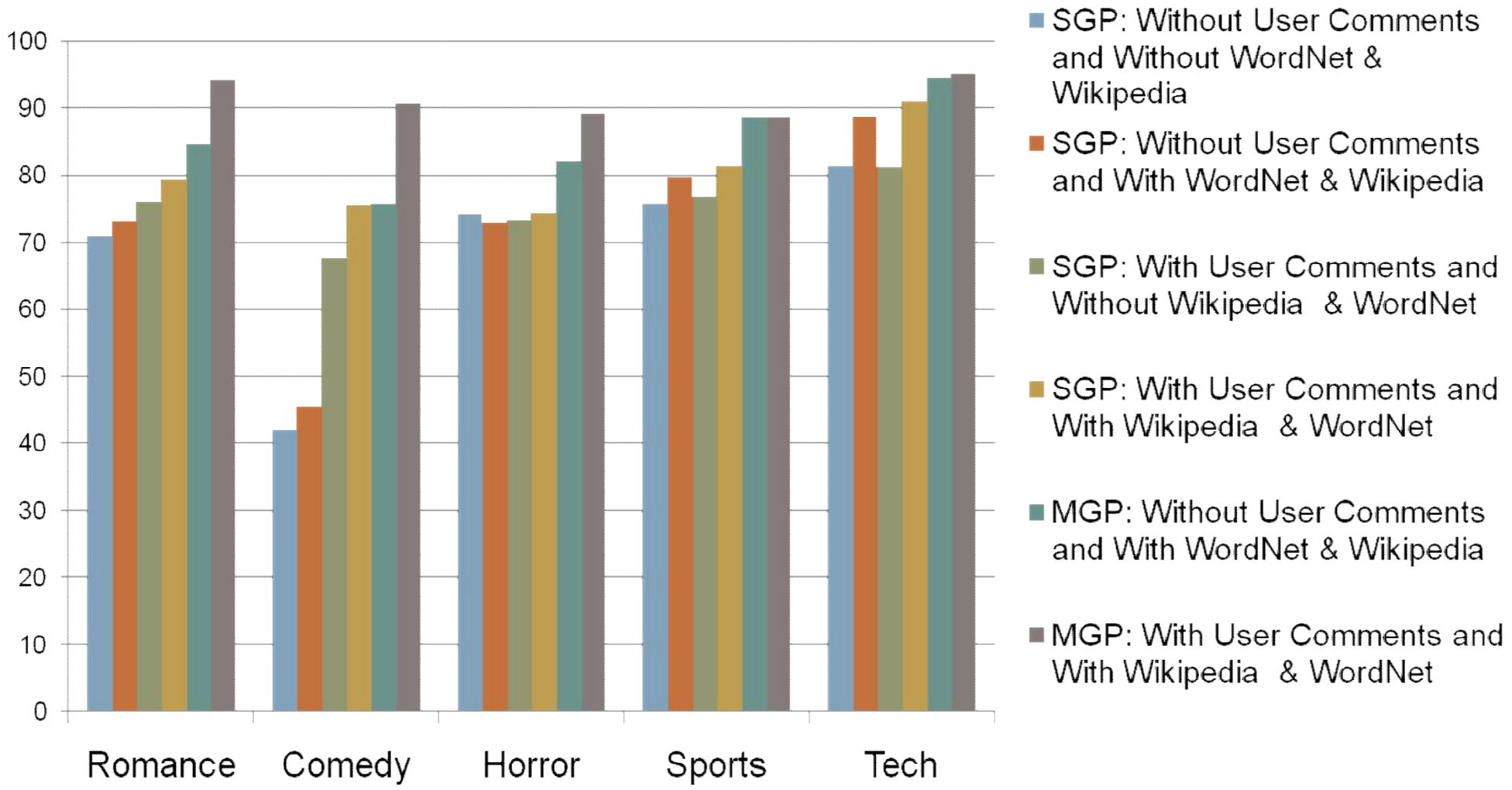
YouCat System Architecture



Tag Prediction

- Feature Scoring
  - $score(f \in genre_k; w_1, w_2) = w_1 \times \sum_{j:word_j \in seed_k} \mathbf{1} + w_2 \times \sum_{j:word_j \in concept_k} \mathbf{1}$
- Video Genre Scoring
  - $score(video \in genre_k; p_1, p_2, p_3) = p_1 \times score(f^{Title} \in genre_k) + p_2 \times score(f^{Meta\ Data} \in genre_k) + p_3 \times score(f^{Comments} \in genre_k)$
- Single Genre Prediction
  - $video_{genre} = argmax_k\ score(video \in genre_k)$
- Multiple Genre Prediction
  - $video_{genre} = k, if\ score(video \in genre_k) \geq \theta$
  - $where\ \theta = \frac{1}{k} \sum_k score(video \in genre_k)$

Genre-wise Results



YouCat Dataset Statistics

Video Statistics

Comedy	Horror	Sports	Romance	Tech	Total
2682	2802	2577	2477	2299	12837

Comment Statistics

Romance	Comedy	Horror	Sports	Tech
103.64	168.94	83.4	56.91	223.45

Supervised Baseline

SVM Features	F_1-Score(%)
All Unigrams	82.5116
Unigrams+Without stop words	83.5131
Unigrams+ Without stop words +Lemmatization	83.8131
Unigrams+Without stop words +Lemmatization+ POS Tags	83.8213
Top Unigrams+Without stop words +Lemmatization+POS Tags	84.0524
All Bigrams	74.2681
Unigrams+Bigrams+Without stop words+Lemmatization	**84.3606**

Overall Results

Model	Average F_1 Score
SVM Baseline: Single Genre Prediction With User Comments	84.3606
Single Genre Prediction : Without User Comments + Without Wikipedia & WordNet	68.76
Single Genre Prediction : With User Comments + Without Wikipedia & WordNet	74.95
Single Genre Prediction : Without User Comments + With Wikipedia & WordNet	71.984
Single Genre Prediction : With User Comments+ With Wikipedia &WordNet	*80.9*
Multiple Genre Prediction : Without User Comments + With Wikipedia & WordNet	84.952
Multiple Genre Prediction : With User Comments + With Wikipedia & WordNet	**91.48**

Comments/Genre and Confusion Matrix

Average Comments/Genre

	Average Tags/Video Without User Comments	Average Tags/Video With User Comments
Romance	1.45	1.55
Comedy	1.67	1.80
Horror	1.38	1.87
Sports	1.36	1.40
Tech	1.29	1.40
Average	**1.43**	**1.60**

Confusion Matrix

	Romance	Comedy	Horror	Sports	Tech
Romance	80.16	8.91	3.23	4.45	3.64
Comedy	3.13	77.08	3.47	9.03	7.29
Horror	10.03	9.34	75.78	3.46	1.38
Sports	0.70	7.30	0	89.05	2.92
Tech	0.72	5.07	0.36	1.81	92.03

Effect of Social Media Content in IR

Genre	Without User Comments + Without Wikipedia & WordNet			With User Comments + Without Wikipedia & WordNet		
	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Romance	76.26	66.27	70.91	77.36	74.60	**75.95**
Comedy	43.96	40.00	41.89	69.23	66.00	**67.58**
Horror	80.47	68.67	74.10	76.45	70.33	**73.26**
Sports	84.21	68.71	75.67	85.07	69.94	**76.77**
Tech	90.83	73.50	81.25	84.09	78.45	**81.17**

Effect of World Knowledge in IR

Genre	Without User Comments + With Wikipedia & WordNet			With User Comments + With Wikipedia & WordNet		
	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Romance	76.06	70.63	73.24	80.16	78.57	**79.36**
Comedy	47.31	44.00	45.6	77.08	74.00	**75.51**
Horror	75.63	70.33	72.88	75.78	73.00	**74.36**
Sports	87.5	73.01	79.60	89.05	74.85	**81.33**
Tech	92.34	85.16	88.60	92.03	89.75	**90.88**

Drawbacks

- User Comments
  - Noisy (slangs, abbreviations, informal language, spams *etc.*)
  - Off-topic conversation (abuses, chit-chats *etc.*)
- Concept Expansion (WordNet + Wikipedia)
  - Topic-drift
  - *good* - gloss of one of its synsets {*dear, good, near -- with or in a close or intimate relationship*}
- Bias towards Comedy
  - *Romantic-Comedy*, *Horror-Comedy*
- Ambiguity in Named-Entity due to shorter context
  - *Manchester Rocks !!!*

- Sentiment Influencing Semantic Similarity Measure

SenSim: Leveraging Sentiment to Compute Similarity

SenSim: Leveraging Sentiment to Compute Similarity

- Introduce Sentiment as another feature in the Semantic Similarity Measure
  - "*Among a set of a similar word pairs, a pair is more similar if their sentiment content is the same*"
  - Is "*enchant*" (hold spellbound) more similar to "*endear*" (make endearing or lovable) than to "*delight* "(give pleasure to or be pleasing to) ?

SenSim: Leveraging Sentiment to Compute Similarity

- Introduce Sentiment as another feature in the Semantic Similarity Measure
  - "*Among a set of a similar word pairs, a pair is more similar if their sentiment content is the same*"
  - Is "*enchant*" (hold spellbound) more similar to "*endear*" (make endearing or lovable) than to "*delight* "(give pleasure to or be pleasing to) ?
- Useful for replacing an unknown feature in test set with a similar feature in training set

SenSim: Leveraging Sentiment to Compute Similarity

- Introduce Sentiment as another feature in the Semantic Similarity Measure
  - "*Among a set of a similar word pairs, a pair is more similar if their sentiment content is the same*"
  - Is "*enchant*" (hold spellbound) more similar to "*endear*" (make endearing or lovable) than to "*delight* "(give pleasure to or be pleasing to) ?
- Useful for replacing an unknown feature in test set with a similar feature in training set
- Given a **word** in a sentence, create its *Similarity Vector*
  - Use Word Sense Disambiguation on context to find its **Synset-id**
  - Create a **Gloss Vector** (*sparse*) using its gloss
  - Extend gloss using *relevant* **WordNet Relations**
    - **Learn the relations to use for different POS tags and the depth in WordNet hierarchy**
  - Incorporate SentiWordNet Scores in the Expanded Vector using Different Scoring

SenSim: Leveraging Sentiment to Compute Similarity

- Introduce Sentiment as another feature in the Semantic Similarity Measure
  - "*Among a set of a similar word pairs, a pair is more similar if their sentiment content is the same*"
  - Is "*enchant*" (hold spellbound) more similar to "*endear*" (make endearing or lovable) than to "*delight* "(give pleasure to or be pleasing to) ?
- Useful for replacing an unknown feature in test set with a similar feature in training set
- Given a **word** in a sentence, create its *Similarity Vector*
  - Use Word Sense Disambiguation on context to find its **Synset-id**
  - Create a **Gloss Vector** (*sparse*) using its gloss
  - Extend gloss using *relevant* **WordNet Relations**
    - **Learn the relations to use for different POS tags and the depth in WordNet hierarchy**
  - Incorporate SentiWordNet Scores in the Expanded Vector using Different Scoring

Sentiment-Semantic Correlation

Annotation Strategy	Overall	NOUN	VERB	ADJECTIVES	ADVERBS
Meaning	0.768	0.803	0.750	0.527	0.759
Meaning + Sentiment	0.799	0.750	0.889	0.720	0.844

WordNet Relations used for Expansion

POS	WordNet relations used for expansion
Nouns	*hypernym, hyponym, nominalization*
Verbs	*nominalization, hypernym, hyponym*
Adjectives	*also see, nominalization, attribute*
Adverbs	*derived*

Scoring Formula

- $Score_{SD}(A) = SWN_{pos}(A) - SWN_{neg}(A)$
- $Score_{SM}(A) = max(SWN_{pos}(A), SWN_{neg}(A))$
- $Score_{TM}(A) =$

$sign(max(SWN_{pos}(A), SWN_{neg}(A))) * (1+abs(max(SWN_{pos}(A), SWN_{neg}(A)))$

$$SenSim_x(A, B) = cosine\ (gloss_{vec}(sense(A)),\ gloss_{vec}(sense(B)))$$

Where,

$gloss_{vec}$ $= 1{:}score_x(1)\ 2{:}score_x(2) \ldots\ n{:}score_x(n)$

$score_x(Y)$ = Sentiment score of word Y using scoring function x

x = Scoring function of type SD/SM/TD/TM

Evaluation on Gold Standard Data: Word Pair Similarity

Evaluation on Gold Standard Data: Word Pair Similarity

- A set of 50 word pairs (with given context) manually marked
- Each word pair is given 3 scores in the form of ratings (1-5):
  - Similarity based on meaning
  - Similarity based on sentiment
  - Similarity based on meaning + sentiment

Evaluation on Gold Standard Data: Word Pair Similarity

- A set of 50 word pairs (with given context) manually marked
- Each word pair is given 3 scores in the form of ratings (1-5):
  - Similarity based on meaning
  - Similarity based on sentiment
  - Similarity based on meaning + sentiment

  Agreement metric: Pearson correlation coefficient

Evaluation on Gold Standard Data: Word Pair Similarity

- A set of 50 word pairs (with given context) manually marked
- Each word pair is given 3 scores in the form of ratings (1-5):
  - Similarity based on meaning
  - Similarity based on sentiment
  - Similarity based on meaning + sentiment

Agreement metric: Pearson correlation coefficient

Metric Used	Overall	NOUN	VERB	ADJECTIVES	ADVERBS
LESK (Banerjee *et al.*, 2003)	0.22	0.51	-0.91	0.19	0.37
LIN (Lin, 1998)	0.27	0.24	0.00	NA	Na
LCH (Leacock *et al.*, 1998)	0.36	0.34	0.44	NA	NA
SenSim (SD)	0.46	0.73	0.55	0.08	0.76
SenSim (SM)	**0.50**	0.62	0.48	0.06	0.54
SenSim (TD)	0.45	0.73	0.55	0.08	0.59
SenSim (TM)	0.48	0.62	0.48	0.06	0.78

Evaluation on Travel Review Data: Feature Replacement

Metric Used	Accuracy (%)	PP	NP	PR	NR
Baseline	89.10	91.50	87.07	85.18	91.24
LESK (Banerjee *et al.*, 2003)	89.36	91.57	87.46	85.68	91.25
LIN (Lin, 1998)	89.27	91.24	87.61	85.85	90.90
LCH (Leacock *et al.*, 1998)	89.64	90.48	88.86	86.47	89.63
SenSim (SD)	89.95	91.39	88.65	87.11	90.93
SenSim (SM)	90.06	92.01	88.38	86.67	91.58
SenSim (TD)	90.11	91.68	88.69	86.97	91.23

Conclusions

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance
- Dependency Relations capture association between words forming opinion expressions, which incorporate feature specificity

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance
- Dependency Relations capture association between words forming opinion expressions, which incorporate feature specificity
- Informal language form and the social media content require different treatment than general SA to harvest their sentiment content

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance
- Dependency Relations capture association between words forming opinion expressions, which incorporate feature specificity
- Informal language form and the social media content require different treatment than general SA to harvest their sentiment content
- Sentiment Analysis of Reviews require extensive world knowledge, which can be harnessed using Wikipedia

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance
- Dependency Relations capture association between words forming opinion expressions, which incorporate feature specificity
- Informal language form and the social media content require different treatment than general SA to harvest their sentiment content
- Sentiment Analysis of Reviews require extensive world knowledge, which can be harnessed using Wikipedia
- Social Media Content and External Knowledge Sources provide a goldmine of information useful for Information Extraction

Conclusions

- Sentiment Analysis is more than just a bag-of-words classification problem
- Discourse Markers, often ignored as stop words in a BOW model, improve its performance
- Dependency Relations capture association between words forming opinion expressions, which incorporate feature specificity
- Informal language form and the social media content require different treatment than general SA to harvest their sentiment content
- Sentiment Analysis of Reviews require extensive world knowledge, which can be harnessed using Wikipedia
- Social Media Content and External Knowledge Sources provide a goldmine of information useful for Information Extraction
- Incorporation of these new perspectives in a traditional BOW Model can improve its performance, while retaining its simplicity

Contributions

Contributions

- Discourse markers can be useful in BOW Model

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent
- TwiSent: Categorization of Spams in Twitter w.r.t SA

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent
- TwiSent: Categorization of Spams in Twitter w.r.t SA
- TwiSent: Categorization of Noisy Text in Twitter

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent
- TwiSent: Categorization of Spams in Twitter w.r.t SA
- TwiSent: Categorization of Noisy Text in Twitter
- TwiSent: Handling Pragmatics and introducing Feature Specificity

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent
- TwiSent: Categorization of Spams in Twitter w.r.t SA
- TwiSent: Categorization of Noisy Text in Twitter
- TwiSent: Handling Pragmatics and introducing Feature Specificity
- YouCat: Analyzing the combined influence of social media content and concept expansion in Information Retrieval

Contributions

- Discourse markers can be useful in BOW Model
- Feature Specificity can be captured by Dependency Relations
- WikiSent: Incorporating World Knowledge through Wikipedia improves SA
- We analyze the movie trend from genre, year of release and polarity information using WikiSent
- TwiSent: Categorization of Spams in Twitter w.r.t SA
- TwiSent: Categorization of Noisy Text in Twitter
- TwiSent: Handling Pragmatics and introducing Feature Specificity
- YouCat: Analyzing the combined influence of social media content and concept expansion in Information Retrieval
- SenSim: Incorporation of sentiment as a feature in semantic similarity improves similarity performance

Future Works

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts
- Improving Spell-Checker using SMT on a parallel corpora

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts
- Improving Spell-Checker using SMT on a parallel corpora
- Pragmatics handling needs to be improved

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts
- Improving Spell-Checker using SMT on a parallel corpora
- Pragmatics handling needs to be improved
- WikiSent needs handling of *synonymous concepts*, *anaphora resolution* and *word sense disambiguation*

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts
- Improving Spell-Checker using SMT on a parallel corpora
- Pragmatics handling needs to be improved
- WikiSent needs handling of *synonymous concepts*, *anaphora resolution* and *word sense disambiguation*
- YouCat needs to be tested on more number of genres for fine distinction, requires a Comment Filter and a WSD module

Future Works

- Incorporation of Parsing to determine scope of discourse markers for structured review analysis
- Ranking of discourse features with positional importance
- Generating a domain-dependent lexicon for better classification of domain-dependent concepts
- Improving Spell-Checker using SMT on a parallel corpora
- Pragmatics handling needs to be improved
- WikiSent needs handling of *synonymous concepts*, *anaphora resolution* and *word sense disambiguation*
- YouCat needs to be tested on more number of genres for fine distinction, requires a Comment Filter and a WSD module

All these perspectives need to be brought in a single canvas and their collective effectiveness needs to be evaluated

Publications (1/2)

- **Sentiment Analysis in Twitter with Lightweight Discourse Analysis**, Subhabrata Mukherjee and Pushpak Bhattacharyya, In Proceedings of the 24th International Conference on Computational Linguistics (**COLING 2012**), IIT Bombay, Mumbai, Dec 8 - Dec 15, 2012

- **YouCat : Weakly Supervised Youtube Video Categorization System from Meta Data & User Comments using WordNet & Wikipedia**, Subhabrata Mukherjee and Pushpak Bhattacharyya, In Proceedings of the 24th International Conference on Computational Linguistics (**COLING 2012**), IIT Bombay, Mumbai, Dec 8 - Dec 15, 2012 (Long Paper)

- **TwiSent: A Multi-Stage System for Analyzing Sentiment in Twitter**, Subhabrata Mukherjee, Akshat Malu, Balamurali A.R. and Pushpak Bhattacharyya, In Proceedings of The 21st ACM Conference on Information and Knowledge Management (**CIKM 2012**), Hawai, Oct 29 - Nov 2, 2012

Publications (2/2)

- **WikiSent : Weakly Supervised Sentiment Analysis Through Extractive Summarization With Wikipedia**, Subhabrata Mukherjee and Pushpak Bhattacharyya, In Proceedings of the European Conference on Machine Learning (**ECML PKDD 2012**), Bristol, U.K., 24-28 Sept, 2012

- **Feature Specific Sentiment Analysis for Product Reviews**, Subhabrata Mukherjee and Pushpak Bhattacharyya, In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Intelligence (**CICLING 2012**), New Delhi, India, March, 2012

- **Leveraging Sentiment to Compute Word Similarity**, Balamurali A.R., Subhabrata Mukherjee, Akshat Malu and Pushpak Bhattacharyya, In Proceedings of the 6th International Global Wordnet Conference (**GWC 2011**), Matsue, Japan, Jan, 2012

Thank You