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
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!

The movie stars Mr. X

The movie was horrible!

[ Sentimental ]

[ Factual ]

[ 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

The movie was fabulous! [Sentimental]

The movie stars Mr. X [Factual]

The movie was horrible! [Sentimental]

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 pretty 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
“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”
“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”
“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”
"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”
“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 \( w_i, w_{i+1} \notin \text{Stopwords} \), then they are directly related. \( w_i, w_{i+1} \in S \)

- **Dependency Relation**
  - Capture **long range dependencies**
  - Let \( \text{Dependency\_Relation} \) be the list of **significant relations**.
  - Any 2 words \( w_i \) and \( w_j \) in \( S \) are directly related, if
    \[ \exists D_i \quad \text{s.t.} \quad D_i(w_i, w_j) \in \text{Dependency\_Relation} \]
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 $C_i \ \forall \ i = 1...n$

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., $k = \arg\min_{i \in n} \text{dist}(w_j, f_i)$, Where $\text{dist}(w_j, f_i)$ 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 $\text{dist}(w_j, f_i) < \theta$, Where $\theta$ 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 the only person that dislikes the software is me.
Graph
Graph
Graph
Graph

I have an ipod

it is a

great buy but im

probably the only person

that dislike the software

7/23/2013
Classification and Datasets
Classification and Datasets

- Lexicon Classification
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
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
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
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 – \((2^{40} - 1)\)
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

<table>
<thead>
<tr>
<th>Relations</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>63.5</td>
</tr>
<tr>
<td>Dep</td>
<td>67.3</td>
</tr>
<tr>
<td>Rcmd</td>
<td>65.4</td>
</tr>
<tr>
<td>xcomp, conj_and ccomp, iobj</td>
<td>61.5</td>
</tr>
<tr>
<td>advcl, appos, csbj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj</td>
<td>63.5</td>
</tr>
</tbody>
</table>
## Ablation test

<table>
<thead>
<tr>
<th>Relations</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>63.5</td>
</tr>
<tr>
<td>Dep</td>
<td>67.3</td>
</tr>
<tr>
<td>Rcmod</td>
<td>65.4</td>
</tr>
<tr>
<td>xcomp, conj_and ccomp, iobj</td>
<td>61.5</td>
</tr>
<tr>
<td>advcl , appos, csubj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj</td>
<td>63.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation Set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Dep+Rcmod</td>
<td>66</td>
</tr>
<tr>
<td>Without Dep</td>
<td>69</td>
</tr>
<tr>
<td>Without Rcmod</td>
<td>67</td>
</tr>
<tr>
<td>Without Dep+Rcmod</td>
<td>68</td>
</tr>
</tbody>
</table>
# Ablation test

<table>
<thead>
<tr>
<th>Relations</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>63.5</td>
</tr>
<tr>
<td>Dep</td>
<td>67.3</td>
</tr>
<tr>
<td>Rcmod</td>
<td>65.4</td>
</tr>
<tr>
<td>xcomp, conj_and ccomp, iobj</td>
<td>61.5</td>
</tr>
<tr>
<td>advcl, appos, csubj, abbrev, infmod, npavmod, rel, acomp, agent, csubjpass, partmod, pobj, purpcl, xsubj</td>
<td>63.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation Set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Dep+Rcmod</td>
<td>66</td>
</tr>
<tr>
<td>Without Dep</td>
<td>69</td>
</tr>
<tr>
<td>Without Rcmod</td>
<td>67</td>
</tr>
<tr>
<td>Without Dep+Rcmod</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>67.85</td>
</tr>
<tr>
<td>3</td>
<td>69.28</td>
</tr>
<tr>
<td>4</td>
<td>68.21</td>
</tr>
<tr>
<td>5</td>
<td>67.40</td>
</tr>
</tbody>
</table>
Lexicon based classification (Dataset 2)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Baseline 1 (%)</th>
<th>Baseline 2 (%)</th>
<th>Proposed System (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antivirus</td>
<td>50.00</td>
<td>56.82</td>
<td>63.63</td>
</tr>
<tr>
<td>Camera 1</td>
<td>50.00</td>
<td>61.67</td>
<td>78.33</td>
</tr>
<tr>
<td>Camera 2</td>
<td>50.00</td>
<td>61.76</td>
<td>70.58</td>
</tr>
<tr>
<td>Camera 3</td>
<td>51.67</td>
<td>53.33</td>
<td>60.00</td>
</tr>
<tr>
<td>Camera 4</td>
<td>52.38</td>
<td>57.14</td>
<td>78.57</td>
</tr>
<tr>
<td>Diaper</td>
<td>50.00</td>
<td>63.63</td>
<td>57.57</td>
</tr>
<tr>
<td>DVD</td>
<td>52.21</td>
<td>63.23</td>
<td>66.18</td>
</tr>
<tr>
<td>IPOD</td>
<td>50.00</td>
<td>57.69</td>
<td>67.30</td>
</tr>
<tr>
<td>Mobile 1</td>
<td>51.16</td>
<td>61.63</td>
<td>66.28</td>
</tr>
<tr>
<td>Mobile 2</td>
<td>50.81</td>
<td>65.32</td>
<td>70.96</td>
</tr>
<tr>
<td>Music Player 1</td>
<td>50.30</td>
<td>57.62</td>
<td>64.37</td>
</tr>
<tr>
<td>Music Player 2</td>
<td>50.00</td>
<td>60.60</td>
<td>67.02</td>
</tr>
<tr>
<td>Router 1</td>
<td>50.00</td>
<td>58.33</td>
<td>61.67</td>
</tr>
<tr>
<td>Router 2</td>
<td>50.00</td>
<td>59.72</td>
<td>70.83</td>
</tr>
</tbody>
</table>
Overall Classification Accuracy (Dataset 2)

Lexicon-based Classification

Supervised Classification
# Overall Classification Accuracy (Dataset 2)

## Lexicon-based Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline$_1$</td>
<td>50.35</td>
</tr>
<tr>
<td>Baseline$_2$</td>
<td>58.93</td>
</tr>
<tr>
<td><strong>Proposed System</strong></td>
<td><strong>70.00</strong></td>
</tr>
</tbody>
</table>

## Supervised Classification
### Overall Classification Accuracy (Dataset 2)

#### Lexicon-based Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline&lt;sub&gt;1&lt;/sub&gt;</td>
<td>50.35</td>
</tr>
<tr>
<td>Baseline&lt;sub&gt;2&lt;/sub&gt;</td>
<td>58.93</td>
</tr>
<tr>
<td><strong>Proposed System</strong></td>
<td><strong>70.00</strong></td>
</tr>
</tbody>
</table>

#### Supervised Classification

<table>
<thead>
<tr>
<th>Domain</th>
<th>Baseline 1 (%)</th>
<th>Proposed System (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile</td>
<td>51.42 (50.72/99.29)</td>
<td><strong>83.82 (83.82/83.82)</strong></td>
</tr>
<tr>
<td>Camera</td>
<td>50</td>
<td><strong>86.99 (84.73/90.24)</strong></td>
</tr>
</tbody>
</table>
## Lexicon Classification Results
(Dataset 1, Lakkaraju et al., SDM 2011)

<table>
<thead>
<tr>
<th>System</th>
<th>Sentiment Evaluation Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline₁</td>
<td>68.75</td>
</tr>
<tr>
<td>Baseline₂</td>
<td>61.10</td>
</tr>
<tr>
<td>CFACTS-R</td>
<td>80.54</td>
</tr>
<tr>
<td>CFACTS</td>
<td>81.28</td>
</tr>
<tr>
<td>FACTS-R</td>
<td>72.25</td>
</tr>
<tr>
<td>FACTS</td>
<td>75.72</td>
</tr>
<tr>
<td>JST</td>
<td>76.18</td>
</tr>
<tr>
<td>Proposed System</td>
<td>80.98</td>
</tr>
</tbody>
</table>
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
Presence of words, at times, weighs more than the frequencies
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
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”
Motivation

- 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 interactive applications
  - Parsing does not work well in presence of noisy text

  "@user share 'em! i'm quite excited bout Tintin, despite not really 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.
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 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 uncertainty

---

*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 uncertainty

- **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 uncertainty

- 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 uncertainty

- Conditionals (If)

I do not like Nokia but I like Samsung.
Algorithm
(but, however, nevertheless, otherwise, yet, still, nonetheless, nevertheless, therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence)
Algorithm

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

- 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, 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

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

- \((\text{till, until, despite, in spite, though, although})\)

- All sentences containing \textit{if} are marked

- All sentences containing \textit{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
  - \textit{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 occurrence 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 occurrence of $w_i$ in the doc
  - $c_i$ is also incremented according to $w_i$’s importance
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 occurrence 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 occurrence 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

<table>
<thead>
<tr>
<th>Manually Annotated Dataset</th>
<th>#Positive</th>
<th>#Negative</th>
<th>#Objective Not Spam</th>
<th>#Objective Spam</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2548</td>
<td>1209</td>
<td>2757</td>
<td>1993</td>
<td>8507</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Auto Annotated Dataset</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Positive</td>
<td>#Negative</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7348</td>
<td>7866</td>
<td>15214</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- **Dataset 1 (Twitter – Manually Annotated)**
  - 8,507 tweets over 2,000 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

- **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)

### Manually Annotated Dataset

<table>
<thead>
<tr>
<th></th>
<th>#Positive</th>
<th>#Negative</th>
<th>#Objective Not Spam</th>
<th>#Objective Spam</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Positive</td>
<td>2,548</td>
<td>1,209</td>
<td>2,757</td>
<td>1,993</td>
<td>8,507</td>
</tr>
</tbody>
</table>

### Auto Annotated Dataset

<table>
<thead>
<tr>
<th></th>
<th>#Positive</th>
<th>#Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Positive</td>
<td>7,348</td>
<td>7,866</td>
<td>15,214</td>
</tr>
</tbody>
</table>

Dataset Domains

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

<table>
<thead>
<tr>
<th></th>
<th>Dataset 1</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-class</td>
<td>68.58</td>
<td>72.81</td>
<td>80.55</td>
</tr>
<tr>
<td>3-class</td>
<td>57.2</td>
<td>61.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-Feel-It</td>
<td>Discourse System</td>
<td></td>
</tr>
</tbody>
</table>

Bar chart showing comparison between C-Feel-It and Discourse System for classification in Twitter datasets 1 and 2.
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

<table>
<thead>
<tr>
<th>Systems</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Accuracy (Only Unigrams)</td>
<td>84.90</td>
</tr>
<tr>
<td>Balamurali et al., 2011 (Only IWSD Sense of Unigrams)</td>
<td>85.48</td>
</tr>
<tr>
<td>Balamurali et al., 2011 (Unigrams+IWSD Sense of Unigrams)</td>
<td>86.08</td>
</tr>
<tr>
<td><strong>Unigrams + IWSD Sense of Unigrams+Discourse Features</strong></td>
<td><strong>88.13</strong></td>
</tr>
</tbody>
</table>
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
“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

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
best remembered for his understated performance as dr. hannibal lec ter 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Unrelated Category

Movie Plot

- John Travolta is considered by many to be a has-been, or a one-hit wonder...
- Leonardo 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Unrelated Category

- Movie Plot
  - 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Unrelated Category

- Movie Plot

  *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

- 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Unrelated Category

Movie Plot
- 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Unrelated Category

Movie Plot

- 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
  - 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

- 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
- Opinion about the Characters in the Movie
- Characteristics of a movie or genre
- Opinion about the Movie and Crew
- Movie Plot

- 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

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.
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.
Harry Potter and the Deathly Hallows – Part 1 is a 2010 fantasy film 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. Part 1 was released in 2D cinemas and IMAX formats worldwide on 19 November 2010.

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.
Harry Potter and the Deathly Hallows – Part 1 is a 2010 fantasy film 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 Barro, and J. K. 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. Part 1 was released in 2D cinemas and IMAX formats worldwide on 19 November 2010.

Further information: Harry Potter and the Deathly Hallows Novel

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.

Daniel Radcliffe as Harry Potter, the film’s protagonist.

Rupert Grint as Ron Weasley, Harry’s best friend and Hermione’s romantic interest.

Emma Watson as Hermione Granger, Harry’s other best friend and Ron’s romantic interest.

Helena Bonham Carter as Bellatrix Lestrange, a Death Eater and Sirius Black’s cousin and murderer.

Robbie Coltrane as Rubeus Hagrid, Harry’s half-giant friend, and gamekeeper at Hogwarts.

Warwick Davis as Griphook, a goblin and former employee at Gringotts Bank. Davis replaced Verne Troyer, who portrayed the character physically in the first film, though Davis had dubbed Griphook’s lines.

Tom Felton as Draco Malfoy, a Death Eater and son of Lucius and Narcissa Malfoy.

Ralph Fiennes as Lord Voldemort, the film’s main antagonist.

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.
Harry Potter and the Deathly Hallows – Part 1 is a 2010 fantasy film 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. Part 1 was released in 2D cinemas and IMAX formats worldwide on 19 November 2010.

**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.

**CHARACTER**

Daniel Radcliffe as Harry Potter, the film’s protagonist.

Rupert Grint as Ron Weasley, Harry’s best friend and Hermione’s romantic interest.

Emma Watson as Hermione Granger, Harry’s other best friend and Ron’s romantic interest.

Helena Bonham Carter as Bellatrix Lestrange, a Death Eater and Sirius Black’s cousin and murderer.

Robbie Coltrane as Rubeus Hagrid, Harry’s half-giant friend, and gamekeeper at Hogwarts.

Warwick Davis as Griphook, a goblin and former employee at Gringotts Bank. Davis replaced Verne Troyer, who portrayed the character physically in the first film, though Davis had dubbed Griphook’s lines.

Tom Felton as Draco Malfoy, a Death Eater and son of Lucius and Narcissa Malfoy.

Ralph Fiennes as Lord Voldemort, the film’s main antagonist.

**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

**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, .....
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 to **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.*

- *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.

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

  - $Acc_{factor_i} = 1$ if $Rel_{factor_i} \geq 0$ and $\exists w_{ij} \in S_i$
    \[s.t. w_{ij} \in \text{Crew or MovieFeature or Movie Title}\]
    \[= 0\] otherwise
Equations 2 can be re-written as:

\[ \text{Rel}_{\text{factor}_i} = \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3} \]

\[ \text{Acc}_{\text{factor}_i} = \text{Rel}_{\text{factor}_i} - \theta \]

\[ = \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3} - \theta \]

\[ = \alpha \times X_{i,1} + \beta \times X_{i,2} - \gamma \times X_{i,3} - \theta \times X_{i,4} \text{ (where } X_{i,4} = 1) \]

Let \( Y_i \) be the binary label information corresponding to each sentence in the development set, where \( Y_i = 1 \) if \( \text{Acc}_{\text{factor}_i} \geq 0 \) and \(-1\) otherwise.

\[ Y_i = W \cdot X_i \quad \text{where,} \]

\[ W = [\alpha \ \beta \ -\gamma \ -\theta]^T \quad \text{and} \quad X_i = [X_{i,1} \ X_{i,2} \ X_{i,3} \ X_{i,4}] \]

or, \( Y = W^T \cdot X \)
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 \( \text{OpinionSummary} = \emptyset \)

for i=1..n
    if \( \text{Acc}_{\text{factor}_i} = 1 \)
        add \( S_i \) to \( \text{OpinionSummary} \)
    end if
end for
WikiSent

Input

Movie Review

Extractive Opinionated Summary

Polarity

Output

Mobile, Printer Reviews

Wikipedia

Frequent Words List

Metadata, Plot, Crew, Cast, Story Characters, Movie Features

Sentiment Lexicon

World Knowledge

Metadata, Plot, Crew, Cast, Story Characters, Movie Features

Frequent Words List

Wikipedia

World Knowledge
Algorithm Demonstration

- In Sentence [1], **Brian Cox** is the only keyword present and it belongs to the Cast list. $\text{Rel}_{\text{factor}} = 2*1 + 1*0 - 0*0 = 2 > -1$ 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. $\text{Rel}_{\text{factor}} = 2*2 + 1 - 1 = 4 > -1$ and is accepted.

- [5] has only the keyword **Big John** from Character and $\text{Rel}_{\text{factor}} = 0 + 0 - 1 = -1$ and is rejected.

- [6] has the keyword **audience** from MovieFeature and **Big John** from Character and its $\text{Rel}_{\text{factor}} = 0 + 1 - 1 = 0 > -1$ and is accepted.

- [7] has the keywords **temper, friend** from Plot and $\text{Rel}_{\text{factor}} = 0 + 0 - 2 = -2$ and is rejected.

- [8] has the keywords **sex, charm** from Plot and $\text{Rel}_{\text{factor}} = 0 + 0 - 2 = -2$ and is rejected.
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. \( Rel_{factor_i} = 2 \times 1 + 1 \times 0 - 0 \times 0 = 2 > -1 \) 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. \( Rel_{factor_i} = 2 \times 2 + 1 - 1 = 4 > -1 \) 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. $Rel_{factor_i} = 2 \times 1 + 1 \times 0 - 0 \times 0 = 2 > -1$ 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. $Rel_{factor_i} = 2 \times 2 + 1 - 1 = 4 > -1$ 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.
In Sentence [1], Brian Cox is the only keyword present and it belongs to the Cast list. \( R_{factor} = 2*1 + 1*0 - 0*0 = 2 > -1 \) 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. \( R_{factor} = 2*2 + 1 - 1 = 4 > -1 \) and is accepted.

[5] has only the keyword Big John from Character and \( R_{factor} = 0 + 0 - 1 = -1 \) and is rejected.

[6] has the keyword audience from MovieFeature and Big John from Character and its \( R_{factor} = 0 + 1 - 1 = 0 > -1 \) and is accepted.

[7] has the keywords temper, friend from Plot and \( R_{factor} = 0 + 0 - 2 = -2 \) and is rejected.

[8] has the keywords sex, charm from Plot and \( R_{factor} = 0 + 0 - 2 = -2 \) 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. \( R_{factor_i} = 2*1 + 1*0 - 0*0 = 2 > -1 \) 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. \( R_{factor_i} = 2*2 + 1 - 1 = 4 > -1 \) and is accepted.
- [5] has only the keyword Big John from Character and \( R_{factor_i} = 0 + 0 - 1 = -1 \) and is rejected.
- [6] has the keyword audience from MovieFeature and Big John from Character and its \( R_{factor_i} = 0 + 1 - 1 = 0 > -1 \) and is accepted.
- [7] has the keywords temper, friend from Plot and \( R_{factor_i} = 0 + 0 - 2 = -2 \) and is rejected.
- [8] has the keywords sex, charm from Plot and \( R_{factor_i} = 0 + 0 - 2 = -2 \) and is rejected.

What bothered some audience members was the presentation of Big John in an oddly empathetic light.
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.).
In Sentence [1], Brian Cox is the only keyword present and it belongs to the Cast list. \( R_{factor_i} = 2*1 + 1*0 - 0*0 = 2 > -1 \) 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. \( R_{factor_i} = 2*2 + 1 - 1 = 4 > -1 \) and is accepted.

[5] has only the keyword Big John from Character and \( R_{factor_i} = 0 + 0 - 1 = -1 \) and is rejected.

[6] has the keyword audience from MovieFeature and Big John from Character and its \( R_{factor_i} = 0 + 1 - 1 = 0 > -1 \) and is accepted.

[7] has the keywords temper, friend from Plot and \( R_{factor_i} = 0 + 0 - 2 = -2 \) and is rejected.

[8] has the keywords sex, charm from Plot and \( R_{factor_i} = 0 + 0 - 2 = -2 \) 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

- B1: Only SWN - 52.1
- B1: Only Subjectivity - 60.3
- B2: SentiWordNet-Full - 61.89
- B1: Only GI - 63.5
- B3: Turney [13] (SO Value) - 65.0
- B2: GI-Basic - 65.68
- B2: Subjectivity-Basic - 68.63
- WikiSent+Subjectivity - 71.1
- B3: Li [33] (40% doc label) - 73.5
- B3: Taboada [1] (SO-CAL Full) - 76.37
- WikiSent+GI - 76.85
- B3: Li [33] (10% doc label) - 60.0
- B3: Lin [32] (LSM Without Prior) - 61.7
- B2: SentiWordNet-Basic - 62.89
- B2: GI-Full - 64.21
- B2: Subjectivity-Full - 65.42
- B3: Taboada [1] (SO-CAL Basic) - 68.05
- B3: Shi [28], Dasgupta [29] (Eigen value Clustering) - 70.9
- WikiSent+SentiWordNet - 73.3
- B3: Lin [32] (LSM With Prior) - 74.1
- B3: Socher [30] (RAE Auto Encoders) - 76.80
Accuracy Comparison with Different Baselines
Accuracy Comparison with Different Baselines

<table>
<thead>
<tr>
<th>Baseline 1 : Simple Bag-of-Words</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Only SentiWordNet</td>
<td>52.1</td>
</tr>
<tr>
<td>Only Subjectivity</td>
<td>60.3</td>
</tr>
<tr>
<td>Only GI</td>
<td>63.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline 2 : Worse, Median and Best Performing Lexicons with SO-CAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiWordNet-Full</td>
<td>61.89</td>
</tr>
<tr>
<td>SentiWordNet-Basic</td>
<td>62.89</td>
</tr>
<tr>
<td>GI-Full</td>
<td>64.21</td>
</tr>
<tr>
<td>GI-Basic</td>
<td>65.68</td>
</tr>
<tr>
<td>Subjectivity-Full</td>
<td>65.42</td>
</tr>
<tr>
<td>Subjectivity-Basic</td>
<td>68.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WikiSent with Different Lexicons</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiSent+Subjectivity</td>
<td>71.1</td>
</tr>
<tr>
<td>WikiSent+SentiWordNet</td>
<td>73.3</td>
</tr>
<tr>
<td><strong>WikiSent+GI</strong></td>
<td><strong>76.85</strong></td>
</tr>
</tbody>
</table>
# Accuracy Comparison with Different Baselines

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

\[
\text{Genre Popularity} = \frac{\text{Positive Movie Reviews per Genre}}{\text{Total Movie Reviews per Genre}}
\]
Trend Analysis

Genre Popularity = \( \frac{\text{Positive Movie Reviews per Genre}}{\text{Total Movie Reviews per Genre}} \)
Trend Analysis

Genre Popularity = \[ \frac{\text{Positive Movie Reviews per Genre}}{\text{Total Movie Reviews per Genre}} \]
Trend Analysis

\[
\text{Genre Popularity} = \frac{\text{Positive Movie Reviews per Genre}}{\text{Total Movie Reviews per Genre}}
\]

Department of Computer Science and Engineering, IIT Bombay  7/23/2013
Trend Analysis

<table>
<thead>
<tr>
<th></th>
<th>WikiSent</th>
<th>Bag-of-Words Baseline 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Reviews (%)</td>
<td>48.95</td>
<td>81.2</td>
</tr>
<tr>
<td>Negative Reviews (%)</td>
<td>51.05</td>
<td>18.79</td>
</tr>
</tbody>
</table>
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
  - Promotional 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

Tweet Fetcher → Spam Filter → Spell Checker

Polarity Detector → Pragmatics Handler → Dependency Extractor

Tweets → Opinion

Department of Computer Science and Engineering, IIT Bombay 7/23/2013
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

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

- 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
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
   for each tweet $t_i \in M$ do
     Compute $\Pr[c_1 | t_i]$, $\Pr[c_2 | t_i]$ using the current NB // $c_1$ - non-spam class, $c_2$ - spam class
     $\Pr[c_2 | t_i] = 1 - \Pr[c_1 | t_i]$
     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)
   end for
2: end loop

$$\Pr[c_j | t_i] = \frac{\Pr[c_j] \prod_k \Pr[f_{i,k} | c_j]}{\sum_r \Pr[c_r] \prod_k P(f_{i,k} | c_r)}$$
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 happyyyyy 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 \( w \in S \) do
  - \( w' = \text{vowel_dropped}(w) \)
  - \( s' = \text{normalize}(s) \)
  - /*diff(s,w) gives difference of length between s and w*/
  - if \( \text{diff}(s', w') < \text{offset} \) then
    - score\[w\] = min(\text{edit_distance}(s, w), \text{edit_distance}(s, w'))
  - else
    - score\[w\] = max_centinel
  - end if
- end for
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, goooooood. More weightage is given by repeating them twice
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
Pragmatics

- **Elongation of a word, repeating alphabets multiple times** - Example: happppyyyyy, goooodle. 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: happpyyyyyyy, gooood. 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

<table>
<thead>
<tr>
<th>Tweets</th>
<th>Total Tweets</th>
<th>Correctly Classified</th>
<th>Misclassified</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>7007</td>
<td>3815</td>
<td>3192</td>
<td>54.45</td>
<td>55.24</td>
</tr>
<tr>
<td>Only spam</td>
<td>1993</td>
<td>1838</td>
<td>155</td>
<td>92.22</td>
<td>92.22</td>
</tr>
<tr>
<td>Only non-spam</td>
<td>5014</td>
<td>2259</td>
<td>2755</td>
<td>45.05</td>
<td>-</td>
</tr>
</tbody>
</table>

### 4-Class Classification

<table>
<thead>
<tr>
<th>Tweets</th>
<th>Total Tweets</th>
<th>Correctly Classified</th>
<th>Misclassified</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>7007</td>
<td>5010</td>
<td>1997</td>
<td>71.50</td>
<td>54.29</td>
</tr>
<tr>
<td>Only spam</td>
<td>1993</td>
<td>1604</td>
<td>389</td>
<td>80.48</td>
<td>80.48</td>
</tr>
<tr>
<td>Only non-spam</td>
<td>5014</td>
<td>4227</td>
<td>787</td>
<td>84.30</td>
<td>-</td>
</tr>
</tbody>
</table>
TwiSent Evaluation
TwiSent Evaluation

Lexicon-based Classification
TwiSent Evaluation

Lexicon-based Classification

![Bar chart showing comparison between C-Feel-It and TwiSent for different datasets.](chart.png)

- **Legend**:
  - C-Feel-It
  - TwiSent

- **Datasets**:
  - 2-Class Classification Dataset
  - 3-Class Classification Dataset
  - 2-Class Classification Dataset

The bar chart compares the performance of C-Feel-It and TwiSent for different datasets, indicating TwiSent's superior performance in the 2-Class Classification Dataset.
TwiSent Evaluation

Lexicon-based Classification

Supervised Classification
**TwiSent Evaluation**

### Lexicon-based Classification

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C-Feel-It</th>
<th>TwiSent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.19</td>
<td>64.92/69.37</td>
</tr>
<tr>
<td>2</td>
<td>50.8</td>
<td>53.16/72.96</td>
</tr>
</tbody>
</table>

### Supervised Classification

<table>
<thead>
<tr>
<th>System</th>
<th>2-class Accuracy</th>
<th>Precision/Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Feel-It</td>
<td>50.8</td>
<td>53.16/72.96</td>
</tr>
<tr>
<td>TwiSent</td>
<td>68.19</td>
<td>64.92/69.37</td>
</tr>
</tbody>
</table>
TwiSent Evaluation

Lexicon-based Classification

Supervised Classification

<table>
<thead>
<tr>
<th>System</th>
<th>2-class Accuracy</th>
<th>Precision/Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Feel-It</td>
<td>50.8</td>
<td>53.16/72.96</td>
</tr>
<tr>
<td>TwiSent</td>
<td>68.19</td>
<td>64.92/69.37</td>
</tr>
</tbody>
</table>
TwiSent Evaluation

Lexicon-based Classification

Ablation Test

<table>
<thead>
<tr>
<th>System</th>
<th>2-class Accuracy</th>
<th>Precision/Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Feel-It</td>
<td>50.8</td>
<td>53.16/72.96</td>
</tr>
<tr>
<td>TwiSent</td>
<td>68.19</td>
<td>64.92/69.37</td>
</tr>
</tbody>
</table>
TwiSent Evaluation

Lexicon-based Classification

Ablation Test

<table>
<thead>
<tr>
<th>Module Removed</th>
<th>Accuracy</th>
<th>Statistical Significance Confidence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity-Specificity</td>
<td>65.14</td>
<td>95</td>
</tr>
<tr>
<td>Spell-Checker</td>
<td>64.2</td>
<td>99</td>
</tr>
<tr>
<td>Pragmatics Handler</td>
<td>63.51</td>
<td>99</td>
</tr>
<tr>
<td>Complete System</td>
<td>66.69</td>
<td>-</td>
</tr>
</tbody>
</table>
Role of Social Media Content and World Knowledge in Information Retrieval

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

Department of Computer Science and Engineering, IIT Bombay    7/23/2013
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. ….
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
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, …
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

**Input:** Youtube Video URL

- **Title, Meta Description, User Comments**

- **System Defined Tags:** $Tag_1, Tag_2, \ldots Tag_n$

- **WordNet**
  - All Words

- **Wikipedia**
  - Named Entities

- **Thesaurus**

- **Seed-List:** $Seed-List_{Tag_1}, Seed-List_{Tag_2}, \ldots Seed-List_{Tag_n}$

- **Word:** Synset Vector

- **Named Entity:** Wiki Definition

- **Concept-List:** $Concept-List_{Tag_1}, Concept-List_{Tag_2}, \ldots Concept-List_{Tag_n}$

- **Unsupervised Tag Prediction**

**Output:** $Tag_p, Tag_q, \ldots Tag_t$
Tag Prediction

Feature Scoring

\[ \text{score}(f \in \text{genre}_k; w_1, w_2) = w_1 \times \sum_{j: \text{word}_j \in \text{seed}_k} 1 + w_2 \times \sum_{j: \text{word}_j \in \text{concept}_k} 1 \]

Video Genre Scoring

\[ \text{score}(\text{video} \in \text{genre}_k; p_1, p_2, p_3) = p_1 \times \text{score}(f^{\text{Title}} \in \text{genre}_k) + p_2 \times \text{score}(f^{\text{MetaData}} \in \text{genre}_k) + p_3 \times \text{score}(f^{\text{Comments}} \in \text{genre}_k) \]

Single Genre Prediction

\[ \text{video}_{\text{genre}} = \arg\max_k \text{score}(\text{video} \in \text{genre}_k) \]

Multiple Genre Prediction

\[ \text{video}_{\text{genre}} = k, \text{if } \text{score}(\text{video} \in \text{genre}_k) \geq \theta \]

\[ \text{where } \theta = \frac{1}{k} \sum_k \text{score}(\text{video} \in \text{genre}_k) \]
Genre-wise Results

- SGP: With User Comments and Without Wikipedia & WordNet
- SGP: With User Comments and With Wikipedia & WordNet
- MGP: With User Comments and With Wikipedia & WordNet

Department of Computer Science and Engineering, IIT Bombay 7/23/2013
## YouCat Dataset Statistics

### Video Statistics

<table>
<thead>
<tr>
<th></th>
<th>Comedy</th>
<th>Horror</th>
<th>Sports</th>
<th>Romance</th>
<th>Tech</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>2682</td>
<td>2802</td>
<td>2577</td>
<td>2477</td>
<td>2299</td>
<td>12837</td>
</tr>
</tbody>
</table>

### Comment Statistics

<table>
<thead>
<tr>
<th></th>
<th>Romance</th>
<th>Comedy</th>
<th>Horror</th>
<th>Sports</th>
<th>Tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>103.64</td>
<td>168.94</td>
<td>83.4</td>
<td>56.91</td>
<td>223.45</td>
</tr>
</tbody>
</table>
## Supervised Baseline

<table>
<thead>
<tr>
<th>SVM Features</th>
<th>F$_1$-Score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Unigrams</td>
<td>82.5116</td>
</tr>
<tr>
<td>Unigrams+Without stop words</td>
<td>83.5131</td>
</tr>
<tr>
<td>Unigrams+ Without stop words +Lemmatization</td>
<td>83.8131</td>
</tr>
<tr>
<td>Unigrams+Without stop words +Lemmatization+ POS Tags</td>
<td>83.8213</td>
</tr>
<tr>
<td>Top Unigrams+Without stop words +Lemmatization+POS Tags</td>
<td>84.0524</td>
</tr>
<tr>
<td>All Bigrams</td>
<td>74.2681</td>
</tr>
<tr>
<td><strong>Unigrams+Bigrams+Without stop words+Lemmatization</strong></td>
<td><strong>84.3606</strong></td>
</tr>
</tbody>
</table>
### Overall Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Average F₁ Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Baseline: Single Genre Prediction With User Comments</td>
<td>84.3606</td>
</tr>
<tr>
<td>Single Genre Prediction : Without User Comments + Without Wikipedia &amp; WordNet</td>
<td>68.76</td>
</tr>
<tr>
<td>Single Genre Prediction : With User Comments + Without Wikipedia &amp; WordNet</td>
<td>74.95</td>
</tr>
<tr>
<td>Single Genre Prediction : With User Comments+ With Wikipedia &amp;WordNet</td>
<td>80.9</td>
</tr>
<tr>
<td>Multiple Genre Prediction : Without User Comments + With Wikipedia &amp; WordNet</td>
<td>84.952</td>
</tr>
<tr>
<td>Multiple Genre Prediction : With User Comments + With Wikipedia &amp; WordNet</td>
<td>91.48</td>
</tr>
</tbody>
</table>
Comments/Genre and Confusion Matrix

Average Comments/Genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>Average Tags/Video Without User Comments</th>
<th>Average Tags/Video With User Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romance</td>
<td>1.45</td>
<td>1.55</td>
</tr>
<tr>
<td>Comedy</td>
<td>1.67</td>
<td>1.80</td>
</tr>
<tr>
<td>Horror</td>
<td>1.38</td>
<td>1.87</td>
</tr>
<tr>
<td>Sports</td>
<td>1.36</td>
<td>1.40</td>
</tr>
<tr>
<td>Tech</td>
<td>1.29</td>
<td>1.40</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>1.43</strong></td>
<td><strong>1.60</strong></td>
</tr>
</tbody>
</table>

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Romance</th>
<th>Comedy</th>
<th>Horror</th>
<th>Sports</th>
<th>Tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romance</td>
<td>80.16</td>
<td>8.91</td>
<td>3.23</td>
<td>4.45</td>
<td>3.64</td>
</tr>
<tr>
<td>Comedy</td>
<td>3.13</td>
<td>77.08</td>
<td>3.47</td>
<td>9.03</td>
<td>7.29</td>
</tr>
<tr>
<td>Horror</td>
<td>10.03</td>
<td>9.34</td>
<td>75.78</td>
<td>3.46</td>
<td>1.38</td>
</tr>
<tr>
<td>Sports</td>
<td>0.70</td>
<td>7.30</td>
<td>0</td>
<td>89.05</td>
<td>2.92</td>
</tr>
<tr>
<td>Tech</td>
<td>0.72</td>
<td>5.07</td>
<td>0.36</td>
<td>1.81</td>
<td>92.03</td>
</tr>
</tbody>
</table>
# Effect of Social Media Content in IR

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Romance</td>
<td>76.26</td>
<td>66.27</td>
</tr>
<tr>
<td>Comedy</td>
<td>43.96</td>
<td>40.00</td>
</tr>
<tr>
<td>Horror</td>
<td>80.47</td>
<td>68.67</td>
</tr>
<tr>
<td>Sports</td>
<td>84.21</td>
<td>68.71</td>
</tr>
<tr>
<td>Tech</td>
<td>90.83</td>
<td>73.50</td>
</tr>
</tbody>
</table>
## Effect of World Knowledge in IR

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Romance</td>
<td>76.06</td>
<td>70.63</td>
</tr>
<tr>
<td>Comedy</td>
<td>47.31</td>
<td>44.00</td>
</tr>
<tr>
<td>Horror</td>
<td>75.63</td>
<td>70.33</td>
</tr>
<tr>
<td>Sports</td>
<td>87.5</td>
<td>73.01</td>
</tr>
<tr>
<td>Tech</td>
<td>92.34</td>
<td>85.16</td>
</tr>
</tbody>
</table>
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
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)?
Introduce Sentiment as another feature in the Semantic Similarity Measure

“Among a set of 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

<table>
<thead>
<tr>
<th>Annotation Strategy</th>
<th>Overall</th>
<th>NOUN</th>
<th>VERB</th>
<th>ADJECTIVES</th>
<th>ADVERBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td>0.768</td>
<td>0.803</td>
<td>0.750</td>
<td>0.527</td>
<td>0.759</td>
</tr>
<tr>
<td>Meaning + Sentiment</td>
<td>0.799</td>
<td>0.750</td>
<td>0.889</td>
<td>0.720</td>
<td>0.844</td>
</tr>
<tr>
<td>POS</td>
<td>WordNet relations used for expansion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td>hypernym, hyponym, nominalization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs</td>
<td>nominalization, hypernym, hyponym</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjectives</td>
<td>also see, nominalization, attribute</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverbs</td>
<td>derived</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Scoring Formula

- $\text{Score}_{SD}(A) = \text{SWN}_{pos}(A) - \text{SWN}_{neg}(A)$
- $\text{Score}_{SM}(A) = \max(\text{SWN}_{pos}(A), \text{SWN}_{neg}(A))$
- $\text{Score}_{TM}(A) =$ 
  
  $\text{sign}(\max(\text{SWN}_{pos}(A), \text{SWN}_{neg}(A))) \times \left(1 + \abs{\max(\text{SWN}_{pos}(A), \text{SWN}_{neg}(A))}\right)$

$\text{SenSim}_x(A, B) = \text{cosine} (\text{gloss}_{vec}(\text{sense}(A)), \text{gloss}_{vec}(\text{sense}(B)))$

Where,

$\text{gloss}_{vec} = 1:\text{score}_x(1) 2:\text{score}_x(2) \ldots n:\text{score}_x(n)$

$\text{score}_x(Y) = \text{Sentiment score of word Y using scoring function } x$

$x = \text{Scoring function of type SD/SM/TD/TM}$
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

<table>
<thead>
<tr>
<th>Metric Used</th>
<th>Overall</th>
<th>NOUN</th>
<th>VERB</th>
<th>ADJECTIVES</th>
<th>ADVERBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LESK (Banerjee et al., 2003)</td>
<td>0.22</td>
<td>0.51</td>
<td>-0.91</td>
<td>0.19</td>
<td>0.37</td>
</tr>
<tr>
<td>LIN (Lin, 1998)</td>
<td>0.27</td>
<td>0.24</td>
<td>0.00</td>
<td>NA</td>
<td>Na</td>
</tr>
<tr>
<td>LCH (Leacock et al., 1998)</td>
<td>0.36</td>
<td>0.34</td>
<td>0.44</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SenSim (SD)</td>
<td>0.46</td>
<td>0.73</td>
<td>0.55</td>
<td>0.08</td>
<td>0.76</td>
</tr>
<tr>
<td>SenSim (SM)</td>
<td><strong>0.50</strong></td>
<td>0.62</td>
<td>0.48</td>
<td>0.06</td>
<td>0.54</td>
</tr>
<tr>
<td>SenSim (TD)</td>
<td>0.45</td>
<td>0.73</td>
<td>0.55</td>
<td>0.08</td>
<td>0.59</td>
</tr>
<tr>
<td>SenSim (TM)</td>
<td>0.48</td>
<td>0.62</td>
<td>0.48</td>
<td>0.06</td>
<td>0.78</td>
</tr>
</tbody>
</table>
## Evaluation on Travel Review Data: Feature Replacement

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


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

Thank You