An Opinion Mining System

Student Name: Robertus Primusanto

Student ID: 40218716

Supervisor: Dr Mark Dras

Date: 13th November 2009

WHAT IS OPINION MINING

- ***Opinion Mining*** is an activity to classify reviews according to their subjectivity and determine the author critical sentiment viewpoint.

AGENDA

- Introduction
- Previous Work
- Objectives
- Methodology
- Result and Discussion
- Conclusion

AGENDA

- Introduction
  - Background
- Previous Work
- Objectives
- Methodology
- Result and Discussion
- Conclusion

1. INTRODUCTION

- Many people published their writing to express their opinion
- The rapid growth of the internet in the last 10 years
- Private and Public organisations sees competitive advantage from opinion mining

1. INTRODUCTION

- Many people published their writing to express their opinion
- The rapid growth of the internet in the last 10 years
- Private and Public organisations sees competitive advantage from opinion mining

facebook twitter IMDb Epinions.com

1. INTRODUCTION

EXAMPLE

- Imagine If you have to classify a review

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format. They look rather like file cards and you can scroll through them and activate the one you want to use. Yeah, impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the task bar that correspond with various windows, huh?”

- Now Imagine If you have to classify many reviews

1. INTRODUCTION

INFORMATION OVERLOAD

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format. They look rather like file cards and you can scroll through them and activate the one you want to use. Yeah, impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the task bar that correspond with various windows, huh?”

Third party software fared worse. Mozilla Firefox became almost unusable under Vista as it simply failed whenever a

1. INTRODUCTION

INFORMATION OVERLOAD

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format. They look rather like file cards and you can scroll through them and activate the one you want to use. Yeah, impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the task bar that correspond with various windows, huh?”

Third party software fared worse. Mozilla Firefox became almost unusable under Vista as it simply failed whenever a Java application was run. At my office, we use a program to called "Zeta Fax" to send faxes directly from our computers. That didn't work under Vista, either. I work for a trade association and the software we use to access our members' information just wouldn't work under Vista. My laser printer wouldn't work under Vista, nor would the network printers we have at the office (one of those machines is vital to what I do on a daily basis and I doubt it will be replaced anytime soon as it is less than a year old and cost about $17,000). In fact, getting Vista to talk to our network at all proved to be a challenge, although the tech guy got that figured out soon enough.

Well, you may be asking yourself about customer support. There's not much to be had with Vista, sadly. Whenever a program would crash, a dialog box would pop up asking if I wanted to search the Internet for a solution to the problem I had. I had to see that message a lot and the only "solution" that was found for anything was for Outlook. Everything else (including a fix for the print spooler that came with Vista) yielded absolutely nothing.

Third party software fared worse. Mozilla Firefox became almost unusable under Vista as it simply failed whenever a Java application was run. At my office, we use a program to called "Zeta Fax" to send faxes directly from our computers. That didn't work under Vista, either. I work for a trade association and the software we use to access our members' information just wouldn't work under Vista. My laser printer wouldn't work under Vista, nor would the network printers we have at the office (one of those machines is vital to what I do on a daily basis and I doubt it will be replaced anytime soon as it is less than a year old and cost about $17,000). In fact, getting Vista to talk to our network at all proved to be a challenge, although the tech guy got that figured out soon enough.

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format. They look rather like file cards and you can scroll through them and activate the one you want to use. Yeah, impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the task bar that correspond with various windows, huh?”

Third party software fared worse. Mozilla Firefox became almost unusable under Vista as it simply failed whenever a

Third party software fared worse. Mozilla Firefox became almost unusable under Vista as it simply failed whenever a Java application was run. At my office, we use a program to called "Zeta Fax" to send faxes directly from our computers That didn't work under Vista, either. I work for a trade association and the software we use to access our members' information just wouldn't work under Vista. My laser printer wouldn't work under Vista, nor would the network printers we have at the office (one of those machines is vital to what I do on a daily basis and I doubt it will be replaced anytime soon as it is less than a year old and cost about $17,000). In fact, getting Vista to talk to our network at all proved to be challenge, although the tech guy got that figured out soon enough.

Well, you may be asking yourself about customer support. There's not much to be had with Vista, sadly. Whenever a program would crash, a dialog box would pop up asking if I wanted to search the Internet for a solution to the problem had. I had to see that message a lot and the only "solution" that was found for anything was for Outlook. Everything e (including a fix for the print spooler that came with Vista) yielded absolutely nothing.

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format They look rather like file cards and you can scroll through them and activate the one you want to use. Yea impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the t bar that correspond with various windows, huh?”

“To sum everything up, think of Vista as a copy of Windows XP that's been broken to the point it's almost unusable and with a gimmick thrown in for good measure. That gimmick is called the Aero graphical user interface -- you get a few dandy, bright icons and a feature that's novel at first but almost useless upon further investigation. That feature, simply put, allows you to click an icon on your task bar then watch in wonder as representations of all the windows you have open are presented in a three-dimensional format They look rather like file cards and you can scroll through them and activate the one you want to use. Yea impressive stuff. I suppose it was too much of a stretch to simply ask the user to click the buttons on the t bar that correspond with various windows, huh?”

Well, you may be asking yourself about customer support. There's not much to be had with Vista, sadly. Whenever a program would crash, a dialog box would pop up asking if I wanted to search the Internet for a solution to the problem had. I had to see that message a lot and the only "solution" that was found for anything was for Outlook. Everything e (including a fix for the print spooler that came with Vista) yielded absolutely nothing.

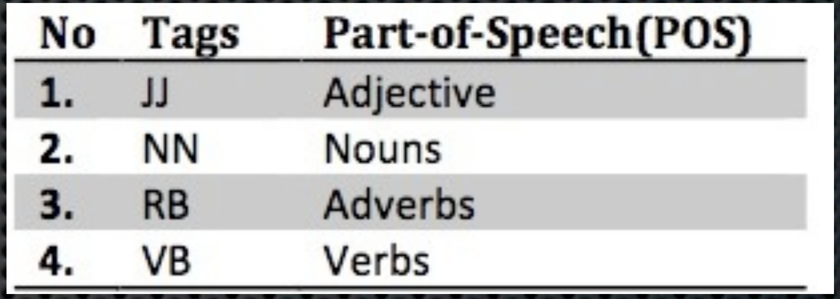
AGENDA

- Introduction
- Previous Work
  - Thumbs up or Thumbs down - an unsupervised review classification
- Objectives
- Methodology
- Result and Discussion
- Conclusion

2. PREVIOUS WORK

Thumbs Up or Thumbs Down

- Turney (2002) introduce an algorithm to classify review recommendation from its Keywords



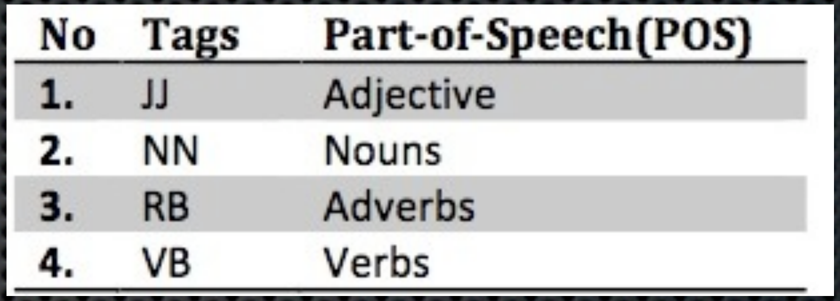
No	Tags	Part-of-Speech(POS)
1.	JJ	Adjective
2.	NN	Nouns
3.	RB	Adverbs
4.	VB	Verbs

- Three steps:
  - Classifying and identifying phrase
  - Estimate Semantic Orientation (SO)
    - PMI - IR (“poor” & “excellent”)
  - Classify a review
- Results:
  - Average 74% accuracy
- Other technique -machine learning described by Pang & Lee (2002)

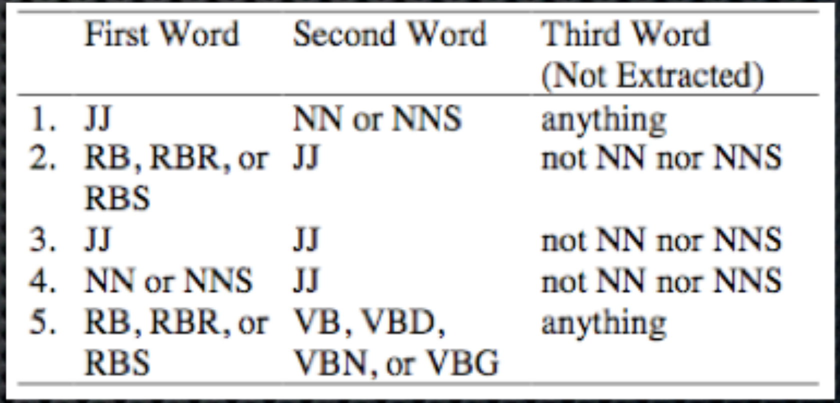
2. PREVIOUS WORK

Thumbs Up or Thumbs Down

- Turney (2002) introduce an algorithm to classify review recommendation from its Keywords
- Three steps:
  - Classifying and identifying phrase
  - Estimate Semantic Orientation (SO)
    - PMI - IR (“poor” & “excellent”)
  - Classify a review
- Results:
  - Average 74% accuracy
- Other technique -machine learning described by Pang & Lee (2002)



No	Tags	Part-of-Speech(POS)
1.	JJ	Adjective
2.	NN	Nouns
3.	RB	Adverbs
4.	VB	Verbs

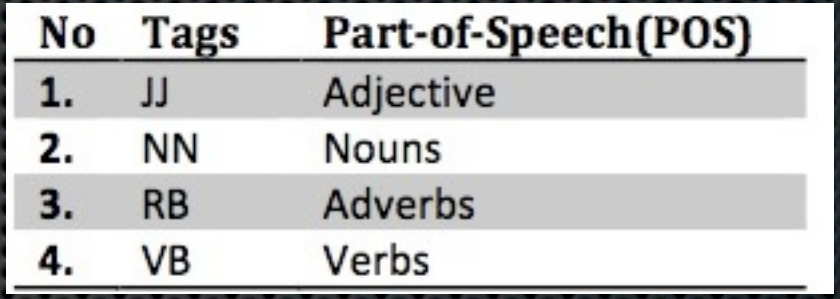


	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

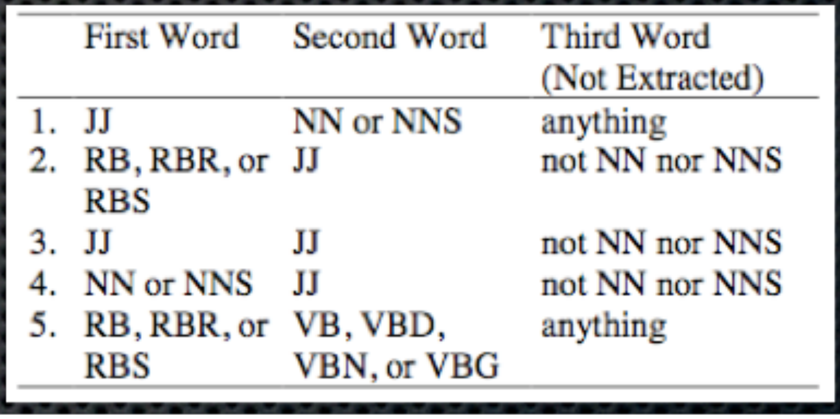
2. PREVIOUS WORK

Thumbs Up or Thumbs Down

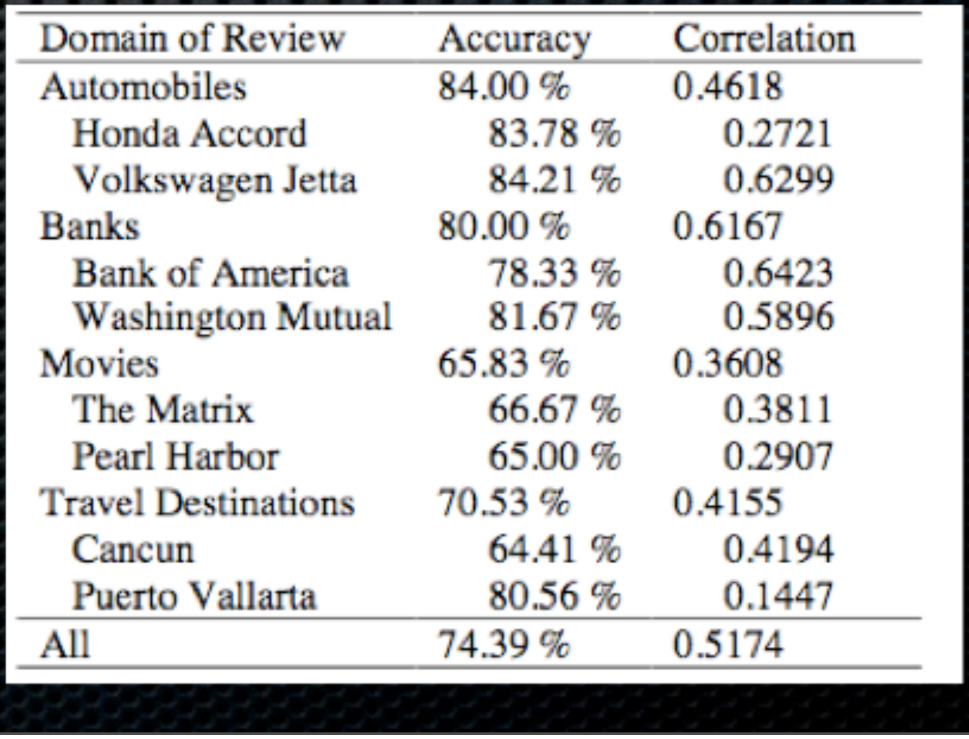
- Turney (2002) introduce an algorithm to classify review recommendation from its Keywords
- Three steps:
  - Classifying and identifying phrase
  - Estimate Semantic Orientation (SO)
    - PMI - IR (“poor” & “excellent”)
  - Classify a review
- Results:
  - Average 74% accuracy
- Other technique -machine learning described by Pang & Lee (2002)



No	Tags	Part-of-Speech(POS)
1.	JJ	Adjective
2.	NN	Nouns
3.	RB	Adverbs
4.	VB	Verbs



	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything



Domain of Review	Accuracy	Correlation
Automobiles	84.00 %	0.4618
Honda Accord	83.78 %	0.2721
Volkswagen Jetta	84.21 %	0.6299
Banks	80.00 %	0.6167
Bank of America	78.33 %	0.6423
Washington Mutual	81.67 %	0.5896
Movies	65.83 %	0.3608
The Matrix	66.67 %	0.3811
Pearl Harbor	65.00 %	0.2907
Travel Destinations	70.53 %	0.4155
Cancun	64.41 %	0.4194
Puerto Vallarta	80.56 %	0.1447
All	74.39 %	0.5174

AGENDA

- Introduction
- Previous Work
- Objectives
- Methodology
- Result and Discussion
- Conclusion

2. OBJECTIVES

- To developed an Opinion Mining System which is based on Turney(2002) algorithm
- To evaluate the effect of approximaton from Google Web1T (will be explained in Methodology) to derived statistical data for PMI Calculation
- To assess the effect of approximation in using different window sizes in Google Web1T

AGENDA

- Introduction
- Previous Work
- Objectives
- Methodology
  - Dataset/Text Classification/Phrase Identification/PMI Calculation/SO Estimation/ Review Recommendation
- Result and Discussion
- Conclusion

3. METHODOLOGY

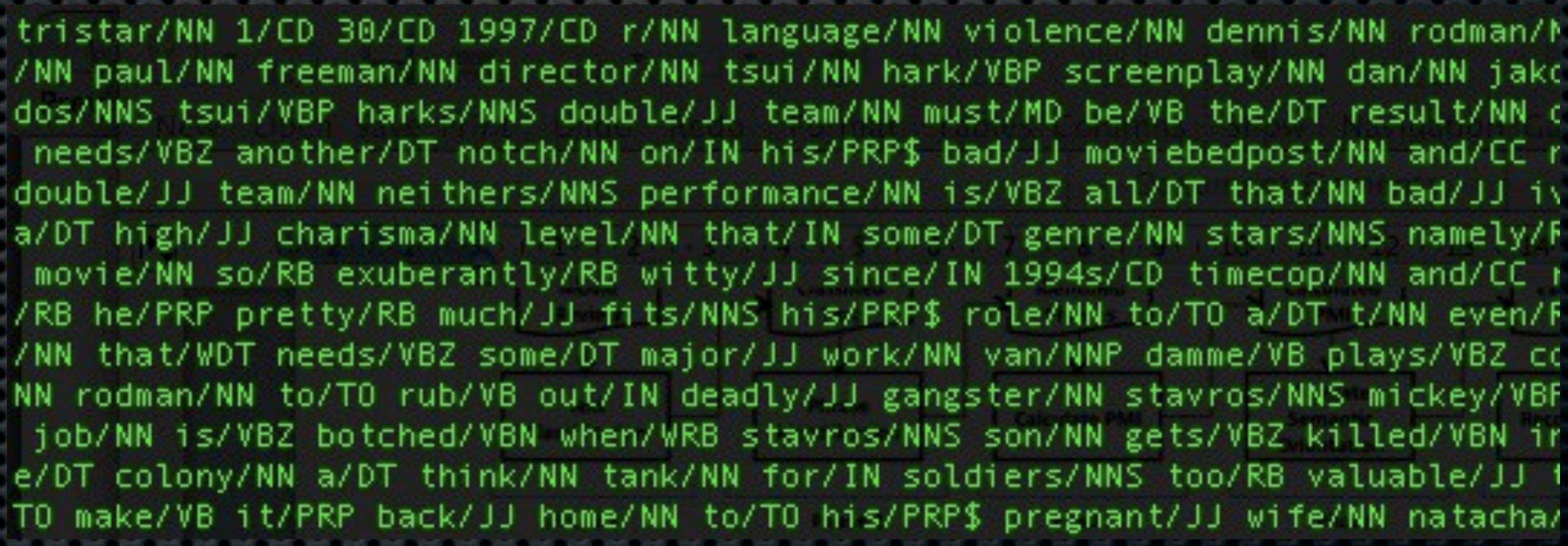
DataSet

- Test Review
  - Data from Pang, Lee, et al. (2002) experiment
  - 1400 reviews of which equally distributed to positive and negative
- Text Corpus
  - Google Web1T data
  - Approx. 24 Gb of data
  - Have different word window unigram to 5gms

3. METHODOLOGY

Text Classification

- Input: Movie Review
- Process: Part of speech Tagging Using Brill Tagger
- Output: Classified Text

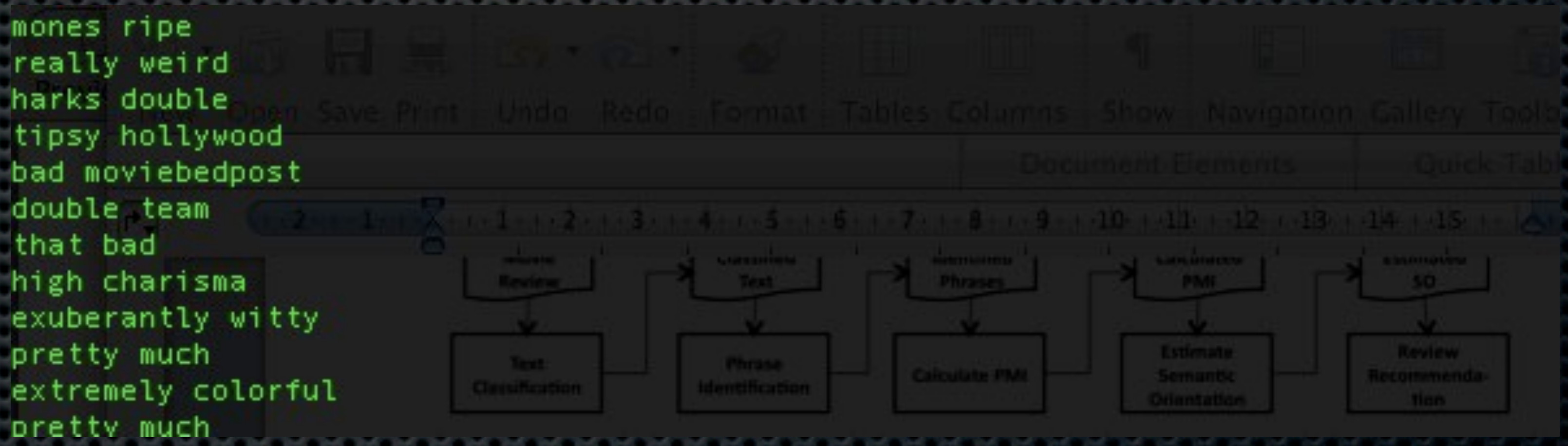


```
tristar/NN 1/CD 30/CD 1997/CD r/NN language/NN violence/NN dennis/NN rodman/N
/NN paul/NN freeman/NN director/NN tsui/NN hark/VBP screenplay/NN dan/NN jakc
dos/NNS tsui/VBP harks/NNS double/JJ team/NN must/MD be/VB the/DT result/NN c
 needs/VBZ another/DT notch/NN on/IN his/PRP$ bad/JJ moviebedpost/NN and/CC r
double/JJ team/NN neithers/NNS performance/NN is/VBZ all/DT that/NN bad/JJ i
a/DT high/JJ charisma/NN level/NN that/IN some/DT genre/NN stars/NNS namely/R
 movie/NN so/RB exuberantly/RB witty/JJ since/IN 1994s/CD timecop/NN and/CC r
/RB he/PRP pretty/RB much/JJ fits/NNS his/PRP$ role/NN to/TO a/DT t/NN even/R
/NN that/WDT needs/VBZ some/DT major/JJ work/NN van/NNP damme/VB plays/VBZ cc
NN rodman/NN to/TO rub/VB out/IN deadly/JJ gangster/NN stavros/NNS mickey/VBF
 job/NN is/VBZ botched/VBN when/WRB stavros/NNS son/NN gets/VBZ killed/VBN ir
e/DT colony/NN a/DT think/NN tank/NN for/IN soldiers/NNS too/RB valuable/JJ t
TO make/VB it/PRP back/JJ home/NN to/TO his/PRP$ pregnant/JJ wife/NN natacha/
```

3. METHODOLOGY

Phrase Identification

- Input: Classified Text
- Process: Extract word according to combination pattern
- Output: Identified Phrase

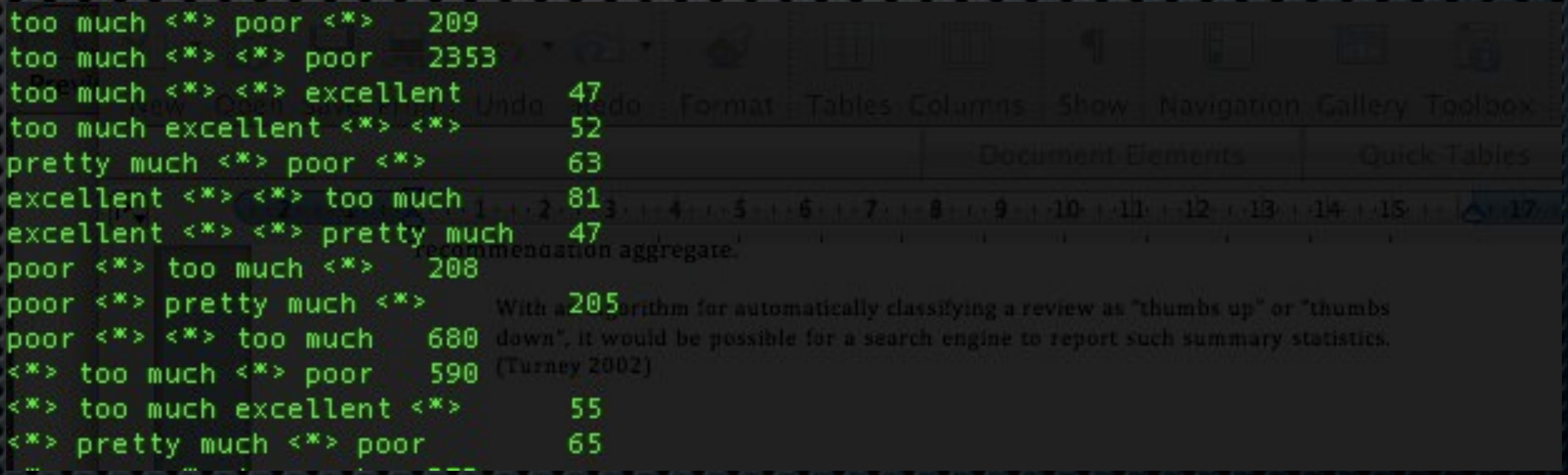


```
mones ripe
really weird
harks double
tipsy hollywood
bad moviebedpost
double team
that bad
high charisma
exuberantly witty
pretty much
extremely colorful
pretty much
```

3. METHODOLOGY

PMI Calculation

- Input: Identified Phrase
- Process: Calculate PMI of a phrase using Google Web1T (Get1T)
- Output: Calculated PMI

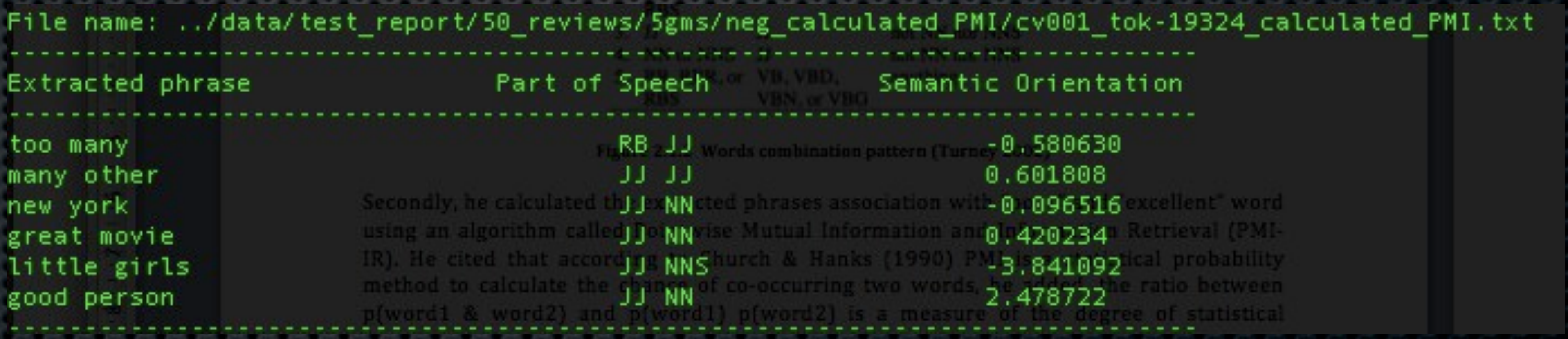


```
too much <*> poor <*>   209
too much <*> <*> poor   2353
too much <*> <*> excellent      47
too much excellent <*> <*>      52
pretty much <*> poor <*>        63
excellent <*> <*> too much      81
excellent <*> <*> pretty much   47
poor <*> too much <*>   208
poor <*> pretty much <*>        205
poor <*> <*> too much   680
<*> too much <*> poor   590
<*> too much excellent <*>      55
<*> pretty much <*> poor        65
```

3. METHODOLOGY

SO Estimation

- Input: Calculated PMI
- Process: Calculate the average of PMI result
- Output: a list of identified phrase, its POS, and its SO

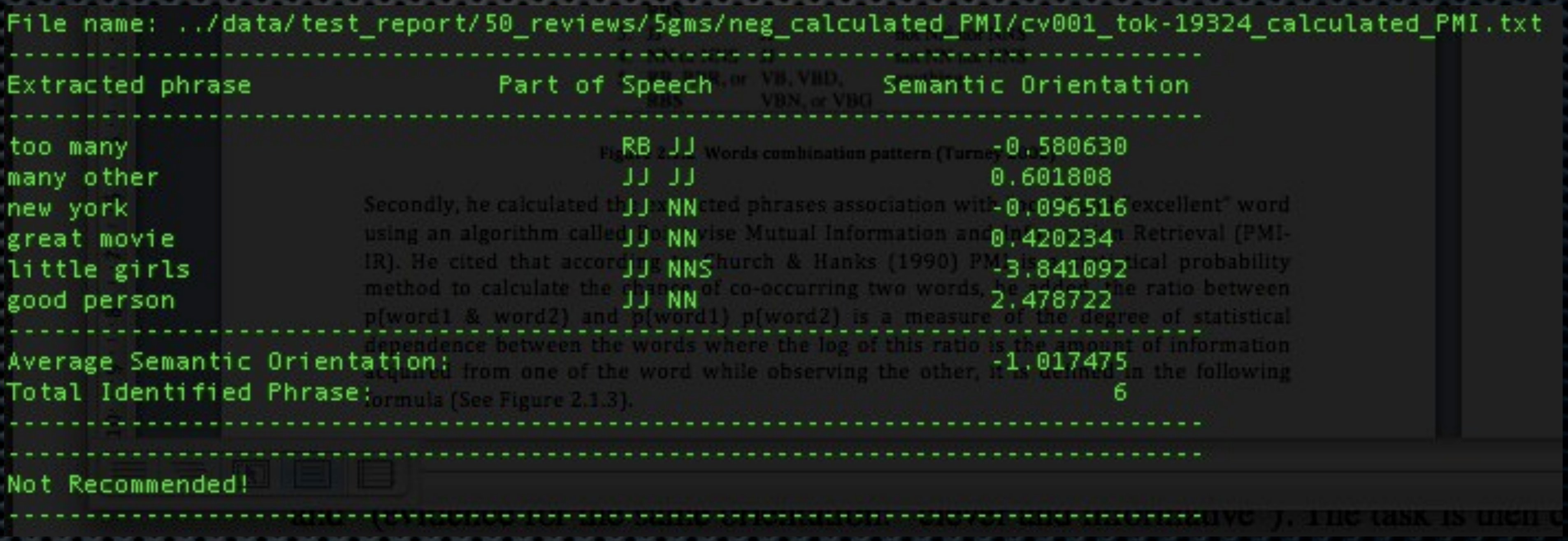


```
File name: ../data/test_report/50_reviews/5gms/neg_calculated_PMI/cv001_tok-19324_calculated_PMI.txt
----------------------------------------------------------------------
Extracted phrase            Part of Speech        Semantic Orientation
----------------------------------------------------------------------
too many                           RB JJ                -0.580630
many other                         JJ JJ                0.601808
new york                           JJ NN                -0.096516
great movie                        JJ NN                0.420234
little girls                       JJ NNS               -3.841092
good person                        JJ NN                2.478722
----------------------------------------------------------------------
```

3. METHODOLOGY

Review Recommendation

- Input: Calculated PMI
- Process: Calculate total SO of each phrase
- Output: review recommendation



```
File name: ../data/test_report/50_reviews/5gms/neg_calculated_PMI/cv001_tok-19324_calculated_PMI.txt
------------------------------------------------------------------------------
Extracted phrase                Part of Speech           Semantic Orientation
------------------------------------------------------------------------------
too many                                RB JJ               -0.580630
many other                              JJ JJ                0.601808
new york                                JJ NN               -0.096516
great movie                             JJ NN                0.420234
little girls                            JJ NNS              -3.841092
good person                             JJ NN                2.478722
------------------------------------------------------------------------------
Average Semantic Orientation:                               -1.017475
Total Identified Phrase:                                            6
------------------------------------------------------------------------------
------------------------------------------------------------------------------
Not Recommended!
------------------------------------------------------------------------------
```

AGENDA

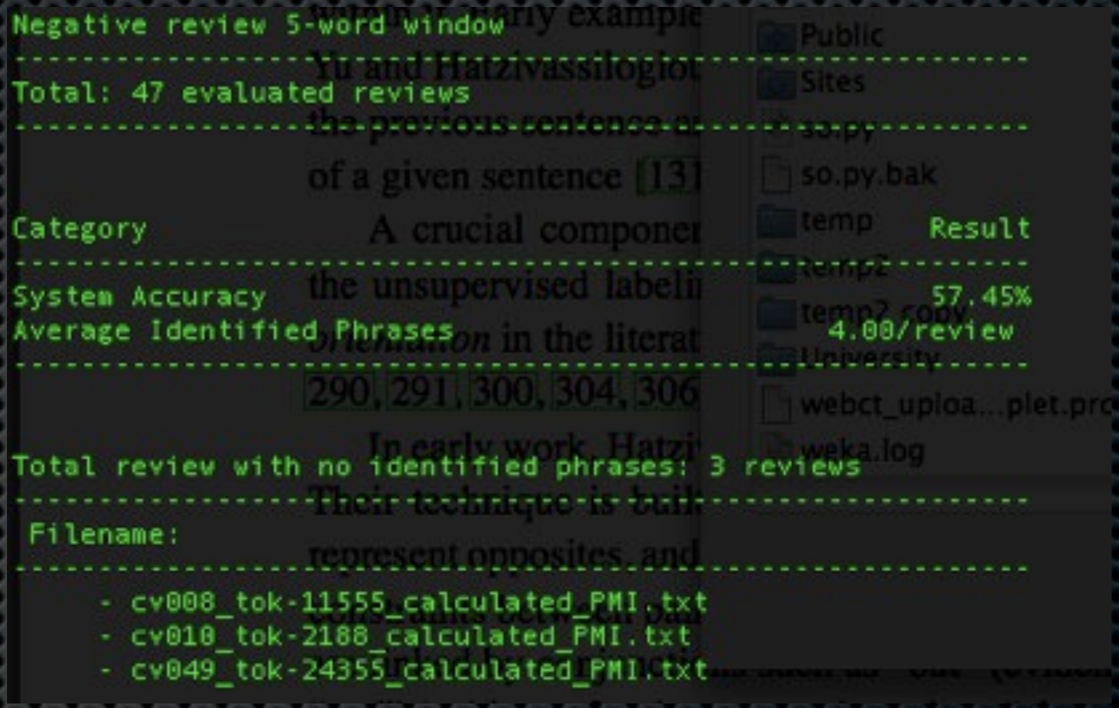
- Introduction
- Previous Work
- Objectives
- Methodology
- Result and Discussion
  - System Accuracy/System Evaluation/Effect of using different word window sizes
- Conclusion

4.RESULTS AND DISCUSSION

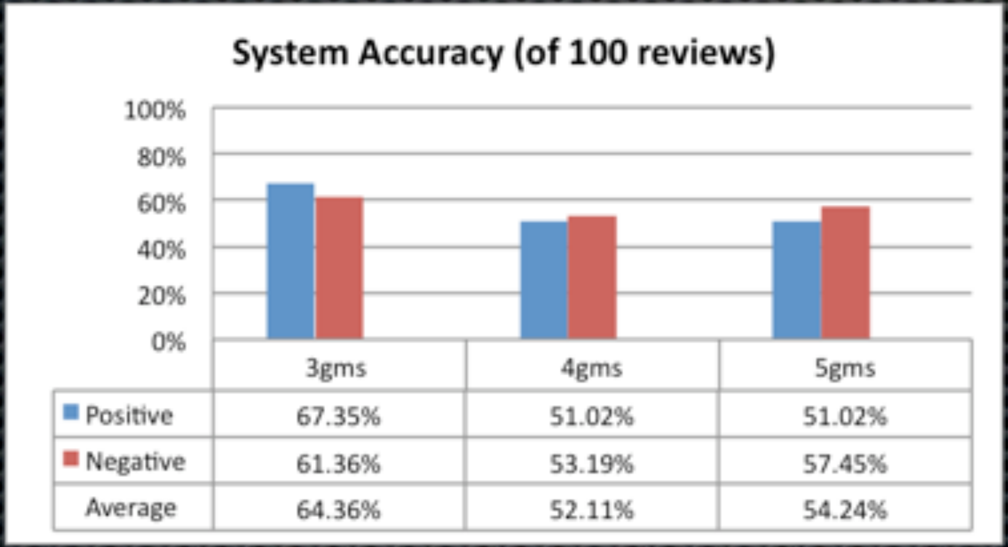
System Accuracy

- Using 50 negative reviews and 5 word windows produce an average accuracy of 54%

- A matrix of 100 reviews and their accuracy

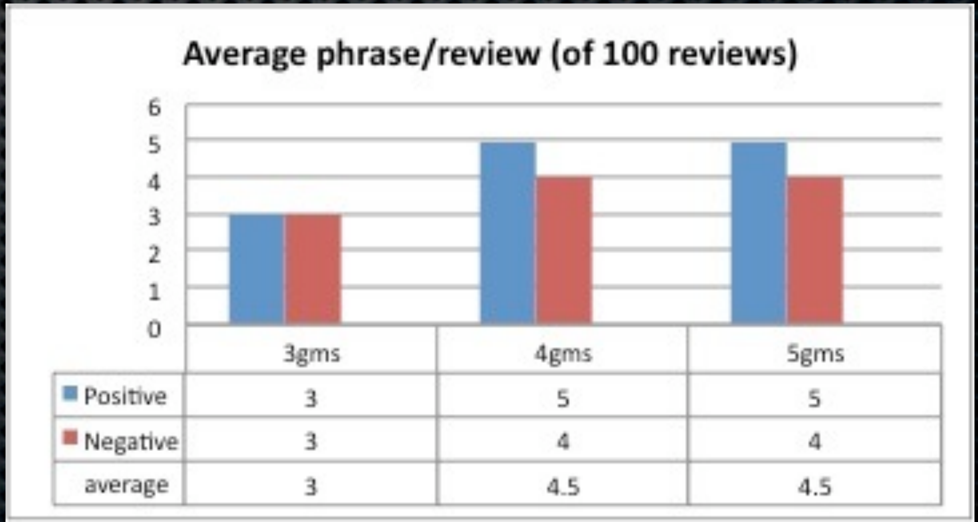


```
Negative review 5-word window
-------------------------------------------------------------------------
Total: 47 evaluated reviews
-------------------------------------------------------------------------

Category                                                           Result
-------------------------------------------------------------------------
System Accuracy                                                    57.45%
Average Identified Phrases                                    4.00/review
-------------------------------------------------------------------------


Total review with no identified phrases: 3 reviews
-------------------------------------------------------------------------
 Filename:
-------------------------------------------------------------------------
     - cv008_tok-11555_calculated_PMI.txt
     - cv010_tok-2188_calculated_PMI.txt
     - cv049_tok-24355_calculated_PMI.txt
```



System Accuracy (of 100 reviews)

	3gms	4gms	5gms
🟦 Positive	67.35%	51.02%	51.02%
🟥 Negative	61.36%	53.19%	57.45%
Average	64.36%	52.11%	54.24%



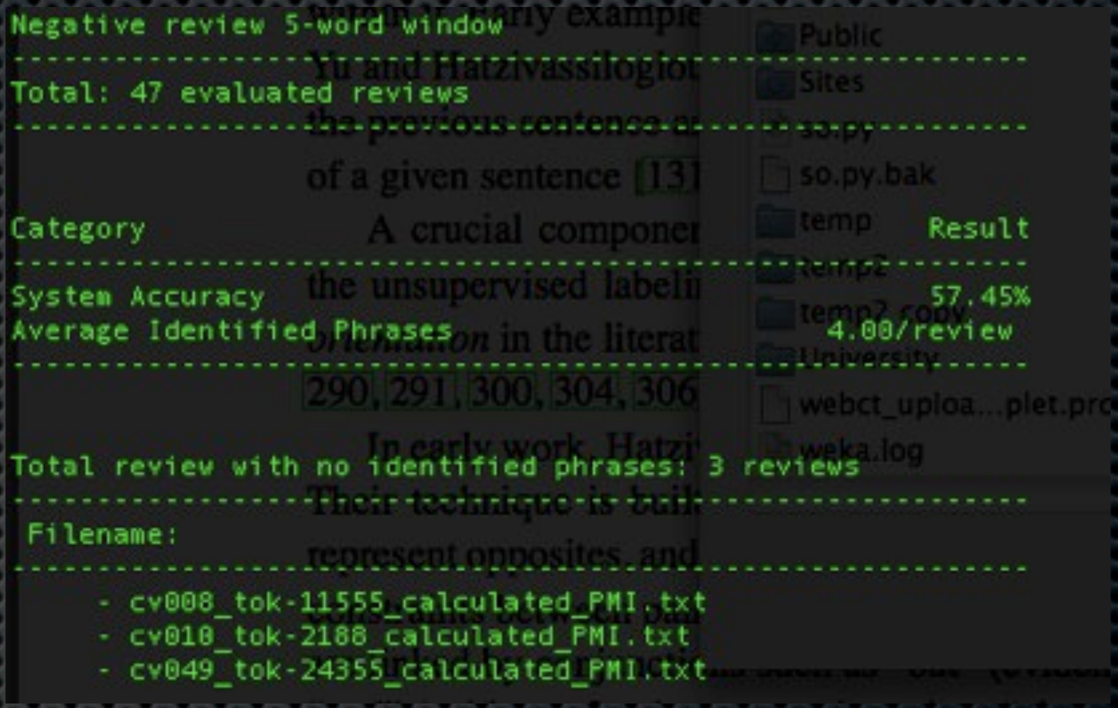
Average phrase/review (of 100 reviews)

	3gms	4gms	5gms
🟦 Positive	3	5	5
🟥 Negative	3	4	4
average	3	4.5	4.5

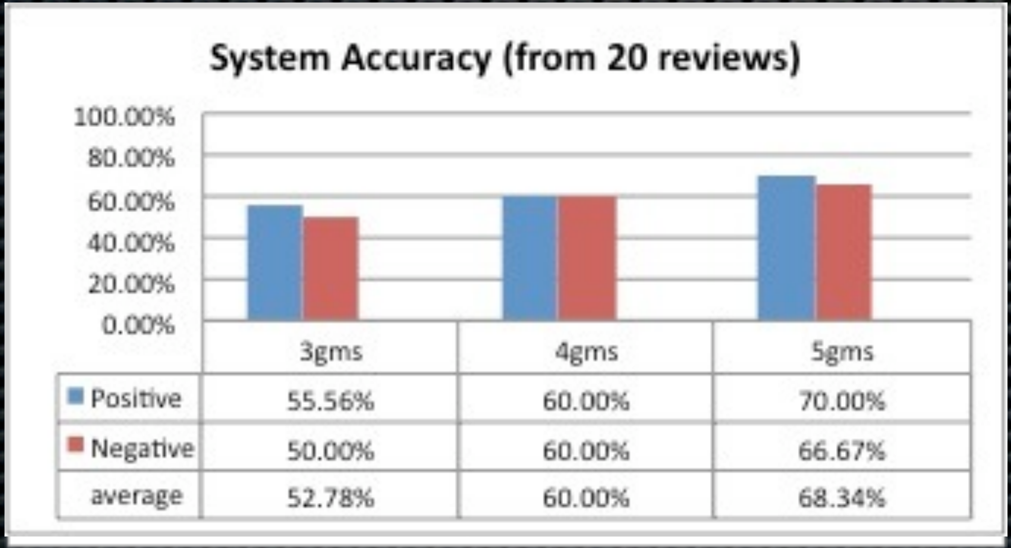
4.RESULTS AND DISCUSSION

System Accuracy

- Using 50 negative reviews and 5 word windows produce an average accuracy of 54%
- A matrix of 100 reviews and their accuracy

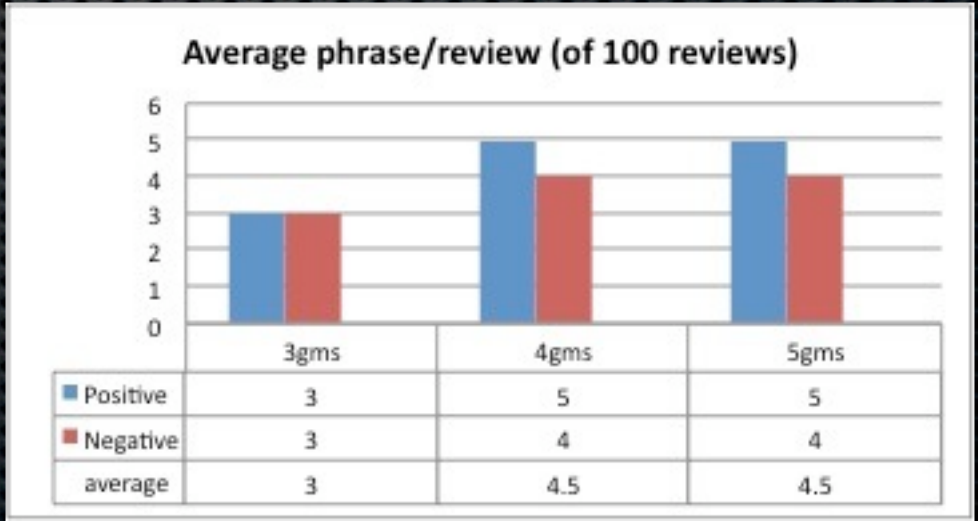


```
Negative review 5-word window
------------------------------------------------------------
Total: 47 evaluated reviews
------------------------------------------------------------

Category                                               Result
------------------------------------------------------------
System Accuracy                                        57.45%
Average Identified Phrases                        4.00/review
------------------------------------------------------------


Total review with no identified phrases: 3 reviews
------------------------------------------------------------
 Filename:
------------------------------------------------------------
     - cv008_tok-11555_calculated_PMI.txt
     - cv010_tok-2188_calculated_PMI.txt
     - cv049_tok-24355_calculated_PMI.txt
```



System Accuracy (from 20 reviews)

	3gms	4gms	5gms
🟦 Positive	55.56%	60.00%	70.00%
🟥 Negative	50.00%	60.00%	66.67%
average	52.78%	60.00%	68.34%

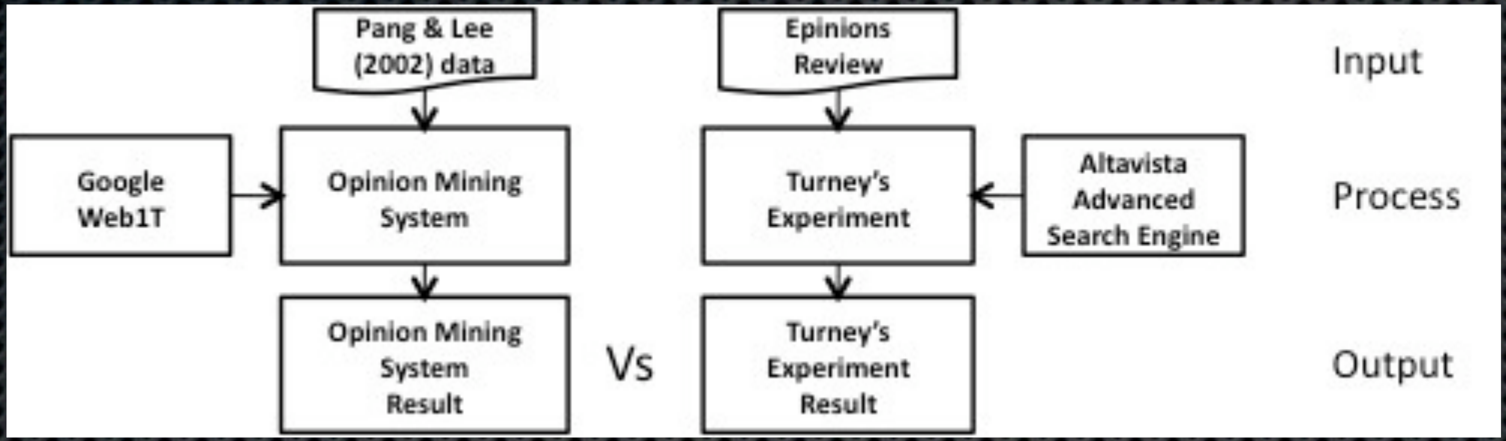


Average phrase/review (of 100 reviews)

	3gms	4gms	5gms
🟦 Positive	3	5	5
🟥 Negative	3	4	4
average	3	4.5	4.5

4.RESULTS AND DISCUSSION

System Evaluation



- Better than deliberately assigning all positive by 4%
- Lower than the golden standard by 10%
- Overal Google Web1T do its job

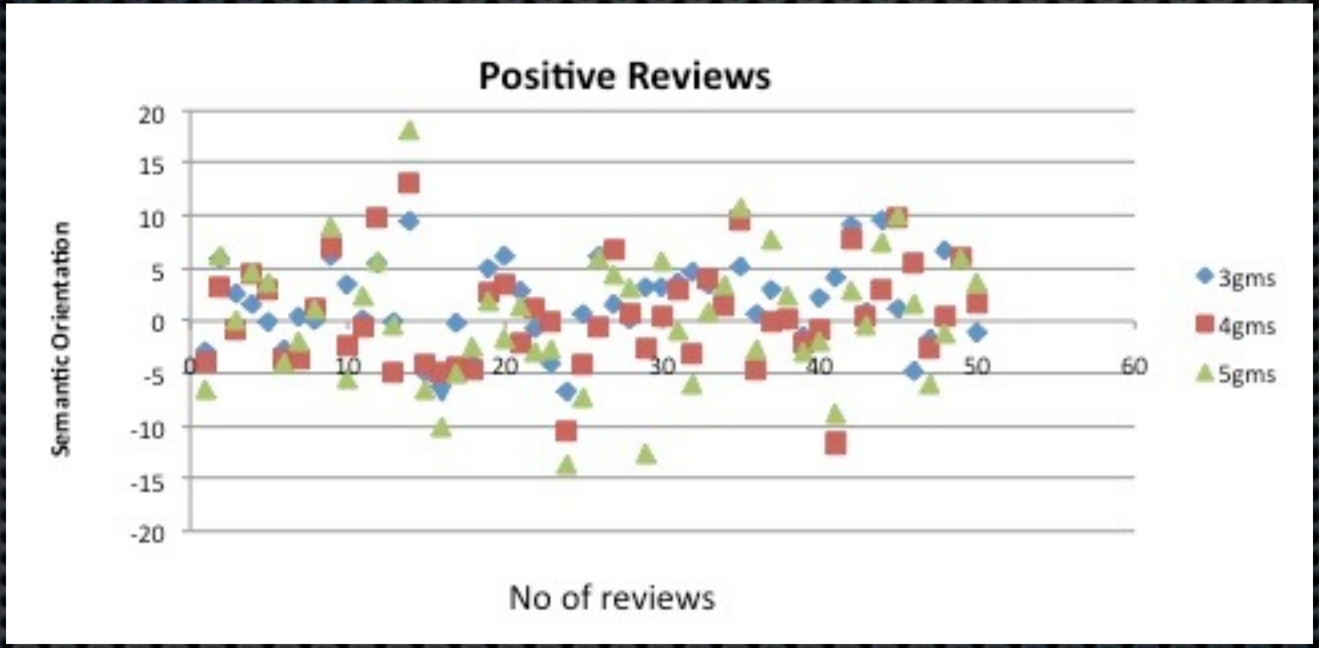
4.RESULTS AND DISCUSSION

Constraints in recreation & Improvement

- Constraints
  - Don’t have same input as previous experiment
  - Less Identified phrases
  - Small sample size
  - Movie review classification is hard
- Improvement
  - Change the association word
  - Less PMI processing time
  - Improve sample size

4.RESULTS AND DISCUSSION

Effect of using different word window sizes



- Surprisingly smaller n-grams cut upper and lower noise
- Pang & Lee (2002) also mentioned that unigram can provide better approximation

AGENDA

- Introduction
- Previous Work
- Objectives
- Methodology
- Result and Discussion
- Conclusion

CONCLUSION

- We have developed an opinion mining system, which replicate Turney (2002) method
- Evaluated the System system accuracy
- Evaluated the effect of approximation using Google Web1T
- Evaluated the effect of using different n-grams

Question?