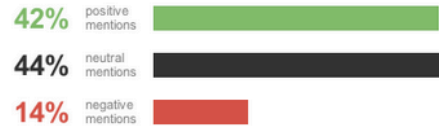


SENTIMENT METER for TRADER JOE'S

28 AUGUST : 10 SEPTEMBER



SENTIMENT METER for TRADER JOE'S

11 SEPTEMBER : 12 SEPTEMBER



Sentiment Analysis of Social Media Texts

Saif M. Mohammad and Xiaodan Zhu

{saif.mohammad,xiaodan.zhu}@nrc-cnrc.gc.ca

National Research Council Canada

Acknowledgment: Thanks to Svetlana Kiritchenko for helpful suggestions.

A tutorial presented at the 2014 Conference on Empirical Methods on Natural Language Processing, October 2014, Doha, Qatar.



INTRODUCTION

Sentiment Analysis

- Is a given piece of text **positive, negative, or neutral**?
 - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Sentiment Analysis

- Is a given piece of text **positive, negative, or neutral**?
 - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Emotion Analysis

- What emotion is being expressed in a given piece of text?
 - Basic emotions: **joy, sadness, fear, anger,...**
 - Other emotions: **guilt, pride, optimism, frustration,...**

Sentiment Analysis

- Is a given piece of text **positive, negative, or neutral**?
 - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Emotion Analysis } **Not in the scope of this tutorial.**

- What emotion is being expressed in a given piece of text?
 - Basic emotions: joy, sadness, fear, anger,...
 - Other emotions: guilt, pride, optimism, frustration,...

Sentiment Analysis: Tasks

- Is a given piece of text **positive, negative, or neutral**?
 - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.

Sentiment Analysis: Tasks

- Is a given piece of text **positive, negative, or neutral**?
 - The text may be a sentence, a tweet, an SMS message, a customer review, a document, and so on.
- Is a word within a sentence positive, negative, or neutral?
 - unpredictable movie plot vs. unpredictable steering
- What is the sentiment towards specific aspects of a product?
 - sentiment towards the food and sentiment towards the service in a customer review of a restaurant
- What is the sentiment towards an entity such as a politician, government policy, company, or product?
 - Stance detection: favorable or unfavorable
 - Framing: focusing on specific dimensions

Sentiment Analysis: Tasks (continued)

- What is the sentiment of the speaker/writer?
 - Is the speaker explicitly expressing sentiment?
- What sentiment is evoked in the listener/reader?
- What is the sentiment of an entity mentioned in the text?

Consider the above questions with the examples below:

General Tapioca was ruthlessly executed today.

Mass-murdered General Tapioca finally killed in battle.

General Tapioca was killed in an explosion.

May God help the people being persecuted by General Tapioca.

Sentiment Analysis: Domains

- Newspaper texts
 - Financial news
 - Entertainment news
- Legal texts
- Novels
- E-mails
- SMS messages
- Customer reviews
- Blog posts
- Tweets
- Facebook posts
- ...and so on.

Sentiment Analysis: Domains

- Newspaper texts
 - Financial news
 - Entertainment news
- Legal texts
- Novels
- E-mails
- SMS messages
- Customer reviews
- Blog posts
- Tweets
- Facebook posts
- ...and so on.

Short informal pieces of text – often called Social media texts.

Quirks of Social Media Texts

- Informal
- Short
 - 140 characters for tweets and SMS messages
- Abbreviations and shortenings
- Wide array of topics and large vocabulary
- Spelling mistakes and creative spellings
- Special strings
 - hashtags, emoticons, conjoined words
- High volume
 - 500 million tweets posted every day
- Often come with meta-information
 - date, links, likes, location
- Often express **sentiment**

Example Applications of Sentiment Analysis and Emotion Detection

- Tracking sentiment towards politicians, movies, products
- Improving customer relation models
- Identifying what evokes strong emotions in people
- Detecting happiness and well-being
- Measuring the impact of activist movements through text generated in social media.
- Improving automatic dialogue systems
- Improving automatic tutoring systems
- Detecting how people use emotion-bearing-words and metaphors to persuade and coerce others

Not in the Scope of this Tutorial

- Sentiment in formal writing such as news, academic publications, etc.
- Application-specific analysis
 - for example, for predicting stocks, election results, public health, machine translation, etc.
- Sentiment analysis in resource-poor languages
 - porting sentiment resources from one language to another
- Detecting sentiment of reader
- Stance detection
- Visualizations of sentiment
- Emotion analysis



SemEval Sentiment Tasks

- SemEval-2013 Task 2
- SemEval-2014 Task 9
- SemEval-2014 Task 4

SemEval-2013, Task 2: Sentiment Analysis in Twitter

- Tasks
 - Is a given **message** positive, negative, or neutral?
 - Is a given **term within a message** positive, negative, or neutral?
- Data
 - test set
 - tweets
 - SMS messages
 - training set
 - tweets
 - sources of data
 - tweets taken from Twitter API
 - search was for certain named entities
 - tweets had to have some words from SentiWordNet

Examples: Message-Level Sentiment

Tweet: The new Star Trek movie is visually spectacular.

Tweet: The new Star Trek movie does not have much of a story.

Tweet: Spock is from planet Vulcan.

Examples: Message-Level Sentiment

Tweet: The new Star Trek movie is visually spectacular.
positive

Tweet: The new Star Trek movie does not have much of a story.
negative

Tweet: Spock is from planet Vulcan.
neutral

Examples: Message-Level Sentiment

Tweet: The new Star Trek movie is visually spectacular.
positive

Tweet: The new Star Trek movie does not have much of a story.
negative

Tweet: Spock is from planet Vulcan.
neutral

When creating annotated data:

- using labels **indeterminate** and **both positive and negative** may be helpful.

Examples: Term-Level Sentiment

Tweet: The new Star Trek does not have much of a story, but it is visually spectacular.
target

Tweet: The movie was so slow it felt like a documentary.
target

Tweet: Spock is watching a documentary.
target

Examples: Term-Level Sentiment

Tweet: The new Star Trek does not have much of a story, but it is visually spectacular.
target is positive

Tweet: The movie was so slow it felt like a documentary.
target is negative

Tweet: Spock is watching a documentary.
target is neutral

Evaluation Metric

- Macro-averaged F-score:

$$F = (F_{pos} + F_{neg})/2$$

where F_{pos} and F_{neg} are the f-scores of the positive and negative sentiment classes, respectively; i.e.,

$$F_{pos} = 2 \frac{P_{pos} R_{pos}}{P_{pos} + R_{pos}}$$

So, the two classes are given the same weight in evaluation.

SemEval-2014, Task 9: Sentiment Analysis in Twitter

(repeat of 2013 task)

- Tasks
 - Is a given **message** positive, negative, or neutral?
 - Is a given **term within a message** positive, negative, or neutral?
- Data
 - test set
 - 2014: tweets set, sarcastic tweets set, blog posts set
 - 2013: tweets set, SMS set
 - training set
 - 2013 tweets set
 - sources of data
 - tweets taken from Twitter API
 - search was for certain named entities
 - tweet had to have some words from SentiWordNet

SemEval-2014, Task 4: Aspect Based Sentiment Analysis

- Tasks
 - In a restaurant or laptop review, identify:
 - aspect terms
 - aspect categories
 - sentiment towards aspect terms
 - sentiment towards aspect categories

The lasagna was great, but we had to wait 20 minutes to be seated.

aspect terms: lasagna (positive)

aspect categories: food (positive), service (negative)

Restaurant domain had five pre-defined aspect categories:

- food, service, price, ambience, anecdotes/miscellaneous

SemEval-2014, Task 4: Aspect Based Sentiment Analysis

- Data
 - test set
 - restaurant reviews
 - laptop reviews
 - training set
 - restaurant reviews
 - laptop reviews
 - source of data
 - restaurant data: customer reviews from Citysearch New York
 - laptop data: customer reviews from Amazon.com

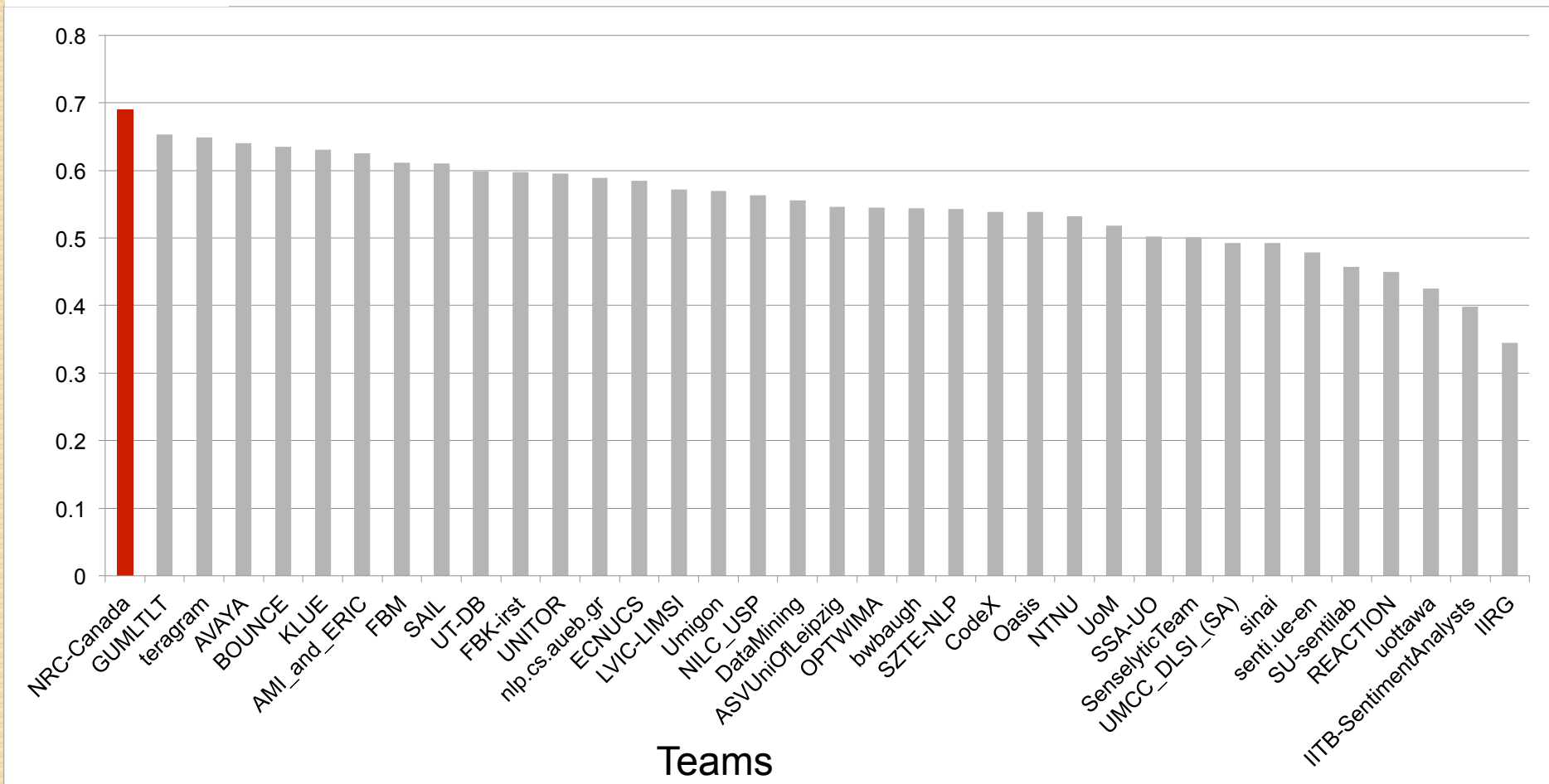
NRC-Canada ranks in SemEval-2013, Task 2

- message-level task (44 teams)
 - tweets set: 1st
 - SMS set: 1st
- term-level task (23 teams)
 - tweets set: 1st
 - SMS set: 2nd

Sentiment Analysis Competition

Classify Tweets

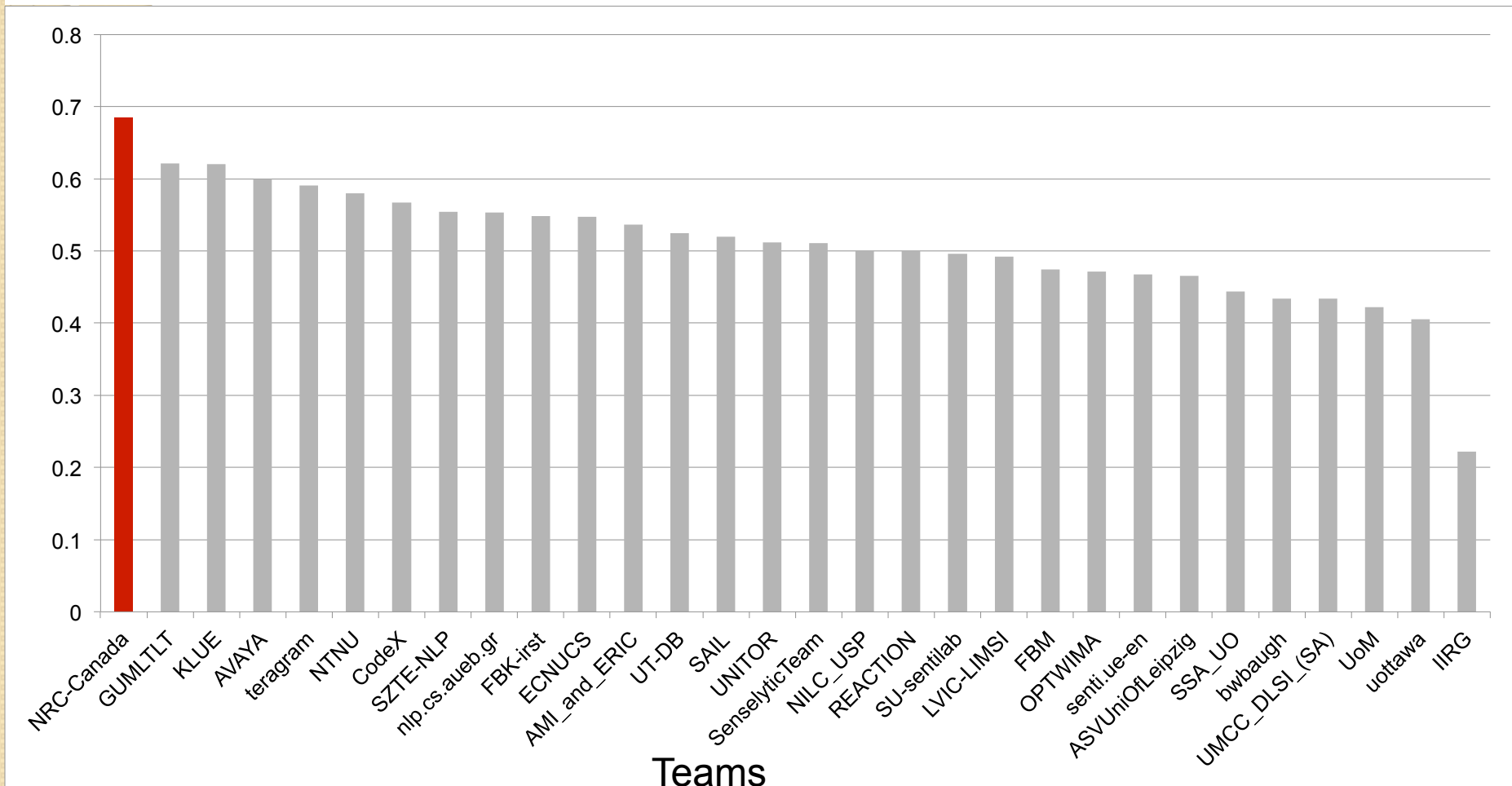
F-score



Sentiment Analysis Competition

Classify SMS

F-score



NRC-Canada ranks in SemEval-2013, Task 2

- message-level task (44 teams)
 - tweets set: 1st
 - SMS set: 1st
- term-level task (23 teams)
 - tweets set: 1st
 - SMS set: 2nd

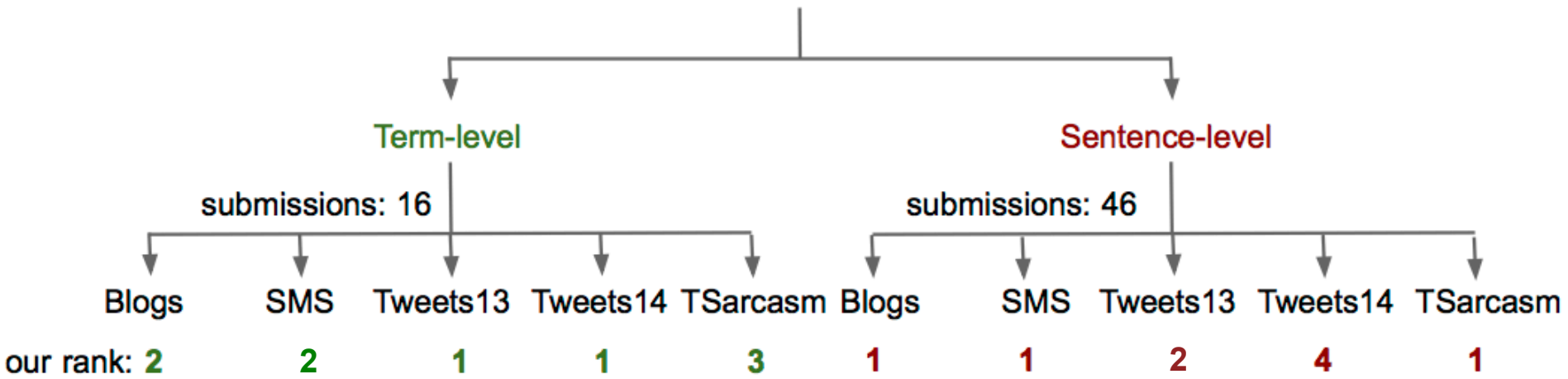
Released description of features.

Released resources created (tweet-specific sentiment lexicons).

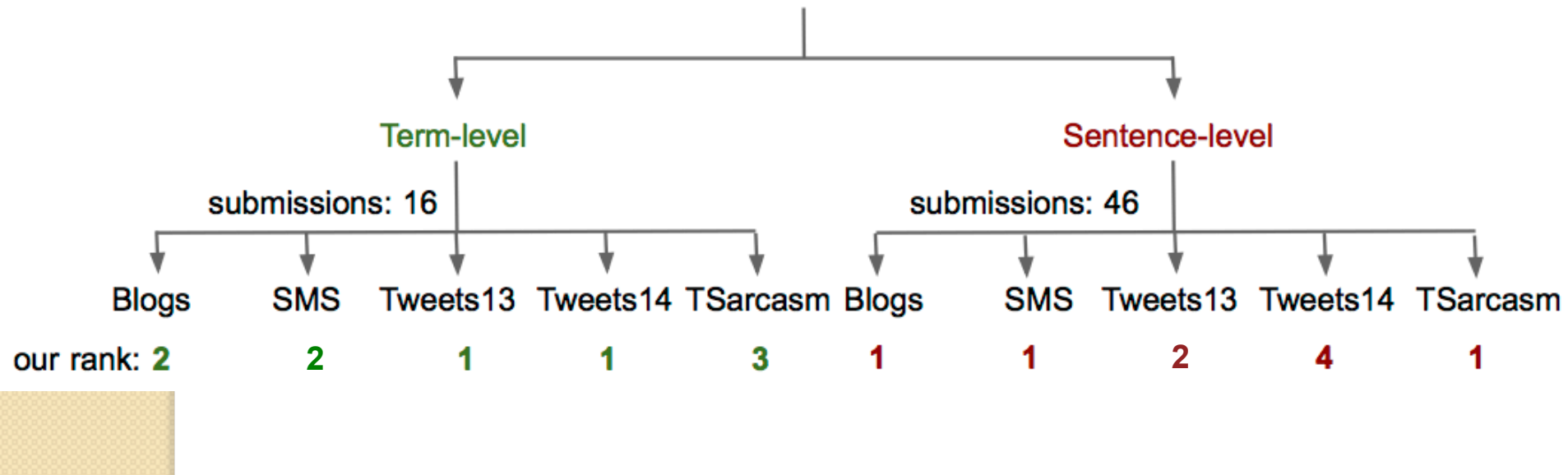
www.purl.com/net/sentimentoftweets

NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu, In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), June 2013, Atlanta, USA.

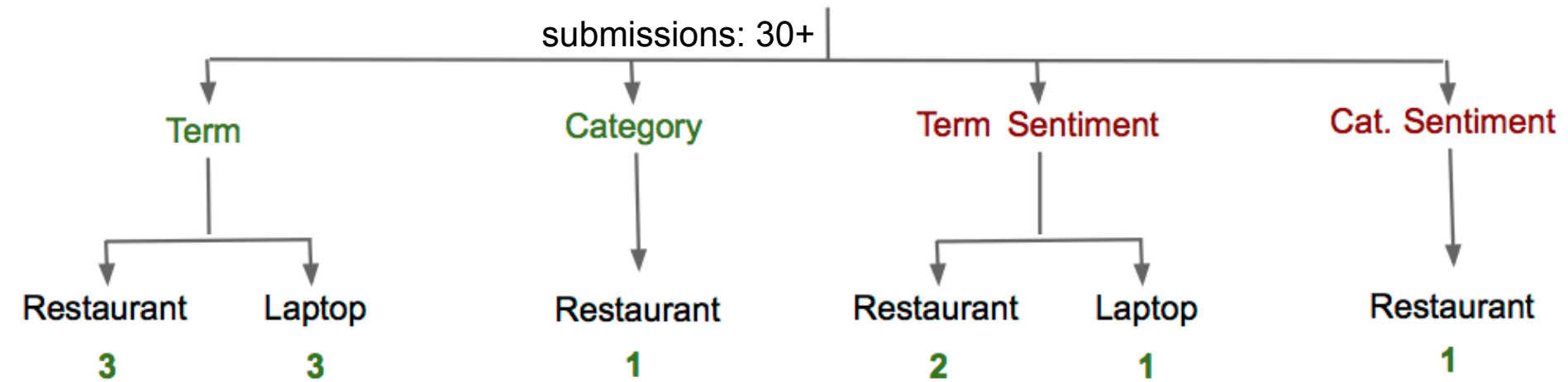
Sentiment Analysis of Social Media Texts (SemEval-2014 Task 9)



Sentiment Analysis of Social Media Texts (SemEval-2014 Task 9)



Aspect-Based Sentiment Analysis (SemEval-2014 Task 4)



SemEval-2015, Sentiment Tasks

- Task 12: Aspect Based Sentiment Analysis
 - repeat of 2014 task
 - new subtask on domain adaptation
- Task 11: Sentiment Analysis of Figurative Language in Twitter
 - metaphoric and ironic tweets
 - intensity of sentiment
- Task 10: Sentiment Analysis in Twitter
 - repeat of 2013 and 2014 task
 - more subtasks
- Task 9: CLIPEval Implicit Polarity of Events
 - Identify polarity and event class



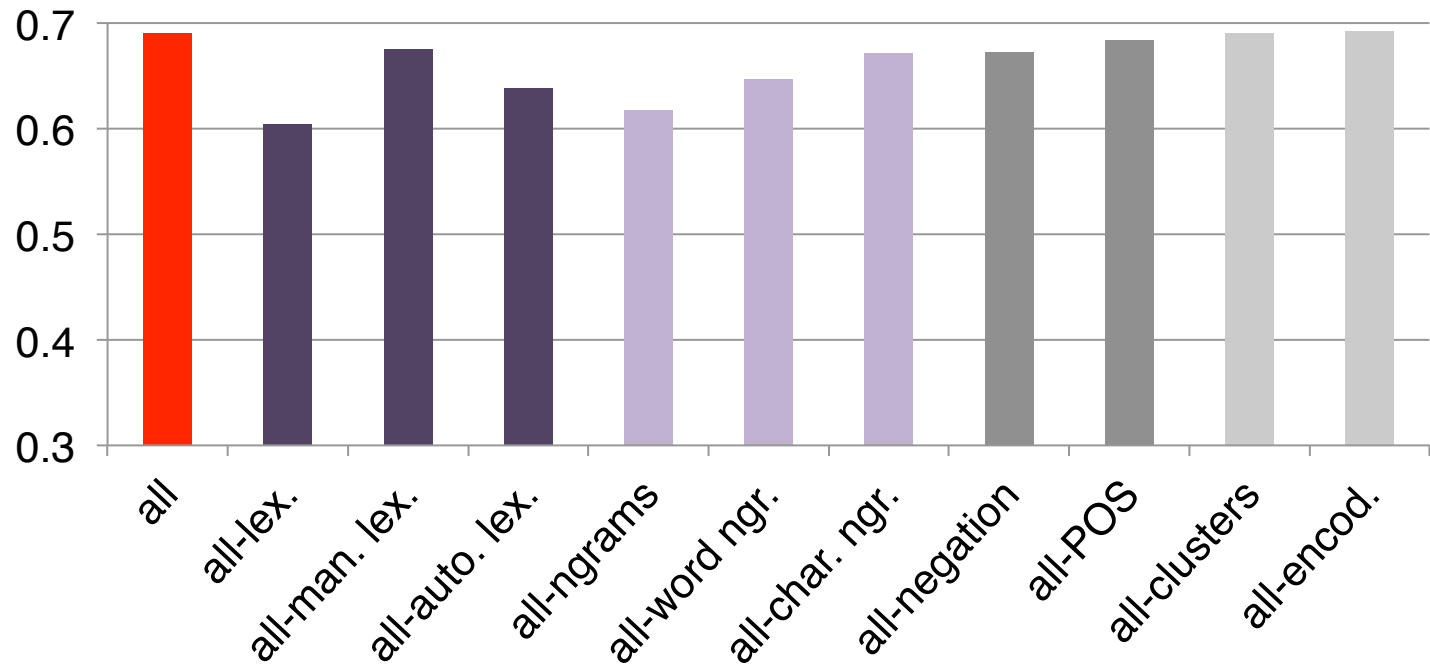
Sentiment Analysis Features

Sentiment analysis features

Features	Examples
sentiment lexicon	#positive: 3, scorePositive: 2.2; maxPositive: 1.3; last: 0.6, scoreNegative: 0.8, scorePositive_neg: 0.4
word n-grams	spectacular, like documentary
char n-grams	spect, docu, visua
part of speech	#N: 5, #V: 2, #A:1
negation	#Neg: 1; ngram:perfect → ngram:perfect_neg, polarity:positive → polarity:positive_neg
word clusters	probably, definitely, def
all-caps	YES, COOL
punctuation	#!+: 1, #?+: 0, #!?: 0
emoticons	:D, >:(
elongated words	soooo, yaayyy

Feature Contributions (on Tweets)

F-scores



N-grams

- Word ngrams
 - unigrams, bigrams, trigrams, fourgrams
 - skip ngrams
 - $w_1 * w_3$
 - Features:
 - whether ngram present or not
- Character ngrams
 - 3-gram, 4-gram, 5-gram
 - Features:
 - whether ngram present or not

Sentiment Lexicons

Lists of positive and negative words.

Positive

spectacular

okay

Negative

lousy

unpredictable

Sentiment Lexicons: **Manually** Created

- General Inquirer (Stone, Dunphy, Smith, Ogilvie, & associates, 1966): ~3,600 words
- MPQA (Wilson, Wiebe, & Hoffmann, 2005): ~8,000 words
- Hu and Liu Lexicon (Hu and Liu, 2004): ~6,800 words
- NRC Emotion Lexicon (Mohammad & Turney, 2010): ~14,000 words and ~25,000 word senses
 - senses are based on categories in a thesaurus
 - has emotion associations in addition to sentiment
- AFINN (by Finn Årup Nielsen in 2009-2011): ~2400 words
- MaxDiff Sentiment Lexicon (Kiritchenko, Zhu, and Mohammad, 2014): about 1,500 terms
 - has intensity scores

Sentiment Lexicons

Lists of positive and negative words.

Positive

spectacular

okay

Negative

lousy

unpredictable

Sentiment Lexicons

Lists of positive and negative words, with scores indicating the degree of association

Positive

spectacular 0.91

okay 0.30

Negative

lousy -0.84

unpredictable -0.17

Sentiment Lexicons

Lists of positive and negative words, with scores indicating the degree of association

Positive

spectacular 0.91

okay 0.30

Negative

lousy -0.84

unpredictable -0.17

spectacular positive 0.91

okay positive 0.30

lousy negative 0.84

unpredictable negative 0.17

How to create sentiment lexicons with intensity values?

- Humans are not good at giving real-valued scores?
 - hard to be consistent across multiple annotations
 - difficult to maintain consistency across annotators
 - 0.8 for annotator may be 0.7 for another
- Humans are much better at comparisons
 - Questions such as: Is one word more positive than another?
 - Large number of annotations needed.

Need a method that preserves the comparison aspect, without greatly increasing the number of annotations needed.

MaxDiff

- The annotator is presented with four words (say, A, B, C, and D) and asked:
 - which word is the **most** positive
 - which is the **least** positive
- By answering just these two questions, five out of the six inequalities are known
 - For e.g.:
 - If A is most positive
 - and D is least positive, then we know:
 $A > B, A > C, A > D, B > D, C > D$

Maximum difference scaling: improved measures of importance and preference for segmentation. Cohen, Steve. *Sawtooth Software Conference Proceedings, Sawtooth Software, Inc.* Vol. 530. 2003.

MaxDiff

- Each of these MaxDiff questions can be presented to multiple annotators.
 - The responses to the MaxDiff questions can then be easily translated into:
 - a ranking of all the terms
 - a real-valued score for all the terms (Orme, 2009)
 - If two words have very different degrees of association (for example, $A \gg D$), then:
 - A will be chosen as most positive much more often than D
 - D will be chosen as least positive much more often than A.
- This will eventually lead to a ranked list such that A and D are significantly farther apart, and their real-valued association scores will also be significantly different.

Dataset of Sentiment Scores

(Kiritchenko, Zhu, and Mohammad, 2014)

- Selected ~1,500 terms from tweets
 - regular English words: peace, jumpy
 - tweet-specific terms
 - hashtags and conjoined words: #inspiring, #happytweet, #needsleep
 - creative spellings: amazzing, goooood
 - negated terms: not nice, nothing better, not sad
- Generated 3,000 MaxDiff questions
- Each question annotated by 10 annotators on CrowdFlower
- Answers converted to real-valued scores (0 to 1) and to a full ranking of terms using the counting procedure (Orme, 2009)

Examples of sentiment scores from the MaxDiff annotations

Term	Sentiment Score 0 (most negative) to 1 (most positive)
awesomeness	0.9133
#happygirl	0.8125
cant waitttt	0.8000
don't worry	0.5750
not true	0.3871
cold	0.2750
#getagrip	0.2063
#sickening	0.1389

Robustness of the Annotations

- Divided the MaxDiff responses into two equal halves
- Generated scores and ranking based on each set individually
- The two sets produced very similar results:
 - average difference in scores was 0.04
 - Spearman's Rank Coefficient between the two rankings was 0.97

Dataset will be used as test set for Subtask E in Task 10 of SemEval-2015: [Determining prior probability](#).

Trial data already available:

<http://alt.qcri.org/semEval2015/task10/index.php?id=data-and-tools>

(Full dataset to be released after the shared task competition in Dec., 2014.)

Sentiment Lexicons: **Automatically** Created

- Turney and Littman Lexicon (Turney and Littman, 2003)
- SentiWordNet (Esuli & Sebastiani, 2006): WordNet synsets
- MSOL (Mohammad, Dunne, and Dorr, 2009): ~60,000 words
- Hashtag Sentiment Lexicon (Mohammad, Kiritchenko, and Zhu, 2013): ~220,000 unigrams and bigrams
- Sentiment140 Sentiment Lexicon (Mohammad, Kiritchenko, and Zhu, 2013): ~330,000 unigrams and bigrams

Turney and Littman (2003) Method

- Pointwise Mutual Information (PMI) based measure
- PMI between two words, w_1 and w_2 (Church and Hanks 1989):

$$PMI(w_1, w_2) = \log_2(p(w_1 \text{ and } w_2) / p(w_1)p(w_2))$$

$p(w_1 \text{ and } w_2)$ is probability of how often w_1 and w_2 co-occur

$p(w_1)$ is probability of occurrence of w_1

$p(w_2)$ is probability of occurrence of w_2

If $PMI > 1$, then w_1 and w_2 co-occur more often than chance

If $PMI < 1$, then w_1 and w_2 co-occur less often than chance

Turney and Littman (2003) Method (continued)

- Created a list of **seed** sentiment words:
 - positive seeds (Pwords): good, nice, excellent, positive, fortunate, correct, superior
 - negative seeds (Nwords): bad, nasty, poor, negative, unfortunate, wrong, inferior
- Polled the AltaVista Advanced Search Engine for number of documents that had both a target word and a seed word within a small window
 - positive seed is assumed to be positive label for co-occurring word w
 - negative seed is assumed to be negative label for co-occurring word w

Turney and Littman (2003) Method (continued)

- For every word w a sentiment association score is generated:

$$score(w) = PMI(w, positive) - PMI(w, negative)$$

PMI = pointwise mutual information

$$PMI(w, positive) = \sum_{pword \in Pwords} PMI(w, Pword)$$

If $score(w) > 0$, then word w is positive

If $score(w) < 0$, then word w is negative

Hashtagged Tweets

- Hashtagged words are good labels of sentiments and emotions

Can't wait to have my own Google glasses **#awesome**

Some jerk just stole my photo on **#tumblr**. **#grr #anger**

- Hashtags are not always good labels:
 - hashtag used sarcastically
 - hashtagged emotion not in the rest of the message

Mika used my photo on tumblr. **#anger**

#Emotional Tweets, Saif Mohammad, In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), June 2012, Montreal, Canada.

Mohammad, Kiritchenko, and Zhu Method

- Created a list of **seed** sentiment words by looking up synonyms of *excellent*, *good*, *bad*, and *terrible*:
 - 30 positive words
 - 46 negative words
- Polled the Twitter API for tweets with seed-word hashtags
 - A set of 775,000 tweets was compiled from April to December 2012

Automatically Generated New Lexicons

- Sentiment lexicons can be generated from sentiment-labeled data
 - Emoticons and hashtag words can be used as labels
- For every word w in the set of millions tweets, an association score is generated:

$$score(w) = PMI(w,positive) - PMI(w,negative)$$

PMI = pointwise mutual information

If $score(w) > 0$, then word w is positive

If $score(w) < 0$, then word w is negative

PMI Method based Lexicons

- Hashtag Sentiment Lexicon
 - created from a large collection of hashtagged tweets
 - has entries for ~215,000 unigrams and bigrams
- Sentiment140 Lexicon
 - created from a large collection of tweets with emoticons
 - Sentiment140 corpus (Alec Go, Richa Bhayani, and Lei Huang, 2009)
<http://help.sentiment140.com/for-students/>
 - has entries for ~330,000 unigrams and bigrams

Features of the Twitter Lexicon

- connotation and not necessarily denotation
 - tears, party, vacation
- large vocabulary
 - cover wide variety of topics
 - lots of informal words
 - twitter-specific words
 - creative spellings, hashtags, conjoined words
- seed hashtags have varying effectiveness
 - study on sentiment predictability of different hashtags
(Kunneman, F.A., Liebrecht, C.C., van den Bosch, A.P.J., 2014)

Negation

- A grammatical category that allows the changing of the truth value of a proposition (Morante and Sporleder, 2012)
- Often expressed through the use of negative signals or negators
 - words like *isn't* and *never*
- Can significantly affect the sentiment of its **scope**
- Examples:

People do **not** like change.

Jack **never** hated the plan, he just has other priorities.

The negator is shown in blue.

The scope is shown by underline.

Conventional methods to handle negation

Reversing hypothesis:

$$s(n,w) = -s(w)$$

where, $s(w)$ is the sentiment of word (or phrase) w ,

$s(n,w)$ is the sentiment of the expression formed by the concatenation of the negator n and word w .

- For example, if $s(\text{honest}) = 0.9$, then $s(\text{not, honest}) = -0.9$

But how good is this hypothesis?

What about:

The movie is not terrible.

If $s(\text{terrible}) = -0.9$. what is the appropriate value for $s(\text{not, terrible})$ in this context?

Negation

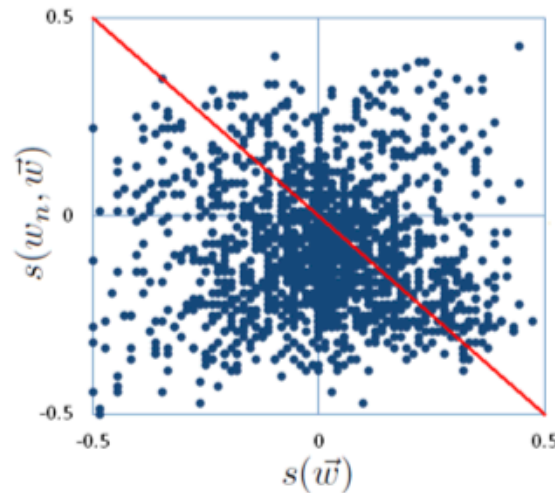


Figure 1: Effect of a list of common negators in modifying sentiment values in Stanford Sentiment Treebank. The x-axis is $s(\vec{w})$, and y-axis is $s(w_n, \vec{w})$. Each dot in the figure corresponds to a text span being modified by (composed with) a negator in the treebank. The red diagonal line corresponds to the sentiment-reversing hypothesis that simply reverses the sign of sentiment values.

An Empirical Study on the Effect of Negation Words on Sentiment. Xiaodan Zhu, Hongyu Guo, Saif Mohammad and Svetlana Kiritchenko. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014, Baltimore, MD.

Negation

Jack was not thrilled at the prospect of working weekends ☹️

The bill is not garbage, but we need a more focused effort ☹️

Negation

Jack was not thrilled at the prospect of working weekends ☹️

↓
negator



need to determine this word's
sentiment when negated

↓
sentiment
label: negative

The bill is not garbage, but we need a more focused effort ☹️

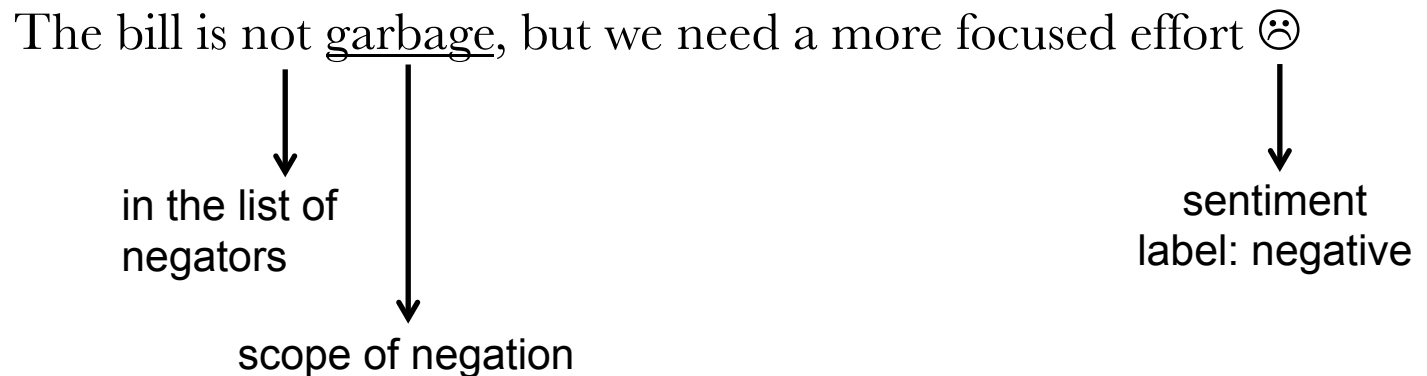
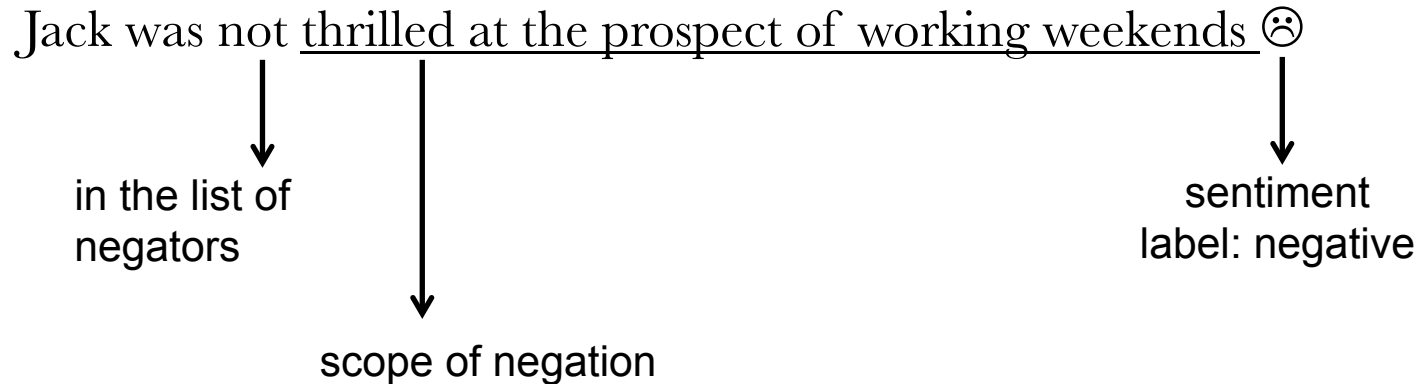
↓
negator



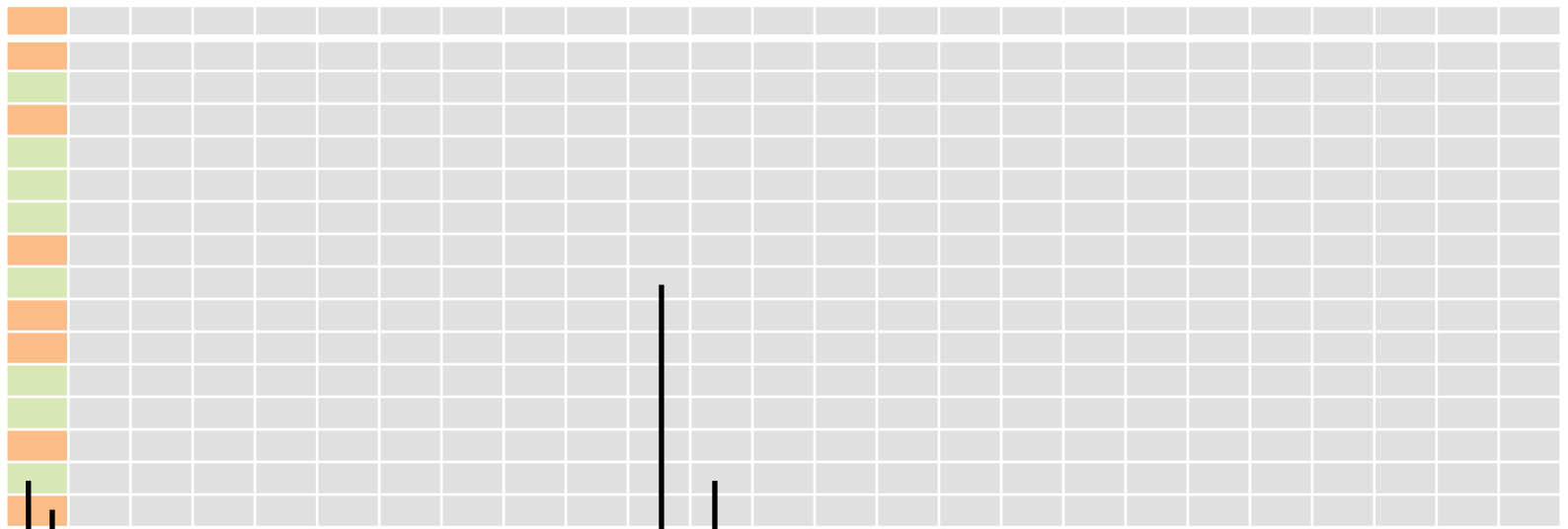
need to determine this word's
sentiment when negated

↓
sentiment
label: negative

Handling Negation



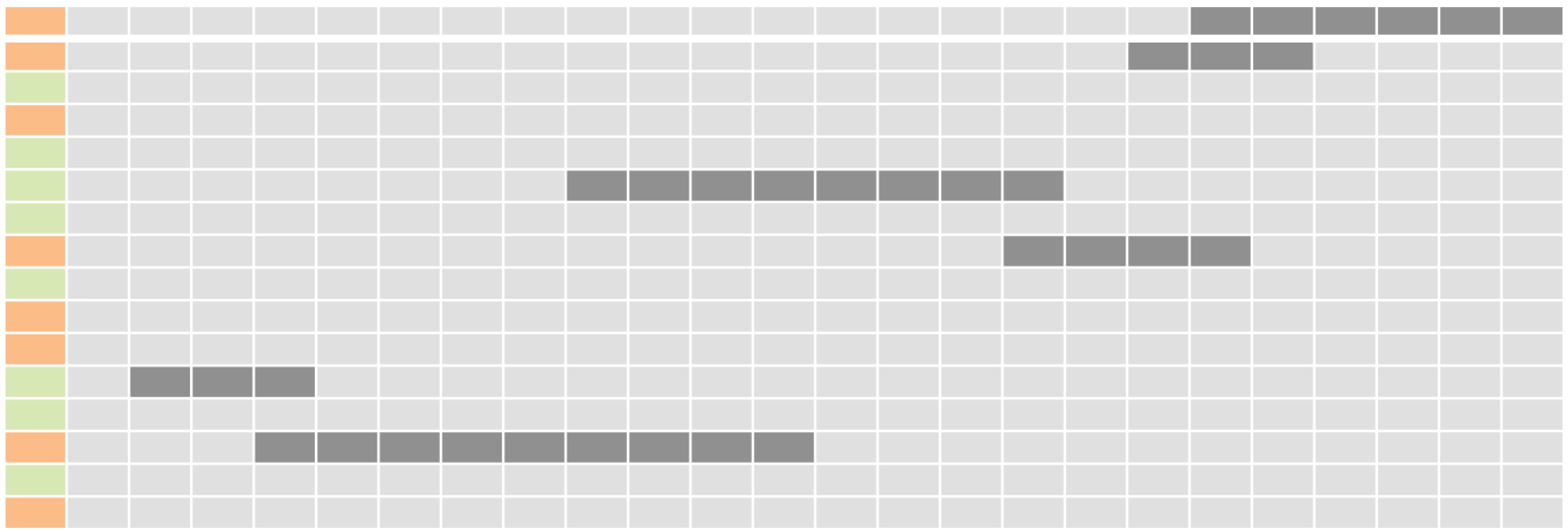
Scope of negation: from negator till a punctuation (or end of sentence)



negative label

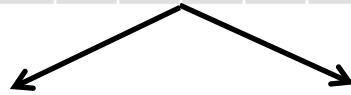
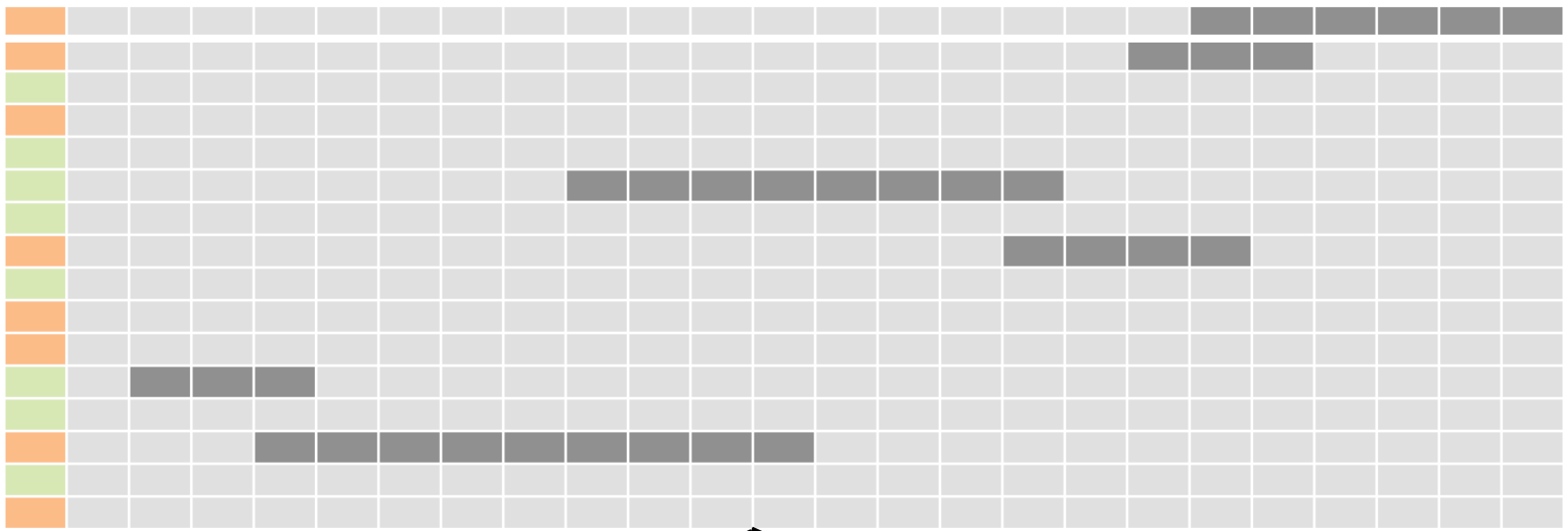
positive label

tweets or sentences

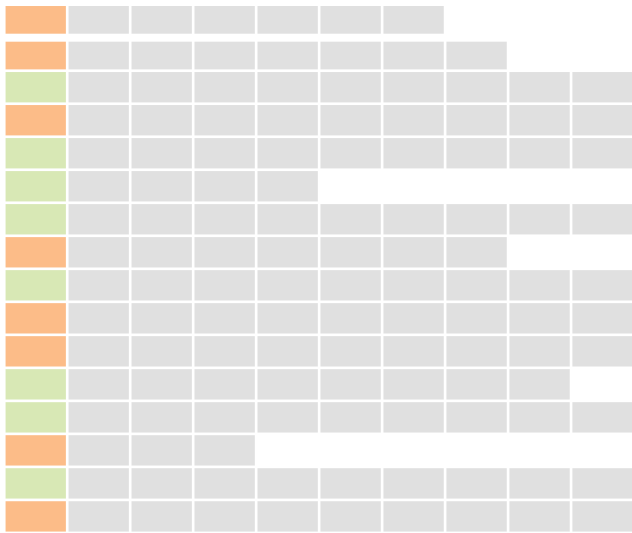


affirmative contexts
(in light grey)

negated contexts
(in dark grey)



All the affirmative contexts



Generate sentiment lexicon for words in affirmative context

All the negated contexts



Generate sentiment lexicon for words in negated context

Term	Sentiment140 Lexicons		
	Base	AffLex	NegLex
Positive terms			
great	1.177	1.273	-0.367
nice	0.974	1.149	-0.912
honest	0.391	0.431	-0.123
Negative terms			
terrible	-1.766	-1.850	-0.890
bad	-1.297	-1.674	0.021
negative	-0.090	-0.261	0.389

Table 3: Example sentiment scores from the Sentiment140 Base, Affirmative Context (AffLex) and Negated Context (NegLex) Lexicons.

The number of positive and negative entries in the sentiment lexicons.

Lexicon	Positive	Negative	Total
NRC Emotion Lexicon	2,312 (41%)	3,324 (59%)	5,636
Bing Liu's Lexicon	2,006 (30%)	4,783 (70%)	6,789
MPQA Subjectivity Lexicon	2,718 (36%)	4,911 (64%)	7,629
Hashtag Sentiment Lexicons (HS)			
HS Base Lexicon			
- unigrams	19,121 (49%)	20,292 (51%)	39,413
- bigrams	69,337 (39%)	109,514 (61%)	178,851
HS AffLex			
- unigrams	19,344 (51%)	18,905 (49%)	38,249
- bigrams	67,070 (42%)	90,788 (58%)	157,858
HS NegLex			
- unigrams	936 (14%)	5,536 (86%)	6,472
- bigrams	3,954 (15%)	22,258 (85%)	26,212
Sentiment140 Lexicons (S140)			
S140 Base Lexicon			
- unigrams	39,979 (61%)	25,382 (39%)	65,361
- bigrams	135,280 (51%)	131,230 (49%)	266,510
S140 AffLex			
- unigrams	40,422 (63%)	23,382 (37%)	63,804
- bigrams	133,242 (55%)	107,206 (45%)	240,448
S140 NegLex			
- unigrams	1,038 (12%)	7,315 (88%)	8,353
- bigrams	5,913 (16%)	32,128 (84%)	38,041

More Data: Restaurant Reviews

- **Yelp Phoenix Academic Dataset**
 - 230,000 customer reviews posted on Yelp
 - 500 business categories
 - multiple categories assigned for a business
 - For e.g., “restaurant, deli, and bakery”
- **Yelp Restaurant Reviews corpus**
 - 58 business categories related to the restaurant domain
 - 183,935 customer reviews
 - Generated sentiment lexicons using the star ratings as labels

More Data: Laptop Reviews

- Amazon Customer Reviews Dataset (McAuley and Leskovec, 2013)
 - 34,686,770 customer reviews posted on Amazon from 1995 to 2013 (11GB of data)
 - 6,643,669 users
 - 2,441,053 products
- Amazon Laptop Reviews corpus
 - Searched for mentions of laptop or notepad in the electronics reviews subset
 - 124,712 customer reviews
 - Generated sentiment lexicons using the star ratings as labels

Other Recent Approaches to Creating Sentiment Lexicons

- Using neural networks and deep learning techniques
 - Duyu Tang, Furu Wei, Bing Qin, Ming Zhou and Ting Liu (2014)
- Constructing domain-specific sentiment
 - Sheng Huang, Zhendong Niu, and Chongyang Shi (2014)
 - Ilia Chetviorkin and Natalia Loukachevitch (2014)
- Others:
 - Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani (2014): SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter.
 - Shi Feng, Kaisong Song, Daling Wang, Ge Yu (2014): A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs.
 - Raheleh Makki, Stephen Brooks and Evangelos E. Milios (2014): Context-Specific Sentiment Lexicon Expansion via Minimal User Interaction.
 - Yanqing Chen and Steven Skiena (2014): Building Sentiment Lexicons for All Major Languages
 - Bandhakavi et al. (2014): Generating a Word-Emotion Lexicon from #Emotional Tweets -- EM with Mixture of Classes Model.

Sentiment analysis features

Features	Examples
sentiment lexicon	#positive: 3, scorePositive: 2.2; maxPositive: 1.3; last: 0.6, scoreNegative: 0.8, scorePositive_neg: 0.4
word n-grams	spectacular, like documentary
char n-grams	spect, docu, visua
part of speech	#N: 5, #V: 2, #A:1
negation	#Neg: 1; ngram:perfect → ngram:perfect_neg, polarity:positive → polarity:positive_neg
word clusters	probably, definitely, def
all-caps	YES, COOL
punctuation	#!+: 1, #?+: 0, #!?: 0
emoticons	:D, >:(
elongated words	soooo, yaayyy

Sentiment Analysis of Short Informal Texts. Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad. *Journal of Artificial Intelligence Research*, 50, August 2014.

Word Clusters

- The CMU Twitter NLP tool provides 1000 token clusters
 - produced with the Brown clustering algorithm on 56 million English-language tweets
 - alternative representation of tweet content
 - reducing the sparsity of the token space
- Feature:
 - the presence or absence of tokens from each of the 1000 clusters.

Other Features

- punctuation:
 - the number of contiguous sequences of exclamation marks, question marks, and both exclamation and question marks, for example, **!!!!**
 - whether the last token contains an exclamation or question mark
- emoticons
 - presence or absence of emoticons at any position in the tweet, for example, **:)**
 - whether the last token is a positive or negative emoticon
- elongated words
 - the number of words with one character repeated more than two times, for example, **yesssss**

Sentiment analysis features

Features	Examples
sentiment lexicon	#positive: 3, scorePositive: 2.2; maxPositive: 1.3; last: 0.6, scoreNegative: 0.8, scorePositive_neg: 0.4
word n-grams	spectacular, like documentary
char n-grams	spect, docu, visua
part of speech	#N: 5, #V: 2, #A:1
negation	#Neg: 1; ngram:perfect → ngram:perfect_neg, polarity:positive → polarity:positive_neg
word clusters	probably, definitely, def
all-caps	YES, COOL
punctuation	#!+: 1, #?+: 0, #!?: 0
emoticons	:D, >:(
elongated words	soooo, yaayyy

Sentiment Analysis of Short Informal Texts. Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad. *Journal of Artificial Intelligence Research*, 50, August 2014.

Overview of Sentiment Analysis Systems

- Rule-based systems
- Conventional statistical systems
- Deep-learning-based models

Overview of Sentiment Analysis Systems

- Rule-based systems
- Conventional statistical systems
- Deep-learning-based models

Teragram: A Rule-Based System (Reckman et al., 2013)

- Develop lexicalized hand-written rules: each rule is a pattern that matches words or sequences of words.

- Examples:

Negative: `_def{Negation} _def{PositiveAdjectives}`
`(SENT, (DIST_4, “_a{ _def{HigherIsBetter}”,`
`“_a{ _def{Lowering} }”))`

Positive: `(ORDDIST_7, “_def{PositiveContext}”,`
`“_a{ _def{PositiveAmbig} }”)`

- Background data: use blogs, forums, news, and tweets to develop the rules.
- Performance:
 - Ranked 3rd on the tweet test data in message-level task (SemEval-2013 Task 2), but ranked 15th on the term-level task.

Remarks

- Carefully developed rule-based systems can sometimes achieve complete performance on the data/domains they are created for.
- Advantages: explicit knowledge representation, so intuitive to develop and maintain.
- Problems
 - Coverage: hand-written rules often have limited coverage, so recall is often low. This can impact the overall performance (as observed in Teragram).
 - Extensibility: not easy to be extended to new data/domains; rule-based models have inherent difficulty in automatically acquiring knowledge.
 - Modeling capability.

Remarks (continued)

- The main stream is statistical approaches, which achieve top performance across different tasks and data sets.
 - Note that knowledge acquired by applying rules can often be easily incorporated as features into statistical approaches.

Overview of Sentiment Analysis Systems

- Rule-based systems
- Conventional statistical systems
 - NRC-Canada systems
 - Key ideas from other top teams
- Deep-learning-based models

Detailed Description of the NRC-Canada Systems

- Message-level sentiment (of tweets, blogs, SMS):
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Term-level sentiment (within tweets, blogs, SMS)
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Aspect-level sentiment (in customer reviews):
 - SemEval-2014 Task 4

Detailed Description of the NRC-Canada Systems

- Message-level sentiment (of tweets, blogs, SMS):
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Term-level sentiment (within tweets, blogs, SMS)
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Aspect-level sentiment (in customer reviews):
 - SemEval-2014 Task 4

Message-Level Sentiment: The Task

Tweet: Happy birthday, Hank Williams. In honor of the Hank turning 88, we'll play 88 Hank songs in a row tonite @The_ZOO_Bar. #honkytonk

positive

Tweet: #Londonriots is trending 3rd worldwide This is NOT something to be proud of United Kingdom!!! Sort it out!!!!

negative

Tweet: On the night Hank Williams came to town.

neutral

Message-Level Sentiment : The Approach

- Pre-processing
 - URL -> http://someurl
 - UserID -> @someuser
 - Tokenization and part-of-speech (POS) tagging (CMU Twitter NLP tool)
- Classifier
 - Linear SVM (An in-house implementation by Colin Cherry)
- Evaluation
 - Macro-averaged F-pos and F-neg

Message-Level Sentiment : The Features

Features	Examples
n-grams	happy, am_very_happy, am*_happy
char n-grams	un, unh, unha, unhap
Emoticons	:D, >:(
hashtags	#excited, #NowPlaying
capitalization	YES, COOL
part of speech	N: 5, V: 2, A:1
negation	Neg:1
word clusters	probably, definitely, def, possibly, prob, ...
lexicons	count: 3; score: 2.45; max: 1.3; last: 0.6

Sentiment Lexicons

- Manual lexicons:
 - NRC Emotion Lexicon
 - MPQA Sentiment Lexicon
 - Bing Liu's Opinion Lexicon
- Automatically created lexicons:
 - Hashtag Sentiment Lexicon
 - Sentiment140 Lexicon

Message-Level Sentiment : The Data (Semeval-2013 Task 2)

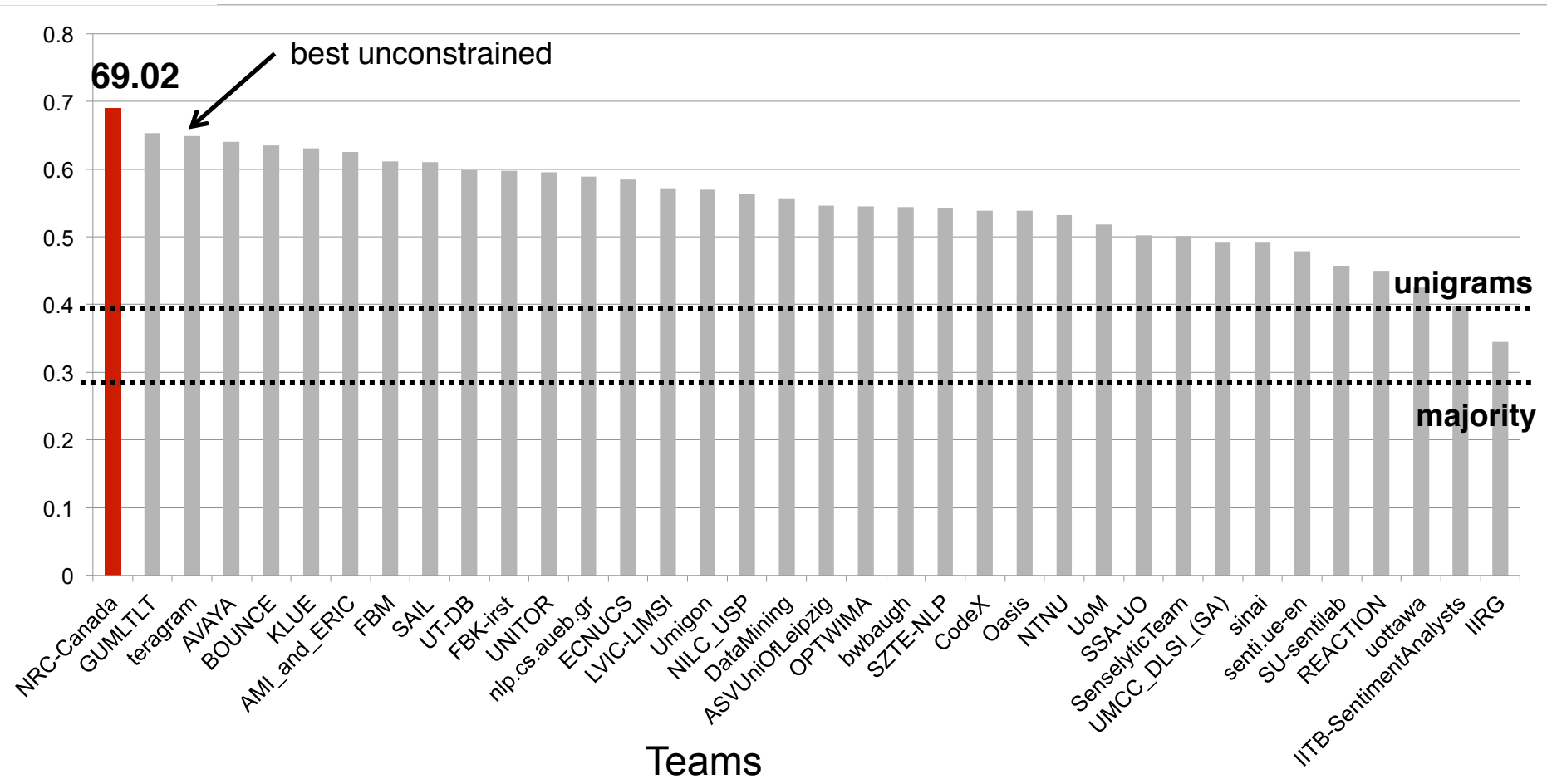
- Training: ~ 10,000 labeled tweets
 - positive: 40%
 - negative: 15%
 - neutral: 45%
- Test:
 - tweets: ~ 4,000
 - SMS: ~ 2,000

Official Performance/Rankings

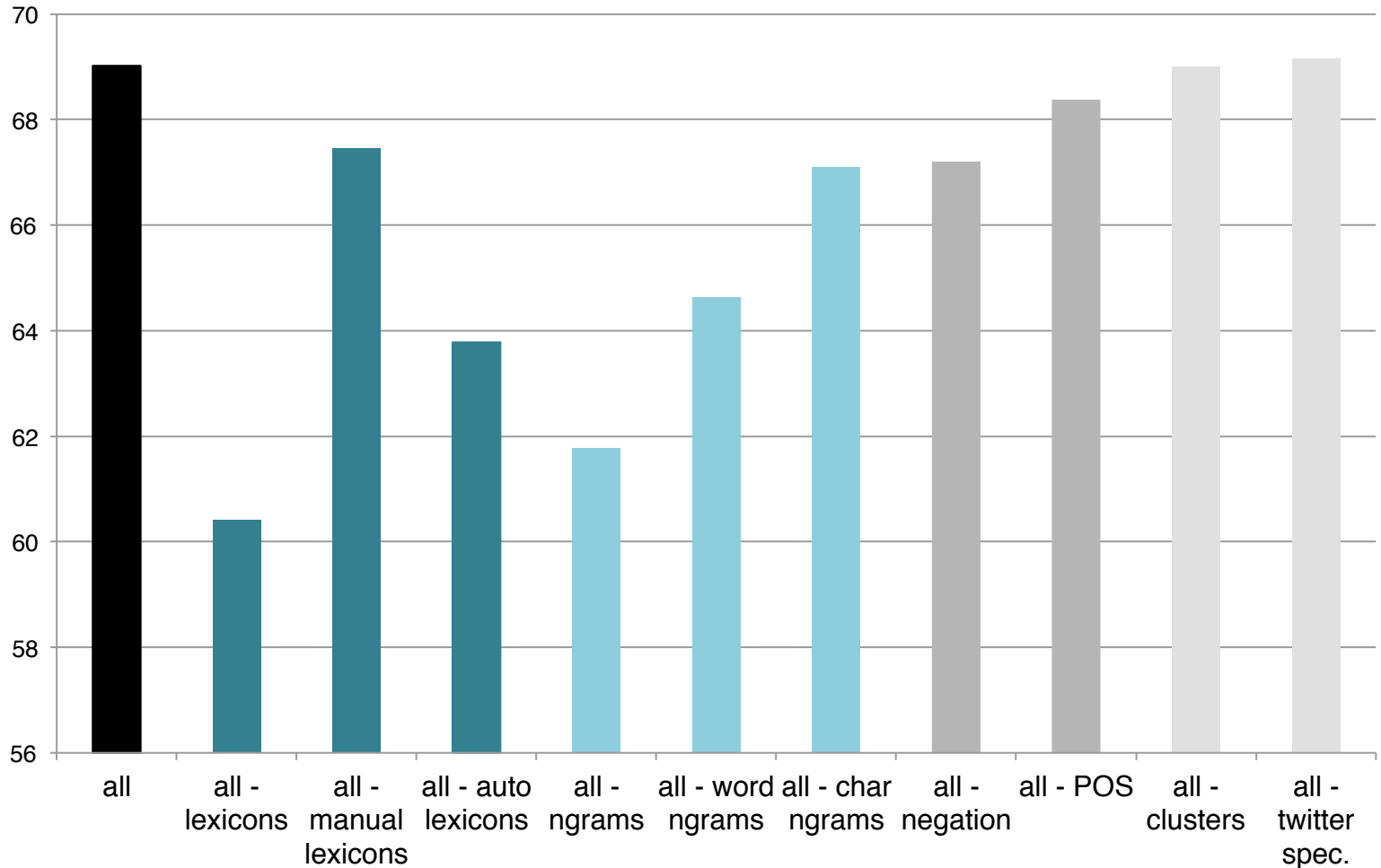
- Tweets
 - Macro-averaged F: 69.02
 - 1st place
- SMS
 - Macro-averaged F: 68.42
 - 1st place

Detailed Results on Tweets

F-score

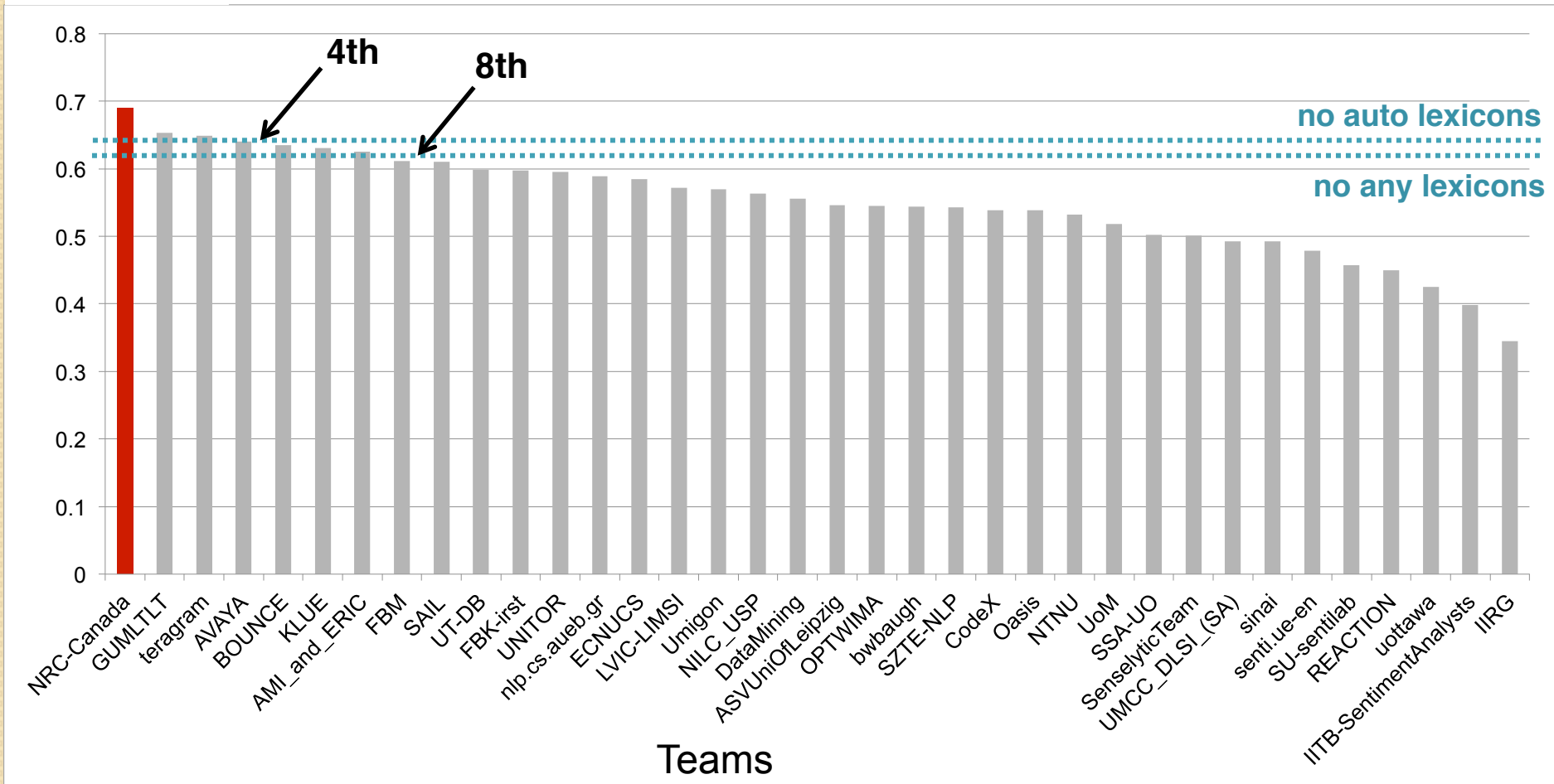


Feature Contributions on Tweets



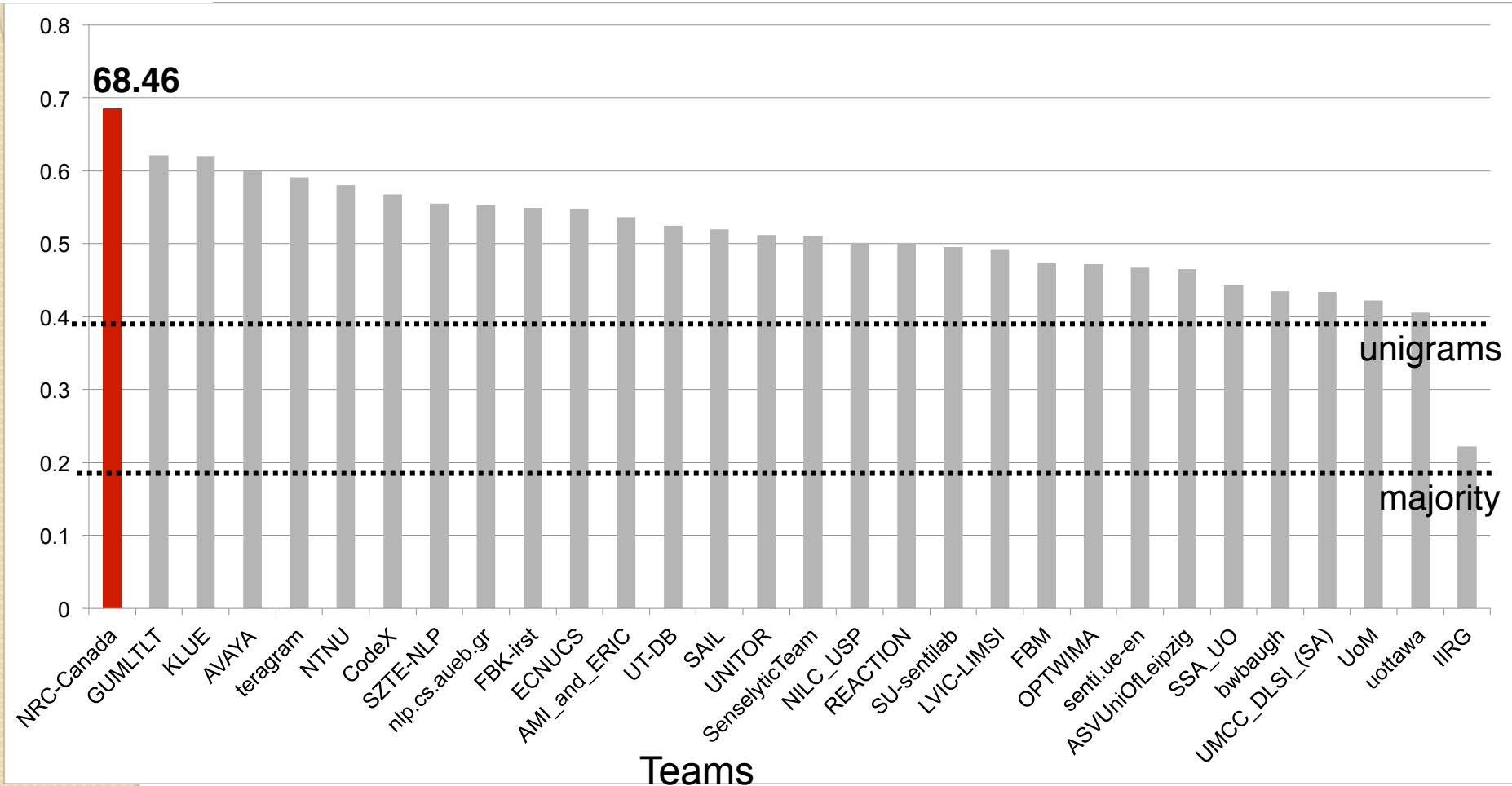
Detailed Results on Tweets (continued)

F-score

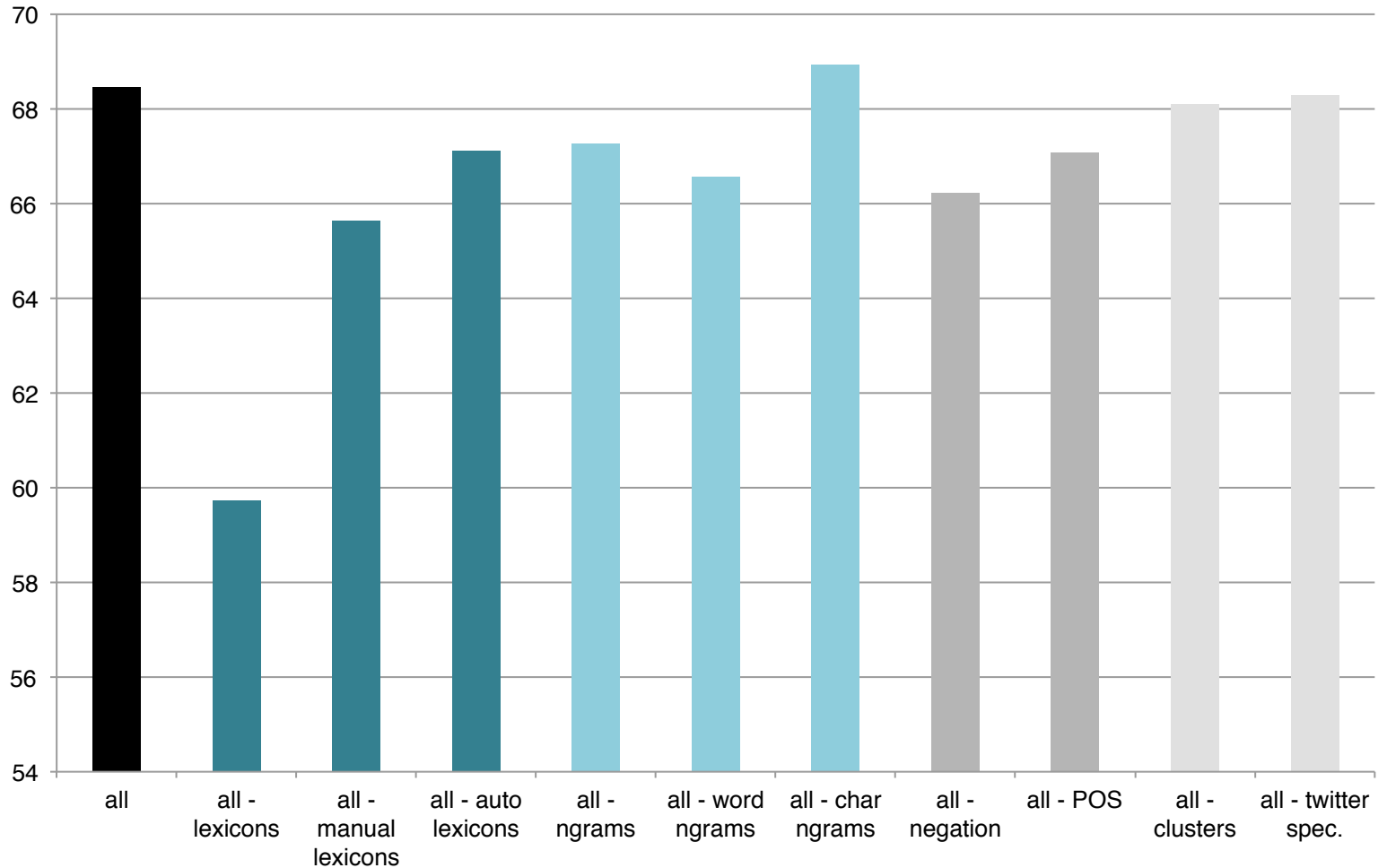


Detailed Results on SMS

F-score



Feature Contributions (on SMS)



Improving our Systems for SemEval-2014 Task 9

Key idea:

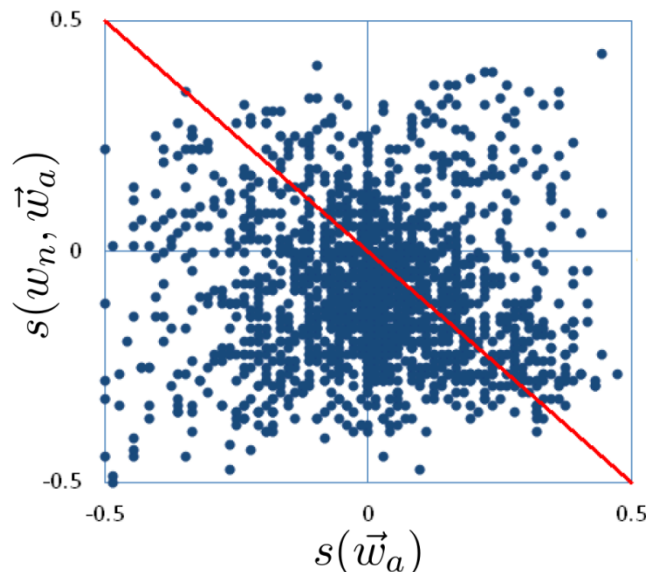
Improving sentiment lexicons to better cope with negation.

Complex Effect of Negation

- Why negation? Negation often significantly affects the sentiment of its scopes.

not very good
↑ ↙
negator w_n **argument** \vec{w}_a

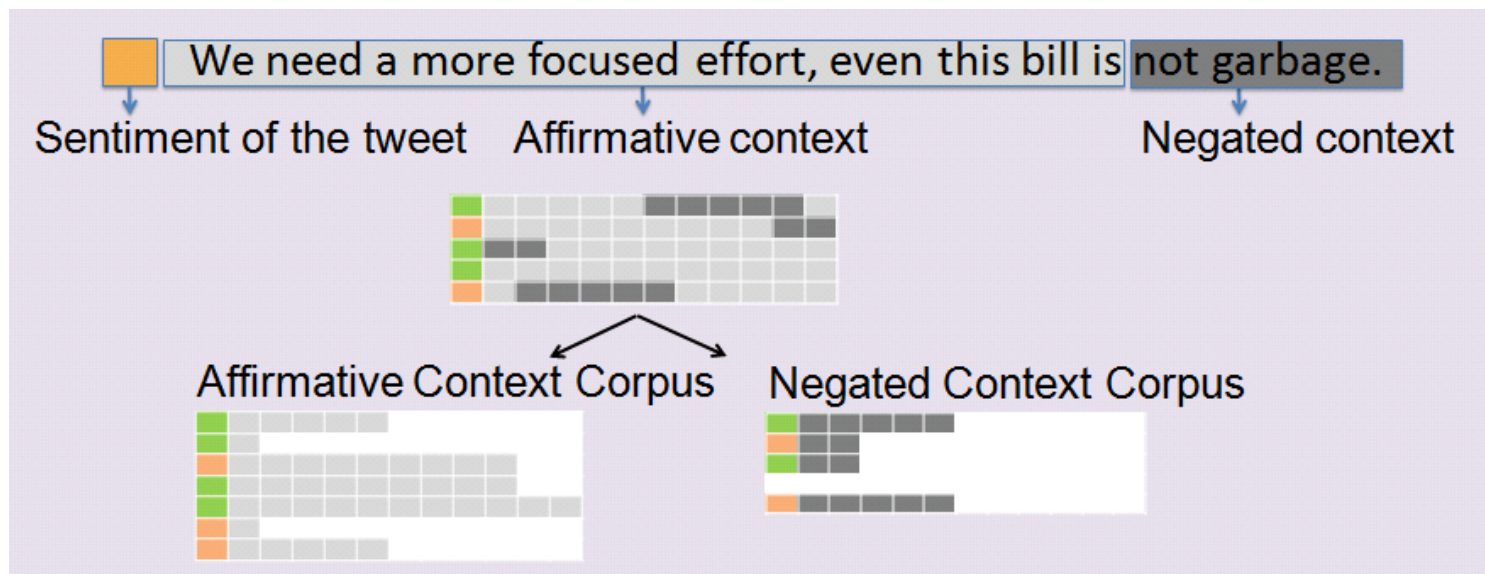
- This complex effect has recently been studied in Stanford Sentiment Tree Bank (Zhu et al., 2014; Socher et al., 2013)



- Non-lexicalized assumptions
 - Reversing
 - Shifting, Polarity-based shifting
- Simple lexicalized assumptions
 - Negator-based shifting
 - Combined shifting
- Sentiment composition
 - Recursive-neural-network-based composition

Improving the Systems for SemEval-2014 Task 9

- In our SemEval-2014 system, we adopted a lexicon-based approach (Kiritchenko et al., 2014) to determine the sentiment of words in affirmative and negated context.

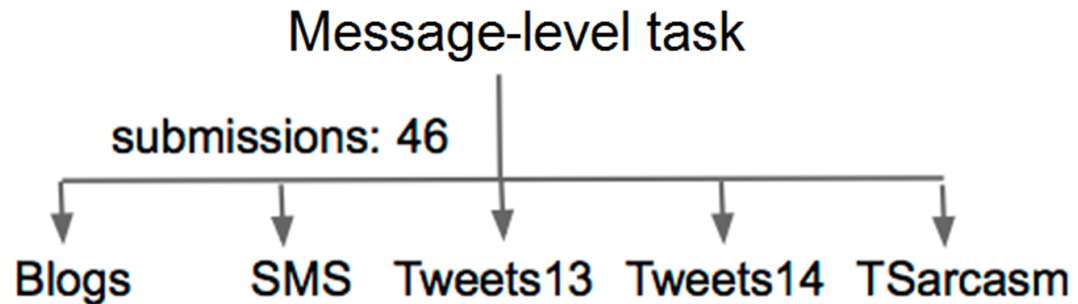


Message-Level Sentiment : The Data (Semeval-2014 Task 9)

- Training (same as in SemEval-2013): ~ 10,000 labeled tweets
 - positive: 40%
 - negative: 15%
 - neutral: 45%
- Test
 - Official 2014 data:
 - tweets: ~ 2,000
 - sarcastic tweets: ~ 100
 - LiveJournal blogs (sentences): ~ 1,000
 - Progress (SemEval-2013 test data):
 - tweets: ~ 4,000
 - SMS: ~ 2,000

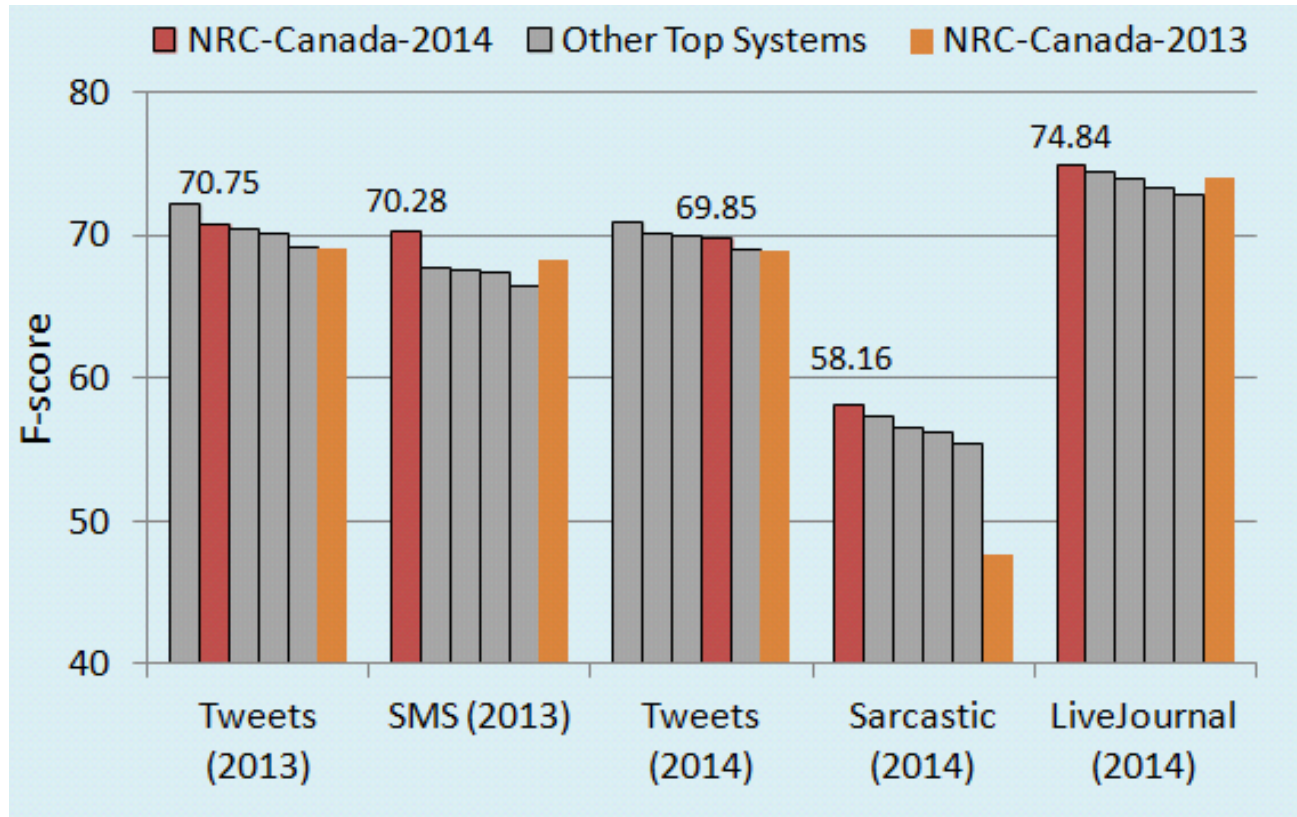
Official Performance/Rankings

- 1st on Micro-averaged F-score over all 5 test sets
- Details

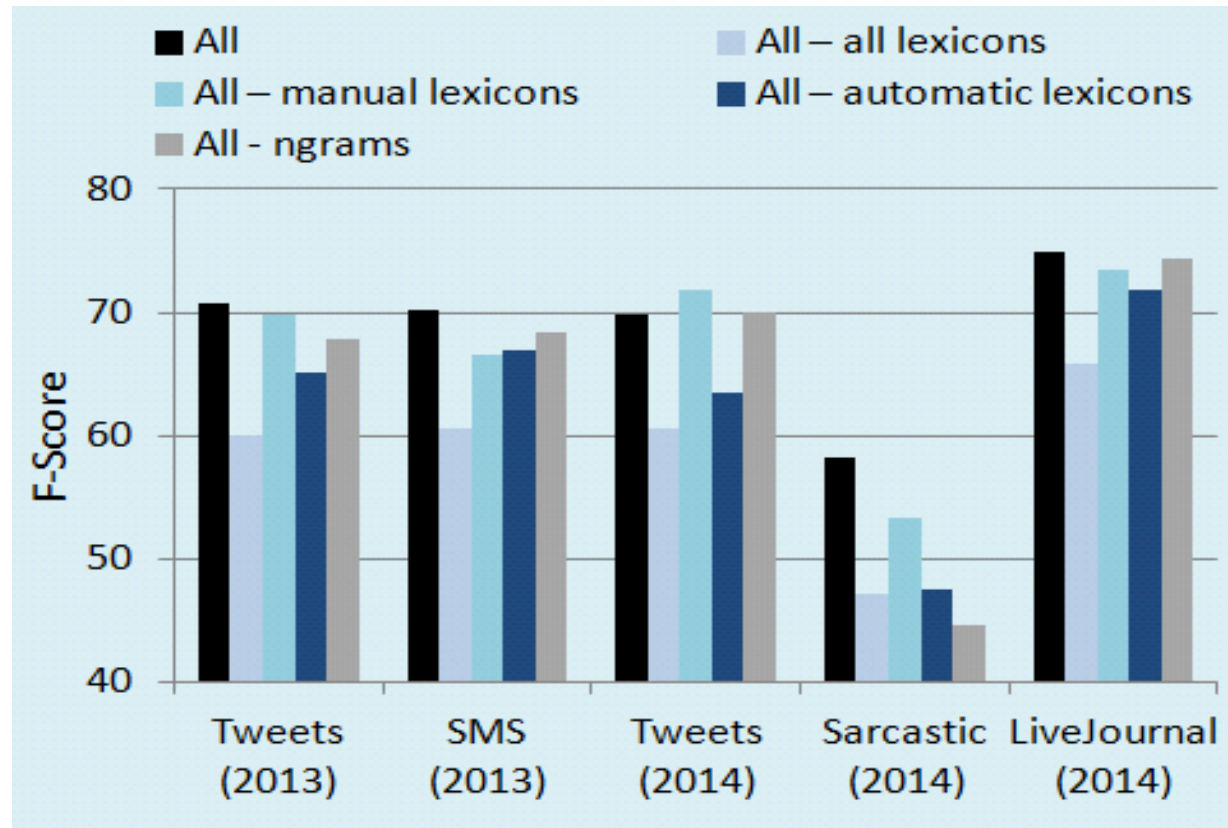


Our rankings: **1** **1** **2** **4** **1**

Official Performance/Rankings



Ablation Effects of Features



Message-Level Sentiment : Summary

- Significant improvement over NRC-Canada-2013 system
- Best micro- and macro-averaged results on all 5 datasets; best results on 3 out of 5 datasets
- System trained on tweets showed similar performance on SMS and LiveJournal blog sentences
- Strong performance on sarcastic tweets
- Most useful features on all datasets:
 - sentiment lexicons, especially automatic tweet-specific lexicons
- N-grams are very helpful for in-domain data (tweets), less helpful for out-of-domain data (SMS and LiveJournal)

Key Ideas from other Top Systems

- Coooolll
 - Use sentiment-specific word embeddings (details will be discussed later)
- TeamX
 - Parameters are fine-tuned towards the tweet datasets
 - This may explain why the system achieved the best results on the tweet sets but showed worse performance on the out-of-domain sets.
- RTRGO
 - Use random subspace learning (Søgaard and Johansen, 2012)
 - Train a classifier on a concatenation of 25 corrupted copies of the training set (each feature is randomly disabled with $\text{prob}=0.2$)

Key Ideas from other Top Systems

- Other ideas
 - Spelling correction
 - Careful normalization (e.g., for the elongated words)
 - Term/cluster weighting (e.g., TF-IDF)

Detailed Description of the NRC-Canada Systems

- Message-level sentiment (of tweets, blogs, SMS):
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Term-level sentiment (within tweets, blogs, SMS)
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Aspect-level sentiment (in customer reviews):
 - SemEval-2014 Task 4

Term-Level Sentiment : The Problem

Tweet: The new Star Trek does not have much of a story, but it is visually spectacular.

target is positive

Tweet: The new Star Trek does not have much of a story, but it is visually spectacular.

target is negative

Tweet: Spock displays more emotions in this Star Trek than the original series.

target is neutral

Further Clarification of the Problem

- The task is not defined as a sequential labeling problem:

Tweet: $\frac{w1}{obj} \frac{w2}{pos} \frac{w3}{neu} \frac{w4}{obj} \frac{w5}{neg} \frac{w6}{neg} \frac{w7}{neg} \frac{w8}{neg} \frac{w9}{neg} .$

- no boundary detection is required
 - no need to label all expressions in a tweet.
- It is an independent classification problem for each sentiment term.

Tweet: $w1 \frac{w2}{pos} \frac{w3}{neu} w4 \frac{w5}{neg} w6 w7 \frac{w8}{neg} \frac{w9}{neg} .$

Basic Feature Categories

Features	Description
term features	extracted from the target terms, including all the features discussed above.
context features	extracted from a window of words around a target term or the entire tweet, depending on features.

Detailed Features

Features	Description
ngrams	
word ngrams	“F-word” + “good”
char. ngrams	dis- un-
encodings	
emoticons	:-) D:< :@ :-
hashtags	#BiggestDayOfTheYear
punctuations	?! !!!
elongated word	sooooo
lexical	
manual lexicons	MPQA, NRC-emo, Liu’s, Turney & Littman's
automatic lexicons	in-house, Osgood
negation	
negations words	can’t n’t cant isnt
interaction w. others	negating lexical words that follow

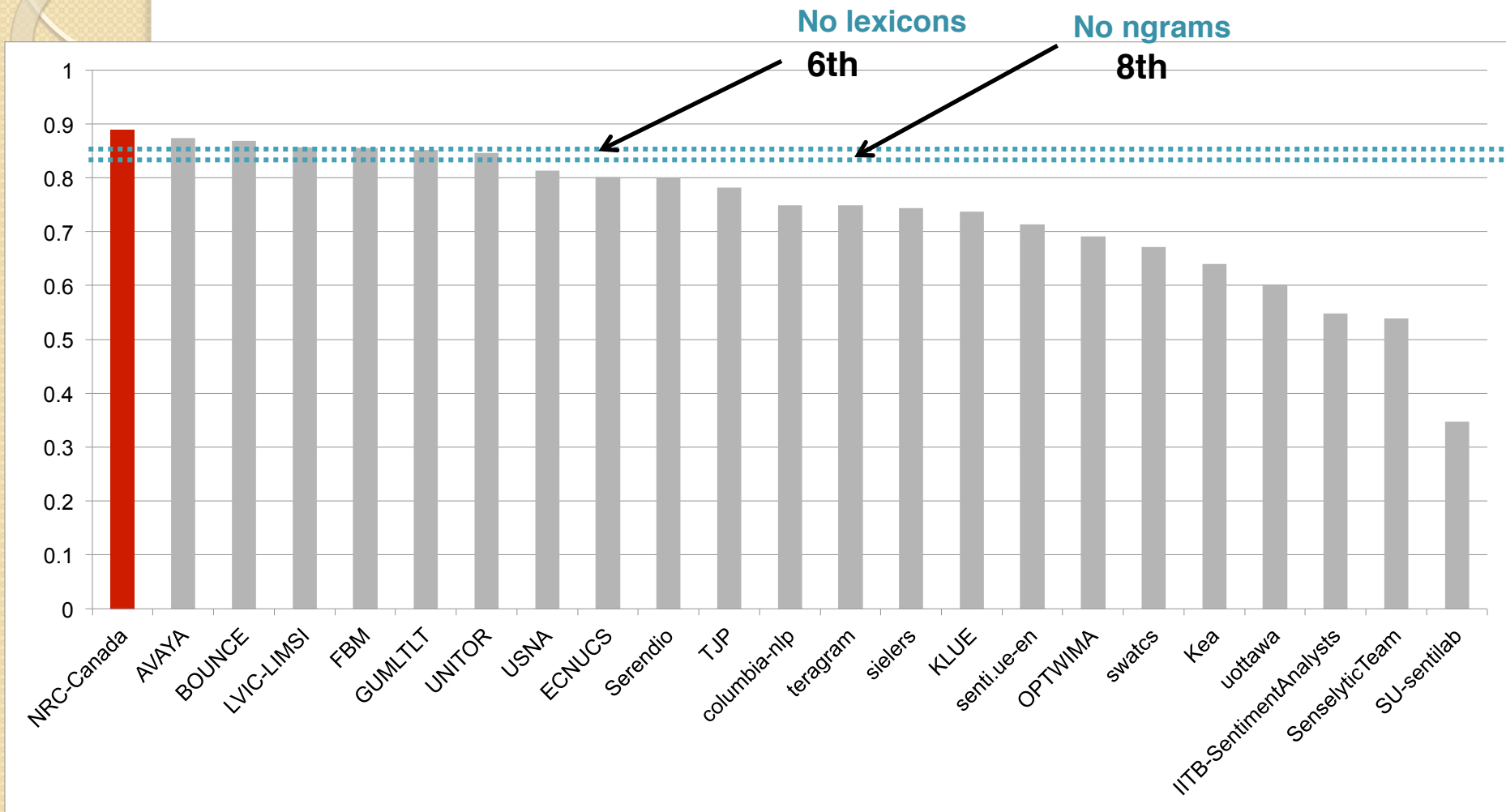
Classifier and Evaluation

- Classifier:
 - Linear SVM (Libsvm)
SVM has performed better than logistic regression (LR) on this task (Trick: the latter is much faster and corresponds well with SVM, so we used LR to quickly test ideas.)
- Evaluation:
 - Macro-averaged F-measure
(same as in the tweet-level task)

Official Performance/Rankings

- Tweets
 - Macro-averaged F: 89.10
 - 1st place
- SMS
 - Macro-averaged F: 88.34
 - 2st place

Detailed Results on Tweets



Term Features vs. Context Features

- Are contexts helpful? How much?

Experiment	Tweets	SMS
all features	89.10	88.34
all - target	72.97 (-16.13)	68.96 (-19.38)
all - context	85.02 (-4.08)	85.93 (-2.41)

- By large, sentiment of terms can be judged by the target terms themselves.
- The contextual features can additionally yield 2-4 points improvement on F-scores.

Discussion

Performance in the term-level task (~ 0.9) markedly higher than in message-level task (~ 0.7)

What does this mean?

- Is it harder for humans to determine sentiment of whole message?
 - Inter-annotator agreement scores will be helpful.
- Does the task set-up favors the term-level task?
 - About 85% of the target terms seen in training data
 - About 81% of the instances of a word in the training and test data have the same polarity

Key Ideas from other Top Systems

- GU-MLT-LT
 - Use on-line classifiers (stochastic gradient decent).
 - Careful normalization: all numbers are normalized to 0; repeated letters are also collapsed (liiike->like)
- AVAYA
 - Dependency parse features:
 - Edges that contain at least one target words
 - Paths between the head of term and the root of the entire message
- BOUNCE
 - Use term length features (intuition: longer terms are more likely to be neutral)
- Teragram
 - Use hand-written rules (as discussed earlier)

Improving the Systems for SemEval-2014 Task 9

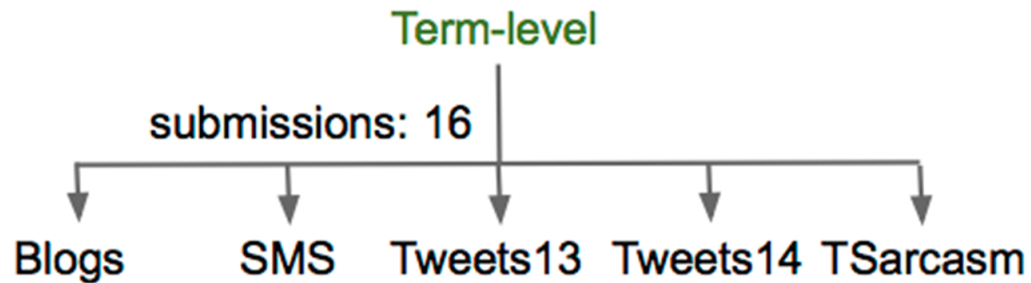
- Improving sentiment lexicons (as in message-level models)
 - Using a lexicon-based approach (Kiritchenko et al., 2014) to determining the sentiment of words in affirmative and negated context.
- Discriminating negation words
 - Different negation words, e.g. *never* and *didn't*, can affect sentiment differently (Zhu et al., 2014).
 - We made a simple, lexicalized modification to our system
This is never acceptable
The word *acceptable* is marked as `acceptable_not` in our old system but as `acceptable_beNever` in our new system.

Term-Level Sentiment : The Data (Semeval-2014 Task 9)

- Training (same as in SemEval-2013): 8,891 terms
 - positive: 62%; negative: 35%; neutral: 3%
- Test
 - Official 2014 data:
 - tweets: 2,473 terms
 - sarcastic tweets: 124
 - LiveJournal blogs: 1,315
 - Progress (SemEval-2013 test data):
 - tweets: 4,435
 - SMS: 2,334

Official Performance/Rankings

- 1st on Micro-averaged F-score over all 5 test sets
- Details



Our rankings:

2

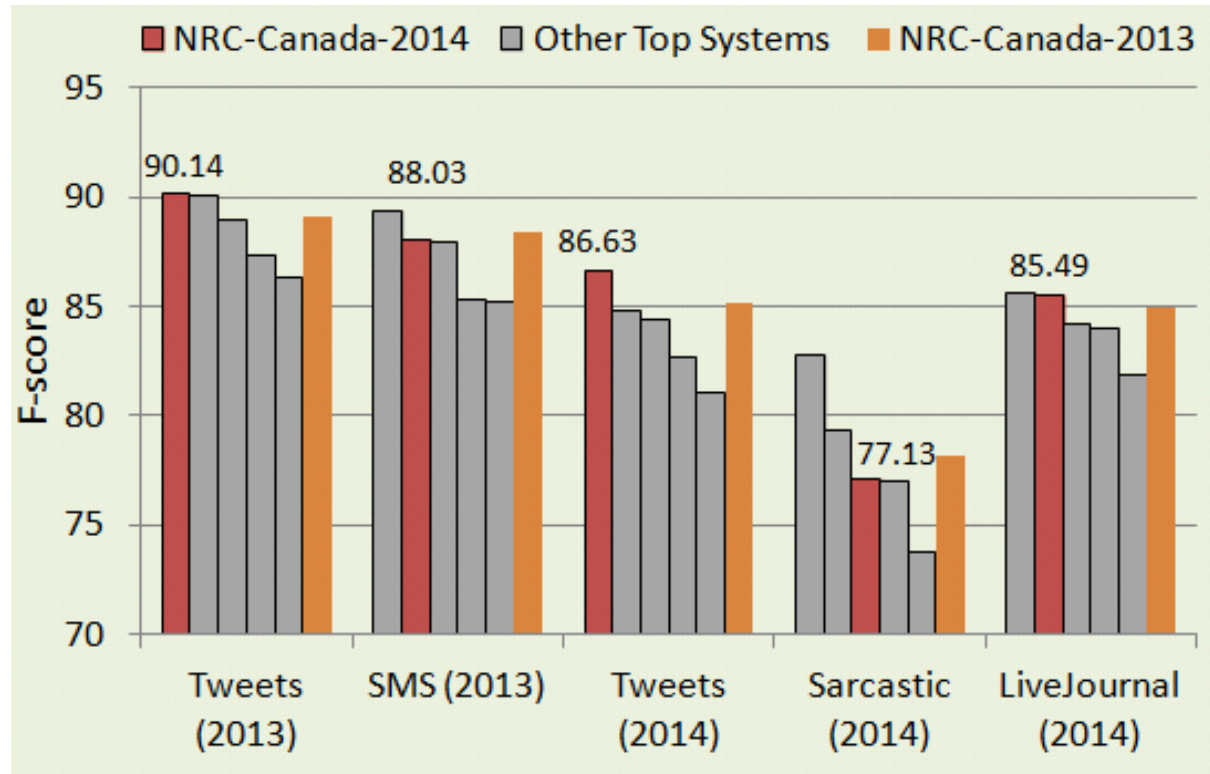
2

1

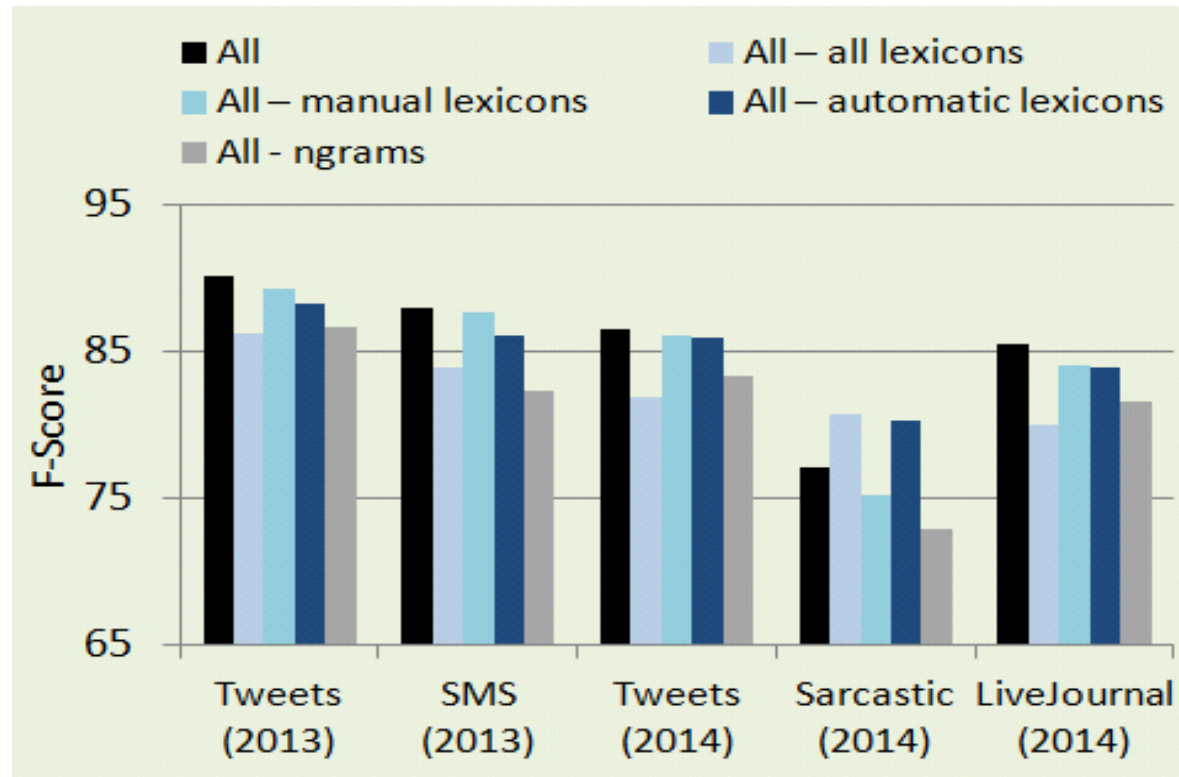
1

3

Official Performance/Rankings



Ablation Effects of Features



Summary

- Significant improvement over NRC-Canada-2013 system.
- Best micro-averaged results on all 5 datasets; best results on 2 out of 5 datasets.
- Effect of lexicon features
 - Sentiment lexicons automatically built from tweets are particularly effective in our models.
- Better handling of negation is helpful.

Key Ideas from other Top Systems

- SentiKLUE
 - Use message-level polarity, which is the 3rd most important feature category (following bag-of-words and sentiment lexicon features.)
- CMUQ-Hybrid
 - RBF kernel found to be the best kernel in the models(so different systems may still need to try different kernels during development.)

Key Ideas from other Top Systems

- CMUQ@Qatar
 - Careful preprocessing (5.2% gain was observed.)
 - Tokenizing also words that stick together (no space between them).
 - For any acronym, keeping both the acronym and the expanded version (e.g., LOL -> laugh out loudly)
- Think-Positive:
 - Apply a deep convolution neural network approach.

Detailed Description of the NRC-Canada Systems

- Message-level sentiment (of tweets, blogs, SMS):
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Term-level sentiment (within tweets, blogs, SMS)
 - SemEval-2013 Task 2, SemEval-2014 Task 9
- Aspect-level sentiment (in customer reviews):
 - SemEval-2014 Task 4

Aspect-Level Sentiment

- Sub-Task 1: **Aspect term extraction**
 - Find terms in a given sentence that are related to aspects of the products.
- Sub-Task 2: **Aspect term polarity**
 - Determine whether the polarity of each aspect term is positive, negative, neutral or conflict.
- Sub-Task 3: **Aspect category detection**
 - Identify aspect categories discussed in a given sentence (e.g., food, service)
- Sub-Task 4: **Aspect category polarity**
 - Determine the polarity of each aspect category.

Sub-Task 1: Aspect Term Extraction

- Find terms in a given sentence that are related to aspects of the products under review
 - I charge it at night and skip taking the **cord** with me because of the good **battery life**.
 - The **service** is outstanding and my **crab-cake eggs benedict** could not have been better.

Aspect Term Extraction: The Approach

- Semi-Markov discriminative tagger
 - Tags phrases, not tokens, can memorize “fish and chips”
 - Trained with MIRA using a basic feature set
 - For each token included in a phrase being tagged:
 - Word identity (cased & lowercased) in a 2-word window
 - Prefixes and suffixes up to 3 characters
 - For each phrase being tagged
 - Phrase identity (cased and lowercased)
- Our current system does not use any word clusters, embeddings or gazetteers/lexicons.

Aspect Term Extraction: The Data

- Restaurants

- Sentences: 3,041
- Term tokens: 3,693
- Term types: 1,212

food	376
service	238
prices	65
place	64
menu	57
staff	57
dinner	56
pizza	51
atmosphere	49
price	42

- Laptops

- Sentences: 3,045
- Term tokens: 2,358
- Term types: 955

screen	64
price	58
battery life	55
use	53
keyboard	52
battery	48
programs	37
features	35
software	34
warranty	31

Aspect Term Extraction: Results

Restaurants	Precision	Recall	F1
DLIREC	85.4	82.7	84.0
XRCE	86.2	81.8	84.0
NRC-Canada	84.4	76.4	80.2
UNITOR	82.4	77.9	80.1
IHS_RD	86.1	74.1	79.6
18 other teams...			

Laptops	Precision	Recall	F1
IHS_RD	84.8	66.5	74.6
DLIREC	82.5	67.1	74.0
NRC-Canada	78.8	60.7	68.6
UNITOR	77.4	57.6	68.0
XRCE	69.7	65.0	67.2
19 other teams...			

Key Ideas from other Top Systems

- DLIREC
 - Clusters (Brown/Word2Vec) built on Amazon and Yelp data.
 - Entity list harvested from "Double Propagation"
 - start with sentiment words, find noun phrases in unsupervised data that are modified by those sentiment words, declare those noun phrases entities, i.e.: "the rice is amazing" extracts "rice".
 - Syntactic heads (from Stanford parser) are important features.
- HIS_RD
 - Some domain-independent word lists appeared to be helpful.

Key Ideas from other Top Systems

- UNITOR
 - Word vectors built using word co-occurrence + LSA on Opinosis (laptop) and TripAdvisor datasets
- XRCE
 - Rule-based post-processing of output from a syntactic parser
 - Parser's lexicon augmented with terms from training data, Wordnet synonyms, and food terms list from Wikipedia "Food Portal".

Aspect-level sentiment: Sub-Tasks

- Sub-Task 1: **Aspect term extraction**
 - Find terms in a given sentence that are related to aspects of the products.
- Sub-Task 2: **Aspect term polarity**
 - Determine whether the polarity of each aspect term is positive, negative, neutral or conflict.
- Sub-Task 3: **Aspect category detection**
 - Identify aspect categories discussed in a given sentence (e.g., food, service)
- Sub-Task 4: **Aspect category polarity**
 - Determine the polarity of each aspect category.

Aspect Term Polarity: The Task

The asian salad of Great Asian is barely eatable.

Task: in the sentence above, what's the sentiment expressed towards the target term “*asian salad*”?

Aspect Term Polarity: The Task

- This is different from the “term-level” problem in Task 9

The asian salad of Great Asian is barely eatable.

Task 4: aspect terms

Task 9: sentiment terms

Task 9: *phrase-level sentiment analysis*

Task 4: *sentiment towards a target*

- System concerns
 - The task-9 systems do not consider syntactic features, but task-4 systems should.
 - Task-9 depends mainly on the sentiment terms themselves, while task 4 relies more on context.
 - A task-9 model can be a component of task 4.

Aspect Term Polarity: The Features

- Surface features
 - Unigrams
 - Context-target bigrams (formed by a word from the surface context and a word from the target term itself)
- Lexicon features
 - Number of positive/negative tokens
 - Sum/maximum of the tokens' sentiment scores

Aspect Term Polarity: The Features

- Syntactic features
 - Consider long-distance sentiment phrases
The ma-po tofu, though not as spicy as what we had last time, is actually great too.
 - Consider local syntax
a serious sushi lover
-
- Word- and POS-ngrams in the parse context
 - Context-target bigrams, i.e., bigrams composed of a word from the parse context and a word from the target term
 - All paths that start or end with the root of the target terms
 - Sentiment terms in parse context

Aspect Term Polarity: The Data

- Customer reviews
 - Laptop data
 - Training: 2358 terms
 - Test: 654 terms
 - Restaurant data
 - Training: 3693 target terms
 - Test: 1134 terms
- Pre-processing
 - We tokenized and parsed the provided data with Stanford CoreNLP Toolkits to obtain (collapsed) typed dependency parse trees ([de Marneffe et al., 2006](#)).

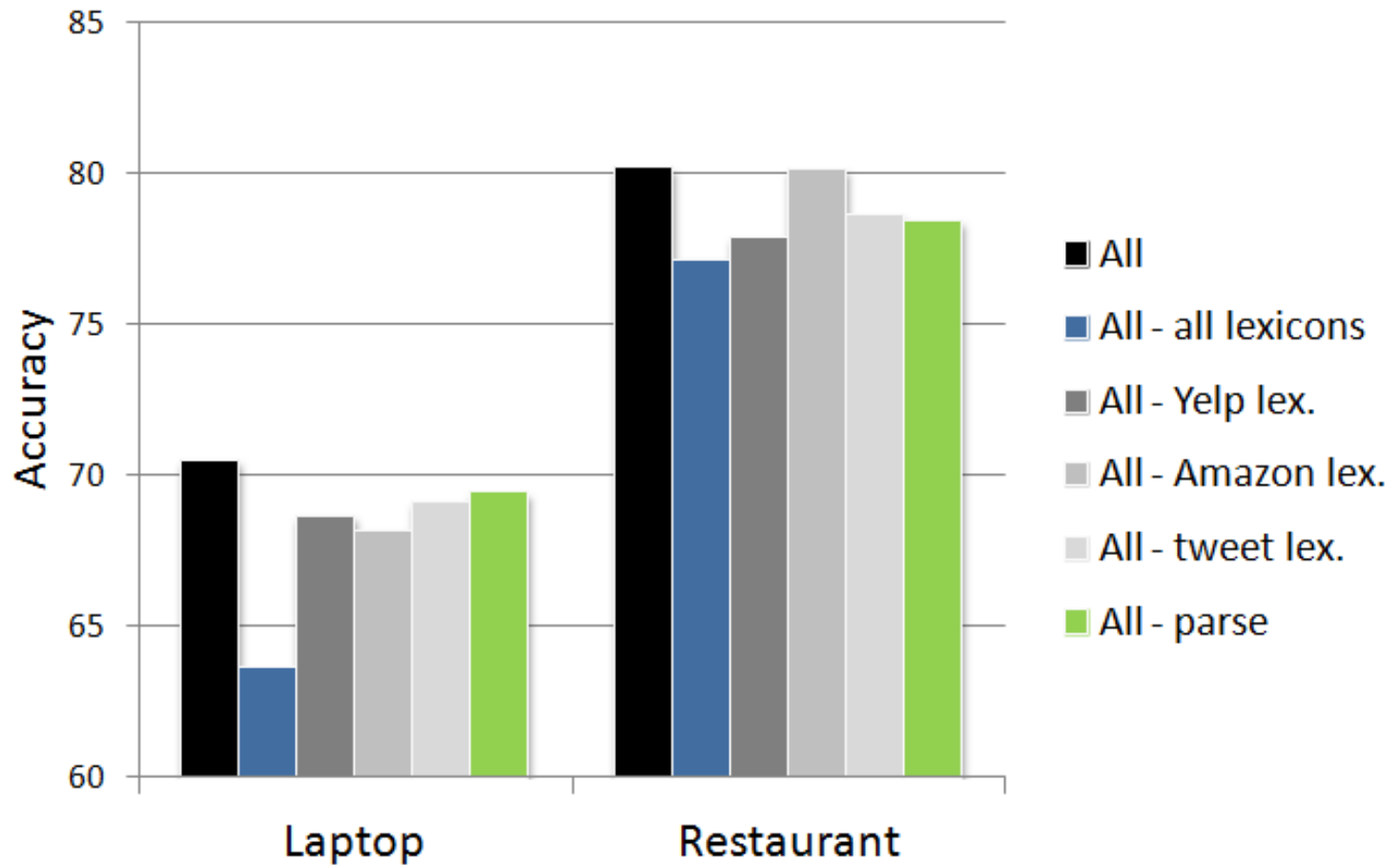
Aspect Term Polarity: Set-Up

- Classifier
 - Libsvm with the linear kernel
- Evaluation metric
 - Accuracy

Aspect Term Polarity: Results

- Laptop reviews
 - Accuracy: 70.49
 - 1st among 32 submissions from 29 teams
- Restaurant reviews
 - Accuracy: 80.16
 - 2nd among 36 submissions from 29 teams

Aspect Term Polarity: Contributions of Features



Aspect Term Polarity: Summary

- Our systems achieve an accuracy of 70.49 on the laptop reviews and 80.16 on the restaurant data, ranking 1st and 2nd, respectively.
- The sentiment lexicon and parse features are critical to help us achieve the performance.
- Carefully designed features are also important: distance-weighted features, normalization & tokenization, etc.

Key Ideas from other Top Systems

- DCU
 - Sentiment scores of a word are reweighted by the distance between the word and the target aspect terms.
 - Different types of distances are calculated: surface, the number of discourse chunks, dependency parse path length.
 - Multiple manual sentiment lexicons are combined together before being used.
 - Manually built domain-specific lexicons: “*mouthwatering*”, “*watering*”.
 - Careful normalization & tokenization: spelling check, multiword processing (e.g., word with the form x-y)

Key Ideas from other Top Systems

- SZTE-NLP
 - Syntax-based features showed to be the best category of features.
 - If a sentence contains multiple aspect terms, identifying the range associated with each target aspect.
 - Each n-gram feature is weighted by the distance of the n-gram to the target aspect term.
 - Dependency parse trees are used to select the words around aspect terms.
 - Use aspect categories as features.
- UWB
 - Each n-gram feature is weighted by the distance of the n-gram to the target aspect term (using a Gaussian distribution.)

Key Ideas from other Top Systems

- XRCE
 - The entire system is built around sentiment-oriented dependency parser
 - Parse trees were annotated with sentiment information.
 - Rules are used to link sentiment on terms based on the parse.
 - Hybridizing rule based parse with machine learning.
- Ubham
 - Detect sentiment of text using lexicon-based methods, then assign that to different clauses using dependency parse trees.

Aspect-level sentiment: Sub-Tasks

- Sub-Task 1: **Aspect term extraction**
 - Find terms in a given sentence that are related to aspects of the products.
- Sub-Task 2: **Aspect term polarity**
 - Determine whether the polarity of each aspect term is positive, negative, neutral or conflict.
- Sub-Task 3: **Aspect category detection**
 - Identify aspect categories discussed in a given sentence (e.g., food, service)
- Sub-Task 4: **Aspect category polarity**
 - Determine the polarity of each aspect category.

Aspect Category Detection: The Task

- Identify aspect categories discussed in a given sentence extracted from a restaurant review

To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.

Aspect categories: food, miscellaneous

- A second stage will assign sentiment to each of these categories

Aspect Category Detection: The Approach

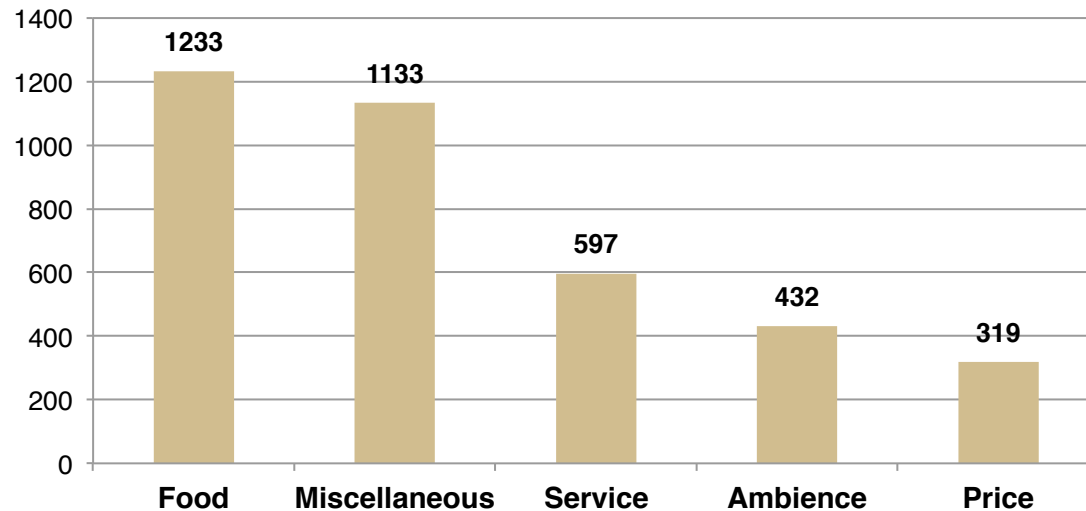
- Pre-processing
 - Tokenization (CMU Twitter NLP tool)
 - Stemming (Porter stemmer)
- Classifier
 - SVM with linear kernel (Colin's implementation)
 - Five binary classifiers (one-vs-all), one for each aspect category
- Evaluation
 - Micro-averaged F1-score

Aspect Category Detection: The Approach

- Features
 - Ngrams
 - Stemmed ngrams
 - Character ngrams
 - Word cluster ngrams
 - Yelp Restaurant Word – Aspect Association Lexicon features
- Post-processing
 - If no category is assigned,
 - $c_{\max} = \operatorname{argmax}_c P(c|d)$
 - assign c_{\max} if $P(c_{\max}|d) \geq 0.4$

Aspect Category Detection: The Data

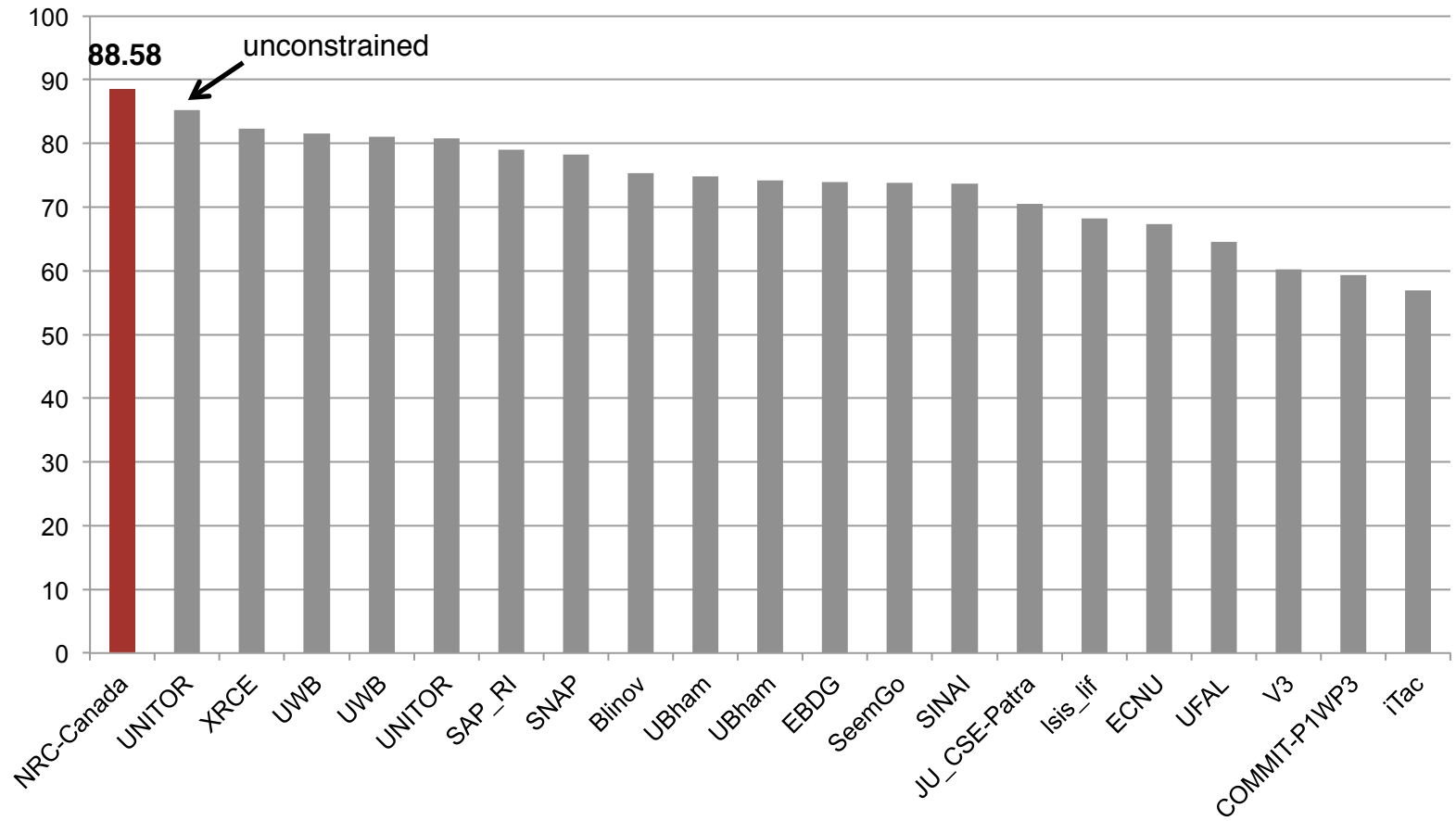
- Training:
 - 3044 sentences with at least one aspect category
 - 574 sentences (19%) with more than one aspect category



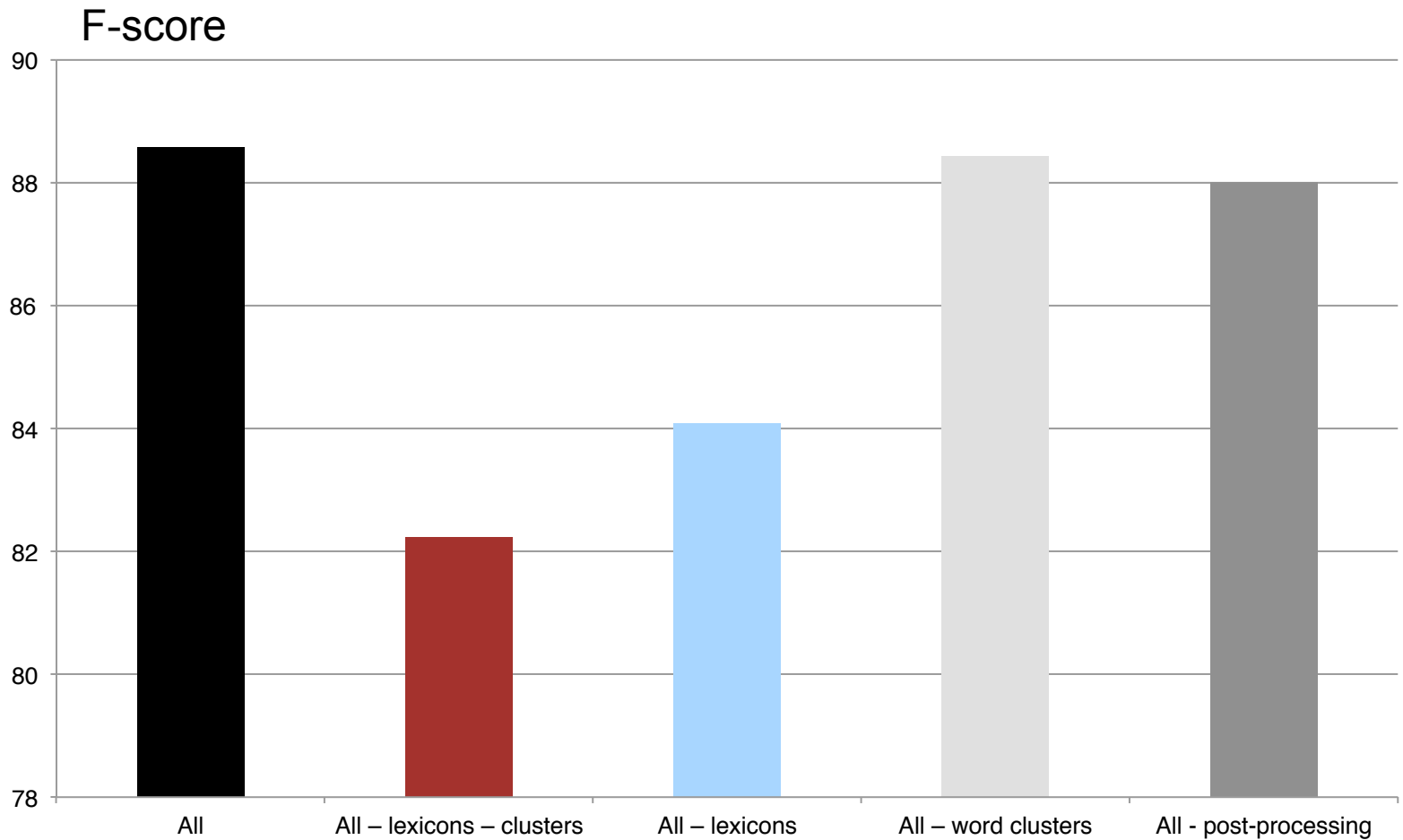
- Test:
 - 800 Sentences with at least one aspect category
 - 189 sentences (24%) with more than one aspect category

Aspect Category Detection: Results

F-score



Aspect Category Detection: Feature Contributions



Aspect Category Detection: Summary

- The system ranked first among 18 teams
- Most useful features:
 - in-domain word – aspect association lexicon

Aspect-level sentiment: Sub-Tasks

- Sub-Task 1: **Aspect term extraction**
 - Find terms in a given sentence that are related to aspects of the products.
- Sub-Task 2: **Aspect term polarity**
 - Determine whether the polarity of each aspect term is positive, negative, neutral or conflict.
- Sub-Task 3: **Aspect category detection**
 - Identify aspect categories discussed in a given sentence (e.g., food, service)
- Sub-Task 4: **Aspect category polarity**
 - Determine the polarity of each aspect category.

Aspect Category Polarity: The Task

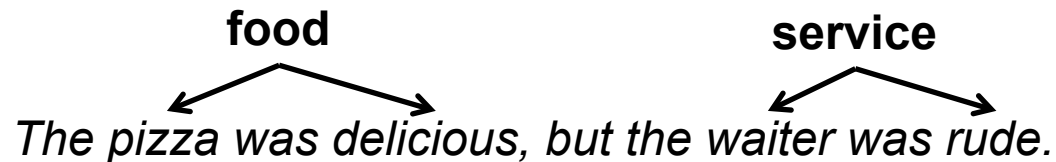
- Determine the polarity (positive, negative, neutral, or conflict) of each aspect category discussed in a given sentence extracted from a restaurant review

To be completely fair, the only redeeming factor was the food, which was above average, but couldn't make up for all the other deficiencies of Teodora.

Aspect categories: food (positive), miscellaneous (negative)

Aspect Category Polarity: The Features

- Standard features
 - ngrams, character ngrams
 - word cluster ngrams
 - sentiment lexicon features
 - negation
- Task-specific features
 - find terms associated with a given aspect category using Yelp Restaurant Word – Aspect Association Lexicon
 - Add standard features generated just for those terms

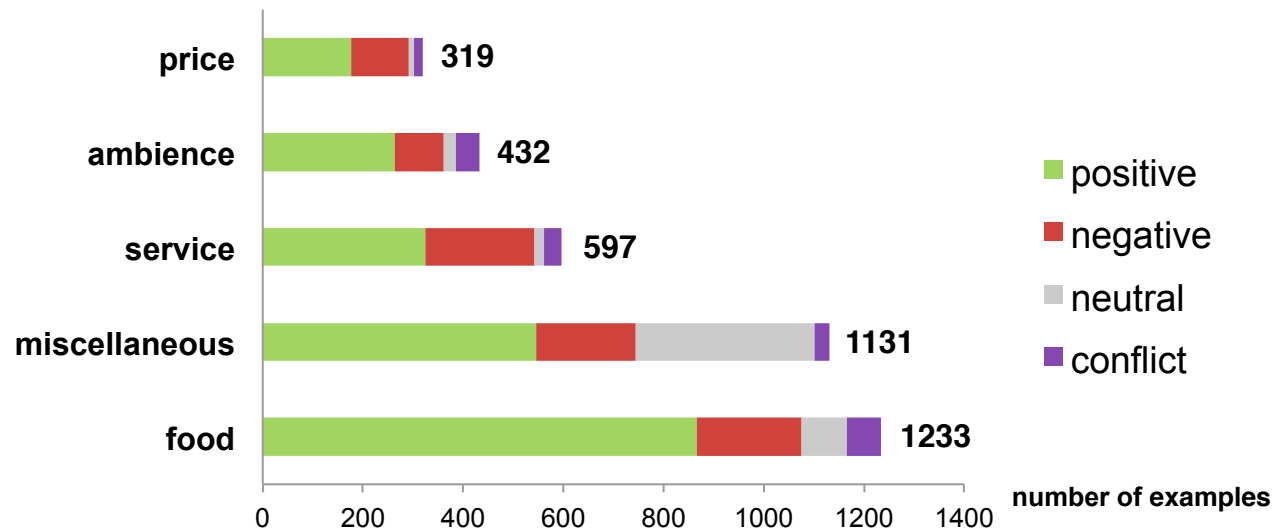


Aspect Category Polarity: Sentiment Lexicons

- Manual lexicons
 - NRC Emotion Lexicon
 - MPQA Sentiment Lexicon
 - Bing Liu's Opinion Lexicon
- Automatically created lexicons
 - **Yelp Restaurant Sentiment Lexicons: AffLex and NegLex**
 - Hashtag Sentiment Lexicons: AffLex and NegLex
 - Sentiment140 Lexicons: AffLex and NegLex

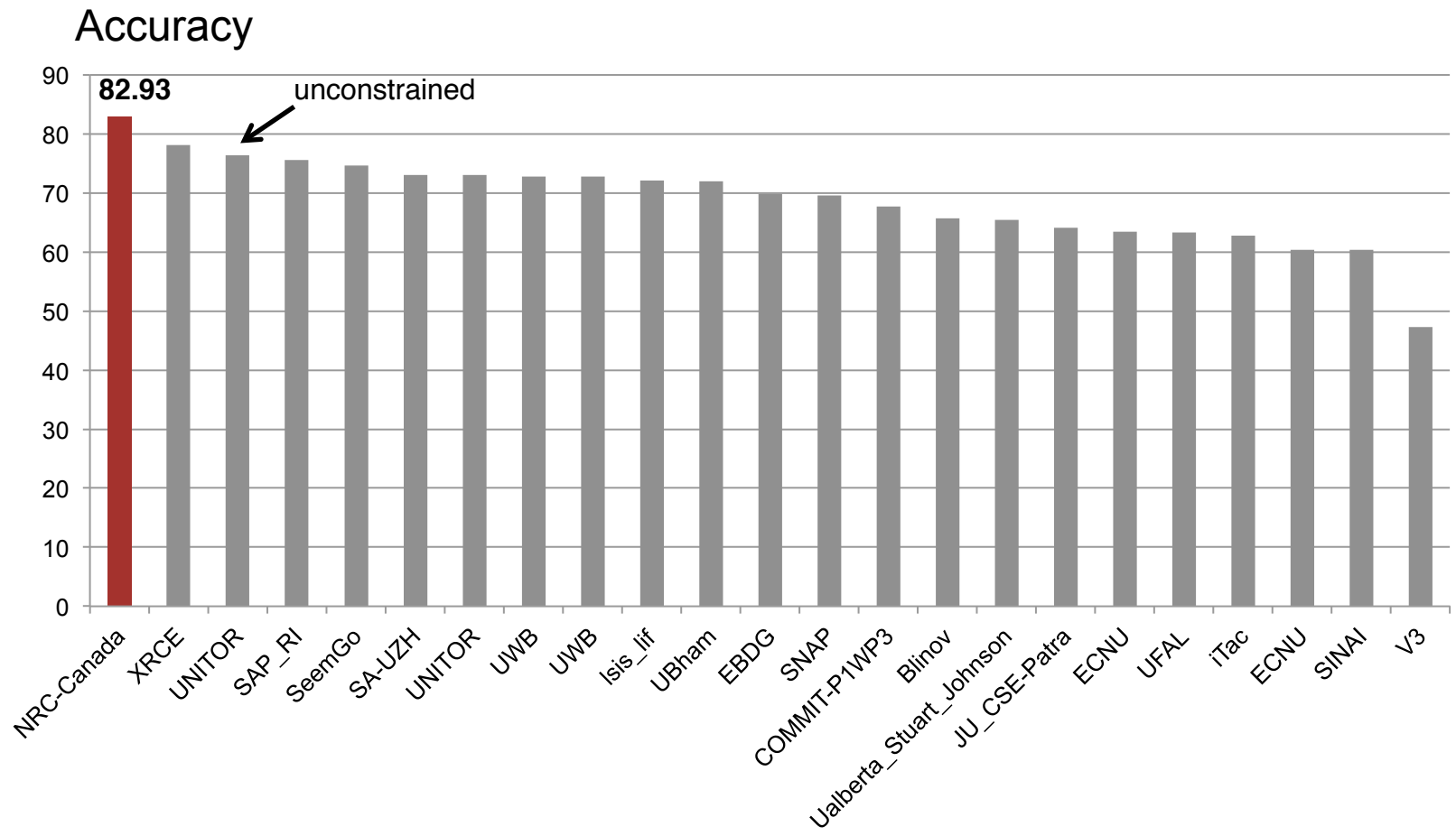
Aspect Category Polarity: The Data

- Training
 - 3044 sentences with at least one aspect category
 - 574 sentences (19%) with more than one aspect category
 - 167 sentences (5%) with aspect categories having different polarities



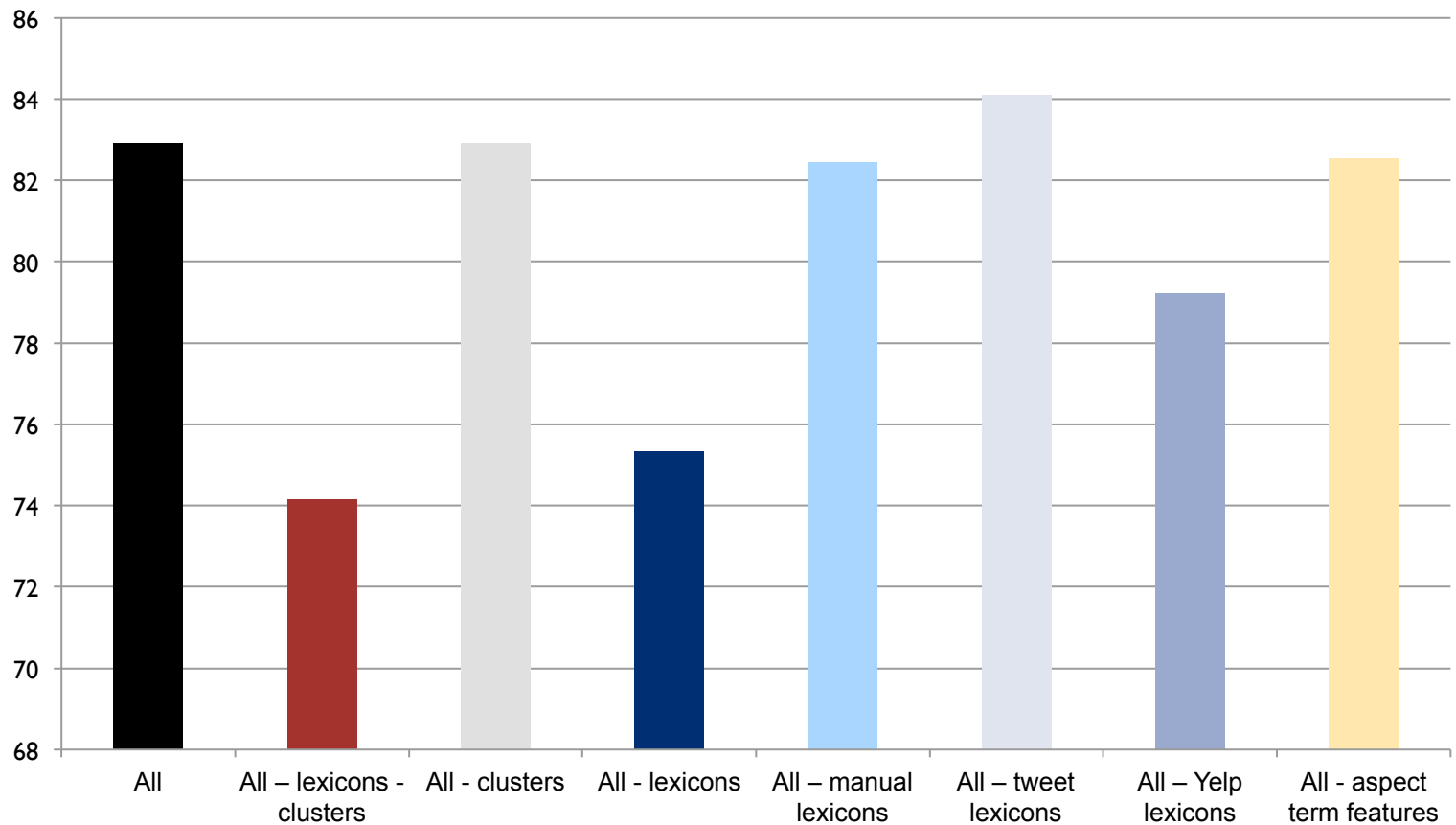
- Test
 - 800 sentences with at least one aspect category
 - 189 sentences (24%) with more than one aspect category
 - 42 sentences (5%) with aspect categories having different polarities

Aspect Category Polarity: Results



Aspect Category Polarity: Feature Contributions

Accuracy



Aspect Category Polarity: Summary

- The system ranked first among 20 teams
- Most useful features:
 - sentiment lexicons, especially in-domain automatic lexicon

Key Ideas from other Top Systems

- XRCE
 - Built around sentiment-oriented dependency parser
 - Parse trees were annotated with sentiment information.
 - Rules are used to link sentiment on terms based on the parse.
 - Hybridizing rule based parse with machine learning.
- UNITOR
 - Linear combination of different kernels
 - LSA features obtained on word-context matrix derived from a large-scale in-domain unlabeled corpus
- UWB
 - Use topic-modeling features obtained with Latent Dirichlet Allocation (LDA)


Overview of Sentiment Analysis Systems

- Rule-based systems
- Conventional statistical systems
- Deep-learning-based models
 - Sentiment word embedding
 - Sentiment composition

General Word Embedding

- Word embedding: representation of lexical items as points in a real-valued (low-dimensional) vector space.
- It is often computed by compressing a larger matrix to smaller one.

new		1		2	6			9		3	...
old	1	1			2	1		4		2	...
good	1		6	3		1		7	1		...
bad	2	1	4			2			3		...
...											...




new	-0.03	0.5	0
old	-0.04	0.3	0
good	1.4	0	2.5
bad	1.3	0	3.6
...			

- Keep (semantically or syntactically) close items in the original matrix/space to be close in the embedding space.

General Word Embedding

- Word embedding: representation of lexical items as points in a real-valued (low-dimensional) vector space.
- It is often computed by compressing a larger matrix to smaller one.

new		1		2	6			9		3	...
old	1	1			2	1		4		2	...
good	1		6	3		1		7	1		...
bad	2	1	4			2			3		...
...											...

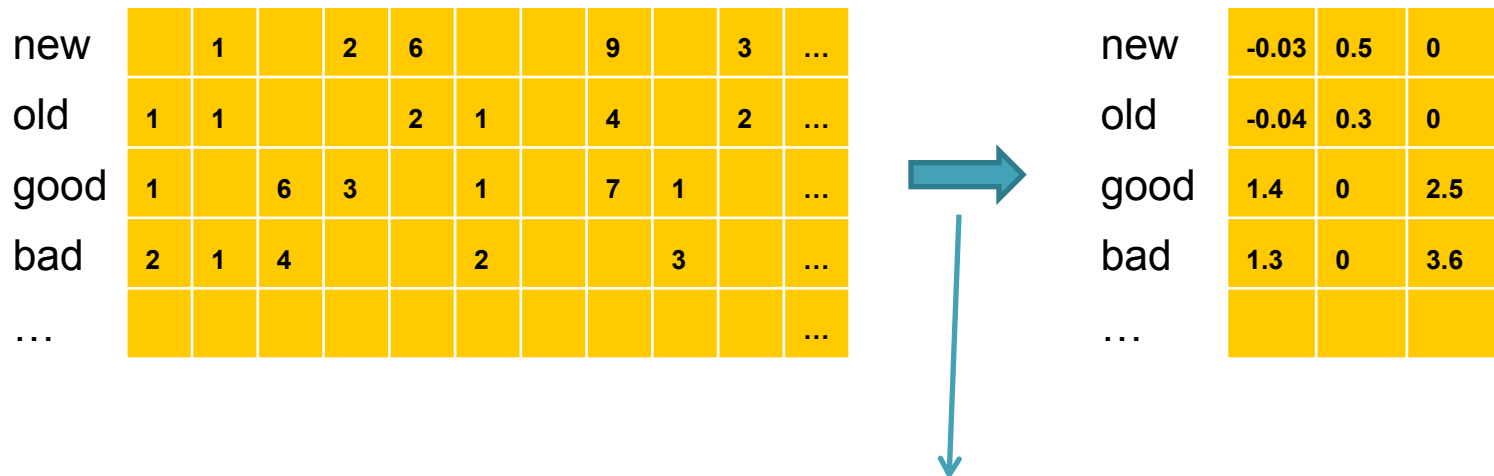


new	-0.03	0.5	0
old	-0.04	0.3	0
good	1.4	0	2.5
bad	1.3	0	3.6
...			

There are many ways to construct this matrix, e.g., using word-context or word-document counts.

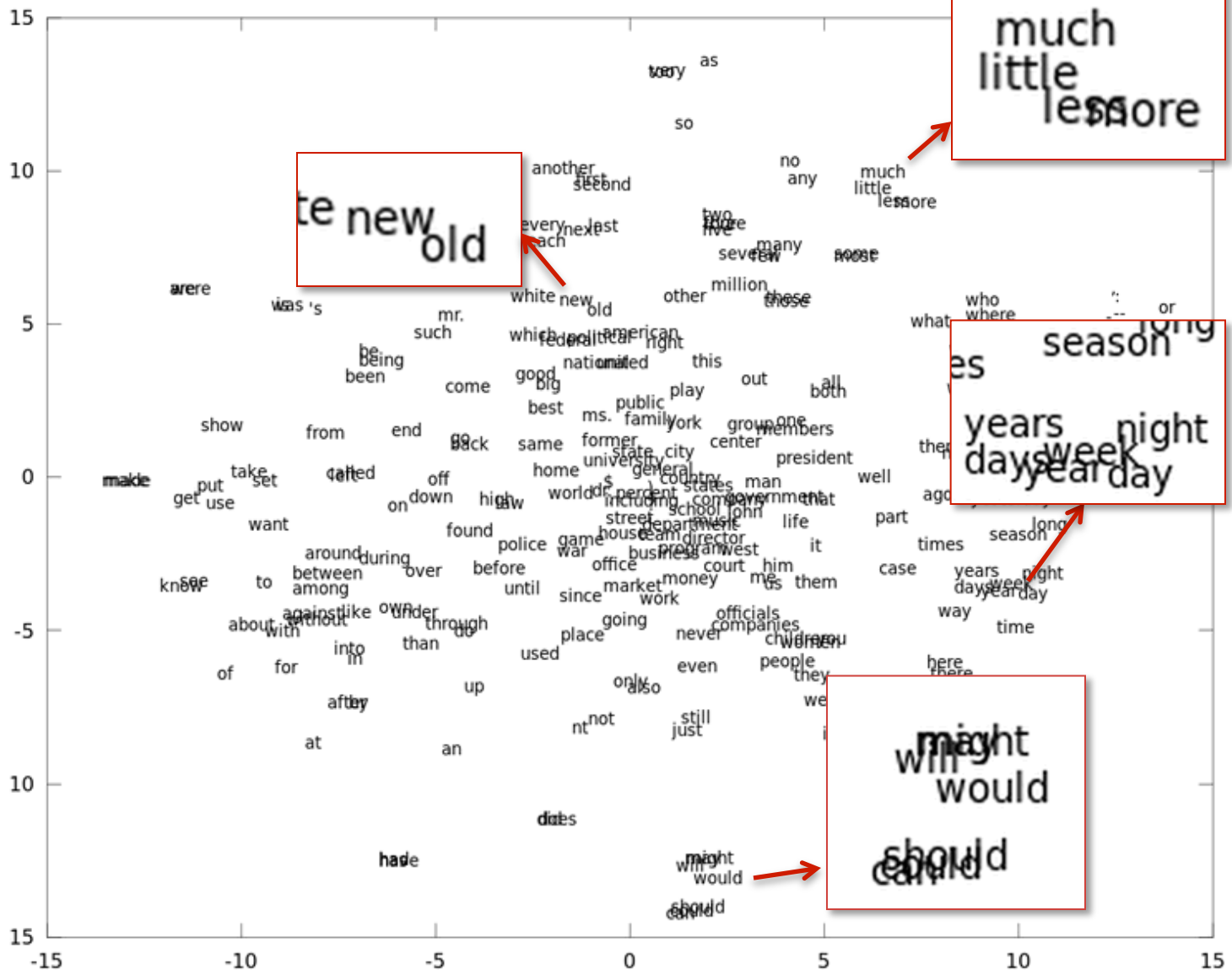
General Word Embedding

- Word embedding: representation of lexical items as points in a real-valued (low-dimensional) vector space.
- It is often computed by compressing a larger matrix to smaller one.



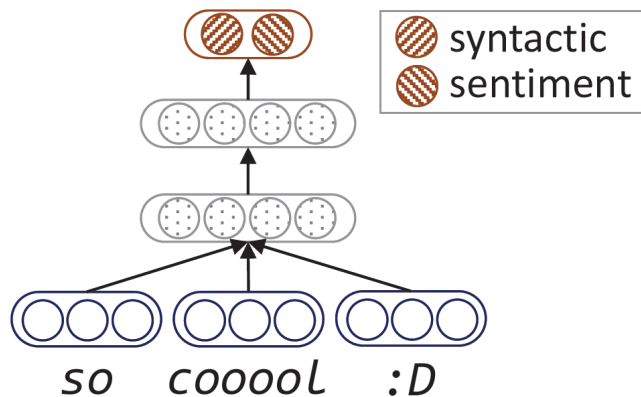
Also, there are many ways to compress the matrix, e.g., PCA, LLE, SNE, C&W, and Word2Vec.

General Word Embedding



Sentiment Word Embedding

- Cooooll (Tang et al., 2014): adapt syntactic/semantic word embedding to consider sentiment information.
 - Motivation: word *new* and *old* often have similar syntactic/semantic embedding but should have different sentiment embedding.
 - Approach: a linear modification of the objective functions.



$$loss_u(t, t^r) = \alpha \cdot loss_{cw}(t, t^r) + (1 - \alpha) \cdot loss_{us}(t, t^r)$$

$$loss_{cw}(t, t^r) = \max(0, 1 - f^{cw}(t) + f^{cw}(t^r))$$

$$loss_{us}(t, t^r) = \max(0, 1 - \delta_s(t) \mathbf{f}_1^u(t) + \delta_s(t) \mathbf{f}_1^u(t^r))$$

Results

Method	Positive/Negative/Neutral				
	T1	T2	T3	T4	T5
SSWE	70.49	64.29	68.69	66.86	50.00
CooooIII	72.90	67.68	70.40	70.14	46.66
STATE	71.48	65.43	66.18	67.07	44.89
W2V	55.19	52.98	52.33	50.58	49.63
Top	74.84	70.28	72.12	70.96	58.16
Average	63.52	55.63	59.78	60.41	45.44

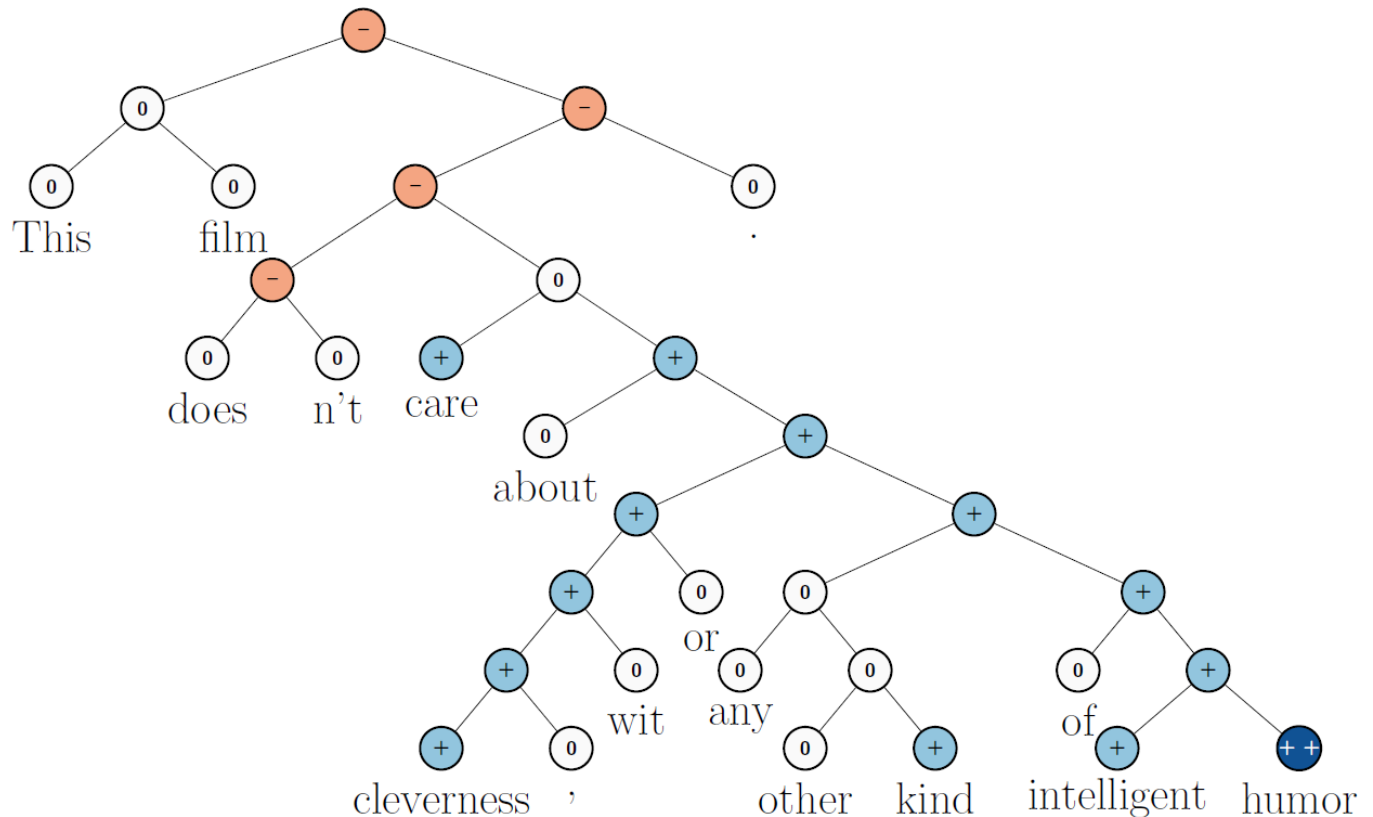
Macro-F scores on five test sets. T1 is LiveJournal2014, T2 is SMS2013, T3 is Twitter2013, T4 is Twitter2014, and T5 is Twitter2014Sarcasm.

Sentiment Composition

- In addition to obtaining sentiment embedding, composing word sentiment to analyze larger pieces of text (e.g., sentences) is another important problem.
- Most work we have discussed so far is based on bag-of-words or bag-of-ngrams assumption.
- More principled models...

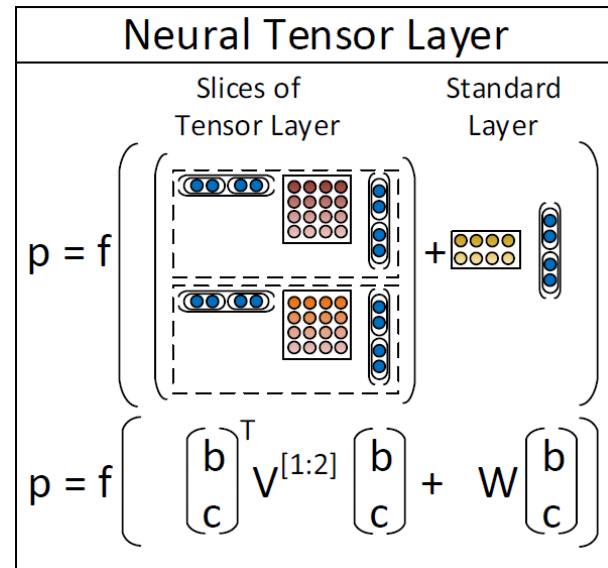
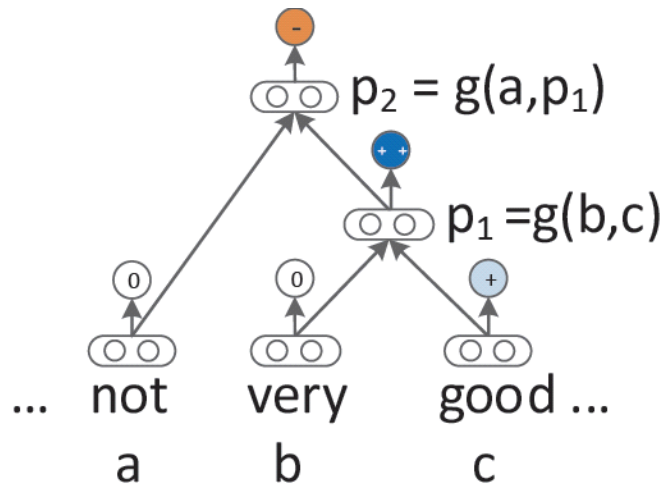
Sentiment Composition

- Socher et al. (2013) proposed a recursive neural network to compose sentiment of a sentence.



Sentiment Composition

- Tensors are critical in capturing interaction between two words/phrases being composed (e.g., a negator and the phrase it modifies.)



- Standard forward/backward propagation was adapted to learn the weights/parameters; main difference lies in the tensor part (V in the figure.)

Results

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	80.7	45.7	87.6	85.4

Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes.

Overview of Sentiment Analysis Systems

- Rule-based systems
- Conventional statistical systems
- Deep-learning-based models



Summary

Summary

- Sentiment analysis subsumes several different problems
 - Important to be aware of the problem pertinent to your task, and the annotation instructions used to create the data
- Sentiment analysis relevant to many domains
 - Not just customer reviews
- Several shared tasks exist
 - Source of data and benchmarks
- Statistical machine learning methods quite popular
- Term-sentiment associations are a key source of information
 - Can obtain this from training data (ngrams)
 - More can be obtained from quasi-annotations, such as from emoticons and hashtags

Summary (continued)

- Other significant features include those from:
 - handling negation
 - handling semantic composition
- Building a competition system involves:
 - careful evaluation of usefulness of features
 - trying various parameters pertaining to both the learning algorithm and the features
 - keep features that obtain improvements consistently on many datasets
 - more labeled data trumps smarter algorithms

Future Directions

- Semantic roles of sentiment
- Sentiment embedding and composition
- Sarcasm, irony, and metaphor
- Sentiment analysis in non-English languages
- Detecting stance, framing, and spin
- Detecting trends and significant changes in sentiment distributions

Future Directions (continued)

- Detecting intensity of sentiment
- Developing better sentiment models for negation, intensifiers, and modality
- Developing better emotion models
- Developing applications for public health, business, social welfare, and for literary and social scientists

SemEval-2015, Sentiment Tasks

- Task 9: **CLIPEval Implicit Polarity of Events**
 - explicit and implicit, pleasant and unpleasant, events
- Task 10: **Sentiment Analysis in Twitter**
 - repeat of 2013 and 2014 task
 - more subtasks
- Task 11: **Sentiment Analysis of Figurative Language in Twitter**
 - metaphoric and ironic tweets
 - intensity of sentiment
- Task 12: **Aspect Based Sentiment Analysis**
 - repeat of 2014 task
 - domain adaptation task

Task 9: CLIPeval Implicit Polarity of Events

- Explicit pleasant event
Yesterday I met a beautiful woman
- Explicit unpleasant event
I ate a bad McRib this week
- Implicit pleasant event
Last night I finished the sewing project
- Implicit unpleasant event
Today, I lost a bet with my grandma

A dataset of first person sentences annotated as instantiations of psychologically grounded pleasant and unpleasant events (MacPhillamy and Lewinsohn 1982):

After that, I started to color my hair and polish my nails.

positive, personal_care

When Swedish security police Saepo arrested me in 2003 I was asked questions about this man.

negative, legal_issue

Task 10: Sentiment Analysis in Twitter

- **Subtask A: Contextual Polarity Disambiguation**
 - Given a message containing a marked instance of a word or phrase, determine whether that instance is positive, negative or neutral in that context.
- **Subtask B: Message Polarity Classification**
 - Given a message, classify whether the message is of positive, negative, or neutral sentiment.
- **Subtask C^{NEW}: Topic-Based Message Polarity Classification**
 - Given a message and a topic, classify whether the message is of positive, negative, or neutral sentiment towards the given topic.
- **Subtask D^{NEW}: Detecting Trends Towards a Topic**
 - Given a set of messages on a given topic from the same period of time, determine whether the dominant sentiment towards the target topic in these messages is (a) strongly positive, (b) weakly positive, (c) neutral, (d) weakly negative, or (e) strongly negative.
- **Subtask E^{NEW}: Determining degree of prior polarity**
 - Given a word or a phrase, provide a score between 0 and 1 that is indicative of its strength of association with positive sentiment.

Task 11: Sentiment Analysis of Figurative Language in Twitter

- **Twitter** is rife with ironic, sarcastic and figurative language.
- How does this creativity impact the perceived affect?
- Do conventional sentiment techniques need special augmentations to cope with this non-literal content?
 - This is not an irony detection task *per se*, but a sentiment analysis task in the presence of irony.
- **Task 11** will test the capability of sentiment systems on a collection of tweets that have a high concentration of sarcasm, irony and metaphor.
 - Tweets are hand-tagged on a sentiment scale ranging from **-5** (*very negative meaning*) to **+5** (*very positive*).

Task 12: Aspect Based Sentiment Analysis

- Subtask 1
 - a set of quintuples has to be extracted from a collection of opinionated documents
 - opinion target
 - target category
 - target polarity
 - “from” and “to” that indicate the opinion target’s starting and ending offset in the text
- Subtask 2
 - same as subtask 1, but on new unseen domain
 - no training data from the target domain

Other Sentiment Challenges

- Kaggle competition on Sentiment Analysis on Movie Reviews
 - website: <http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews>
 - deadline: 11:59 pm, Saturday 28 February 2015 UTC
 - # of teams: 395
 - The sentiment labels are:
 - 0 - negative
 - 1 - somewhat negative
 - 2 - neutral
 - 3 - somewhat positive
 - 4 - positive

Email

Saif M. Mohammad

saif.mohammad@nrc-cnrc.gc.ca

Xiaodan Zhu

xiaodan.zhu@nrc-cnrc.gc.ca

Svetlana Kiritchenko

svetlana.kiritchenko@nrc-cnrc.gc.ca

References

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM'11, pp. 30--38, Portland, Oregon.
- Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012). Textual and contextual patterns for sentiment analysis over microblogs. In Proceedings of the 21st International Conference on World Wide Web Companion, WWW '12 Companion, pp. 453--454, New York, NY, USA.
- Almquist, E., & Lee, J. (2009). What do customers really want? Harvard Business Review.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceeding of the 7th International Conference on Language Resources and Evaluation, Vol. 10 of LREC '10, pp. 2200--2204.
- Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. In Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '12, pp. 11--18, Jeju, Republic of Korea.
- Becker, L., Erhart, G., Skiba, D., & Matula, V. (2013). Avaya: Sentiment analysis on Twitter with self-training and polarity lexicon expansion. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pp. 333{340, Atlanta, Georgia, USA.
- Bellegarda, J. (2010). Emotion analysis using latent aective folding and embedding. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Los Angeles, California.
- Anil Bandhakavi; Nirmalie Wiratunga; Deepak P; Stewart Massie. *Generating a Word-Emotion Lexicon from #Emotional Tweets. SemEval-2014.*
- Boucher, J. D., & Osgood, C. E. (1969). The Pollyanna Hypothesis. Journal of Verbal Learning and Verbal Behaviour, 8, 1--8.

- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanagan, J., & Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '11.
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Tech. rep., Stanford University.
- Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, 143, 29-36.
- Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the 8th Conference of European Chapter of the Association for Computational Linguistics, EACL '97, pp. 174-181, Madrid, Spain.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 168-177, New York, NY, USA. ACM.
- Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, pp. 1827-1830, New York, NY, USA. ACM.
- Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, ACL '11, pp. 151-160.
- Johansson, R., & Moschitti, A. (2013). Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39 (3), 473-509.
- John, D., Boucouvalas, A. C., & Xu, Z. (2006). Representing emotional momentum within expressive internet communication. In Proceedings of the 24th International Conference on Internet and Multimedia Systems and Applications, pp. 183-188, Anaheim, CA. ACTA Press.
- Jurgens, D., Mohammad, S. M., Turney, P., & Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval '12, pp. 356-364, Montreal, Canada. Association for Computational Linguistics.

- Kennedy, A., & Inkpen, D. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. In Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations, Ottawa, Ontario, Canada.
- Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22 (2), 110-125.
- Kiritchenko, S., Zhu, X., Cherry, C., & Mohammad, S. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the International Workshop on Semantic Evaluation, SemEval '14, Dublin, Ireland.
- Kunneman, F. A., C. C. Liebrecht, and A. P. J. van den Bosch. The (Un) Predictability of Emotional Hashtags in Twitter. 2014.
- Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, volume 50, pages 723-762.
- Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The Good the Bad and the OMG!. In Proceedings of the 5th International AAI Conference on Weblogs and Social Media.
- Laponi, E., Read, J., & Ovreid, L. (2012). Representing and resolving negation for sentiment analysis. In Vreeken, J., Ling, C., Zaki, M. J., Siebes, A., Yu, J. X., Goethals, B., Webb, G. I., & Wu, X. (Eds.), *ICDM Workshops*, pp. 687-692. IEEE Computer Society.
- Li, J., Zhou, G., Wang, H., & Zhu, Q. (2010). Learning the scope of negation via shallow semantic parsing. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pp. 671-679, Beijing, China.
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Aggarwal, C. C., & Zhai, C. (Eds.), *Mining Text Data*, pp. 415-463. Springer US.
- Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI '03, pp. 125-132, New York, NY. ACM.
- Louviere, J. J. (1991). Best-worst scaling: A model for the largest difference judgments. Working Paper.
- Raheleh Makki, Stephen Brooks and Evangelos E. Milios, Context-Specific Sentiment Lexicon Expansion via Minimal User Interaction. 2014.

- Mart nez-Camara, E., Mart n-Valdivia, M. T., Ure~nalopez, L. A., & Montejoraez, A. R. (2012). Sentiment analysis in Twitter. *Natural Language Engineering*, 1-28.
- Mihalcea, R., & Liu, H. (2006). A corpus-based approach to nding happiness. In *Pro-ceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 139-144. AAAI Press.
- Mohammad, S. (2012). #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM '12*, pp. 246-255, Montreal, Canada. Association for Computational Linguistics.
- Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic ori-entation lexicons from overtly marked words and a thesaurus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 2, EMNLP '09*, pp. 599-608.
- Mohammad, S. M., & Kiritchenko, S. (2014). Using hashtags to capture ne emotion categories from tweets. To appear in *Computational Intelligence*.
- Mohammad, S. M., Kiritchenko, S., & Martin, J. (2013). Identifying purpose behind elec-toral tweets. In *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '13*, pp. 1-9.
- Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, LA, California.
- Mohammad, S. M., & Yang, T. W. (2011). Tracking sentiment in mail: How genders di er on emotional axes. In *Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '11*, Portland, OR, USA.
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). A ect analysis model: novel rule-based approach to a ect sensing from text. *Natural Language Engineering*, 17, 95-135.
- Orme, B. (2009). Maxdi analysis: Simple counting, individual-level logit, and HB. Saw-tooth Software, Inc.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation, LREC '10*, Valletta, Malta. European Language Resources Association (ELRA).

- Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '05, pp. 115-124.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2 (1-2), 1-135.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '02, pp. 79-86, Philadelphia, PA.
- Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. In Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series).
- Porter, M. (1980). An algorithm for suffix stripping. Program, 3, 130-137.
- Proisl, T., Greiner, P., Evert, S., & Kabashi, B. (2013). Klue: Simple and robust methods for polarity classification. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pp. 395-401, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Reckman, H., Baird, C., Crawford, J., Crowell, R., Micciulla, L., Sethi, S., & Veress, F. (2013). teragram: Rule-based detection of sentiment phrases using sentiment analysis. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pp. 513-519, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani (2014). SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter. In Proceeding of the 11th Extended Semantic Web Conference (ESWC), Crete, Greece.
- Sauper, C., & Barzilay, R. (2013). Automatic aggregation by joint modeling of aspects and values. Journal of Artificial Intelligence Research, 46, 89-127.
- Sheng Huang and Zhendong Niu and Chongyang Shi. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems, 2014, pp. 191-200.
- Shi Feng, Kaisong Song, Daling Wang, Ge Yu. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. World Wide Web. Pg 1-19. 2014.

- Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '12. Association for Computational Linguistics.
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '13. Association for Computational Linguistics.
- Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., & associates (1966). The General Inquirer: A Computer Approach to Content Analysis. The MIT Press.
- Taboada, M., Brooke, J., To loski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37 (2), 267-307.
- Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, Ting Liu. Building Large-Scale Twitter-Specific Sentiment Lexicon : A Representation Learning Approach, COLING 2014.
- Duyu Tang, Furu Wei, Nan Yang, Bing Qin, Ting Liu, Ming Zhou: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 208–212, Dublin, Ireland, August 23-24, 2014.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62 (2), 406-418.
- Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning, pp. 491-502, Freiburg, Germany.
- Turney, P., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21 (4).
- Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP '10, pp. 60-68, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., & Ritter, A. (2013). SemEval-2013 Task 2: Sentiment analysis in Twitter. In Proceedings of the International Workshop on Semantic Evaluation, SemEval '13, Atlanta, Georgia, USA.

- Wilson, T., Wiebe, J., & Ho mann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pp. 347-354, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yang, B., & Cardie, C. (2013). Joint inference for fine-grained opinion extraction. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '13.
- Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING '00, pp. 947-953, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Xiaodan Zhu, Hongyu Guo, Saif Mohammad and Svetlana Kiritchenko. An Empirical Study on the Effect of Negation Words on Sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014, Baltimore, MD.

SemEval Proceedings

- SemEval-2013 proceedings are available here:
http://www.aclweb.org/anthology/siglex.html#2013_1
- SemEval-2014 proceedings are available here:
<http://alt.qcri.org/semeval2014/cdrom/>