NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu
National Research Council Canada

SemEval-2013, Task 2

- Is a given message positive, negative, or neutral?
  - tweet or SMS
- Is a given term within a message positive, negative, or neutral?

International competition on sentiment analysis of tweets:

- SemEval-2013 (co-located with NAACL-2013)
- 44 teams

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.

Sentiments in Tweets

Examples at Message Level (Task B)

Tweet: The new Star Trek is spectacular. #InNY #ILoveMovies
target is positive

Tweet: The new Star Trek has no story. #dumbmovie
target is negative

Tweet: Spock displays emotions in the new Star Trek.
target is neutral

Sentiments in Tweets

Examples at Term Level (Task A)

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

Applications of Sentiment Analysis

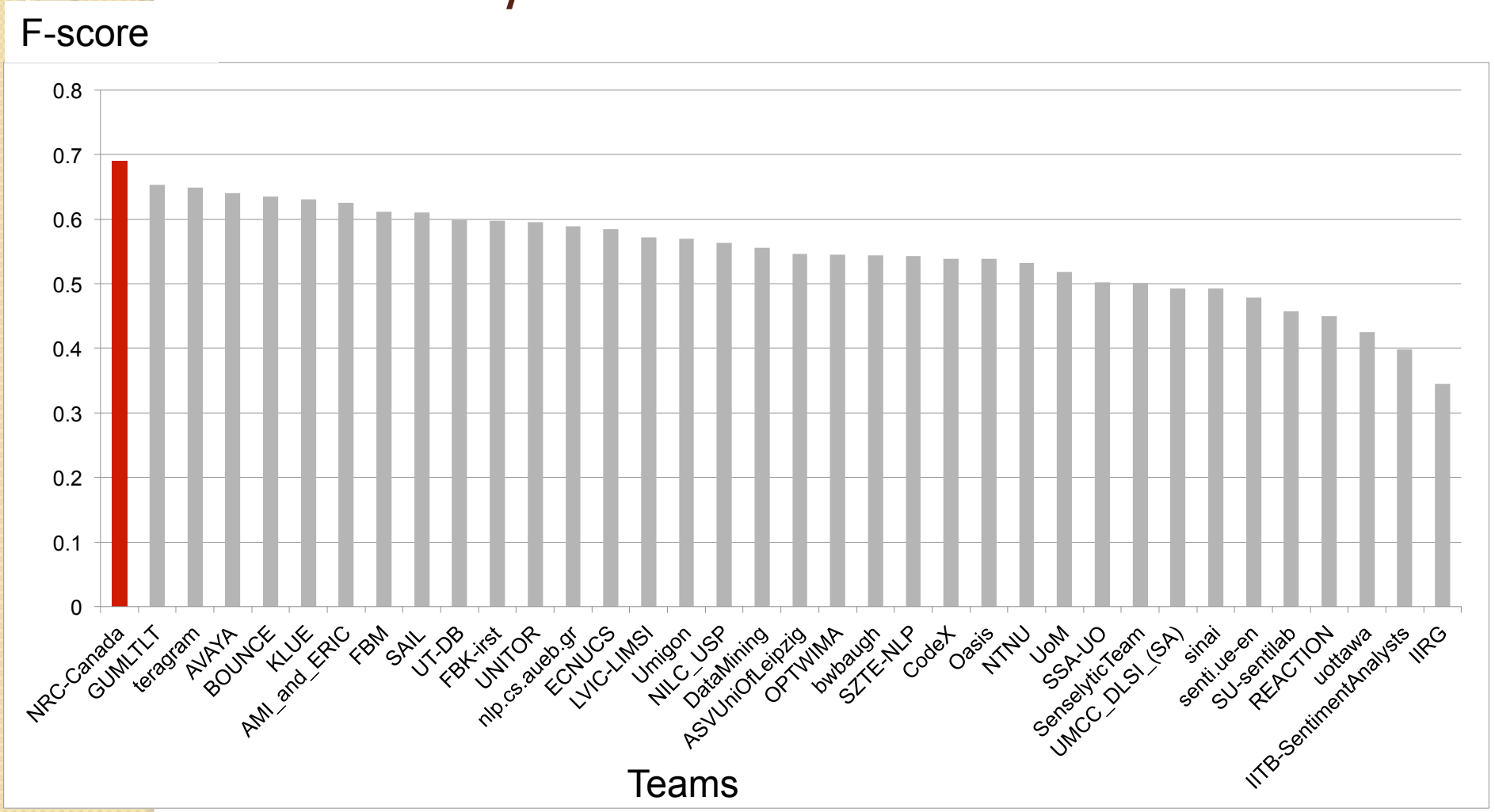
- Tracking sentiment towards politicians, movies, products
- Improving customer relation models
- Identifying what evokes strong emotions in people
- Detecting personality
- Detecting happiness and well-being
- Measuring the impact of activist movements through text generated in social media.
- Improving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

Challenges

- Sentiment may not be explicitly stated
  - Need world knowledge and context
- No tone, pitch, or other prosodic information
- Text may have sarcasm, exaggeration, etc

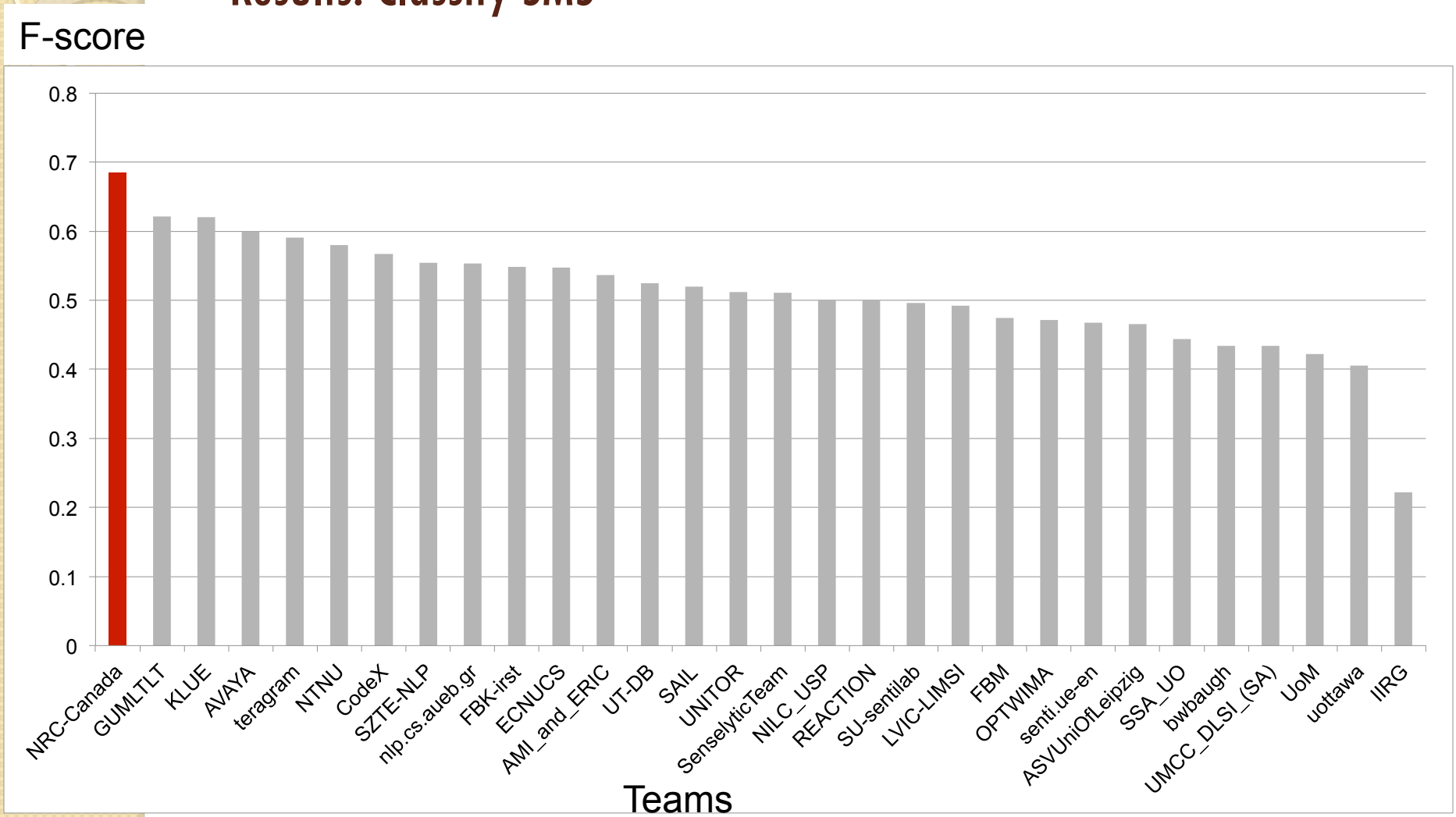
Sentiment Analysis Competition

Results: Classify Tweets



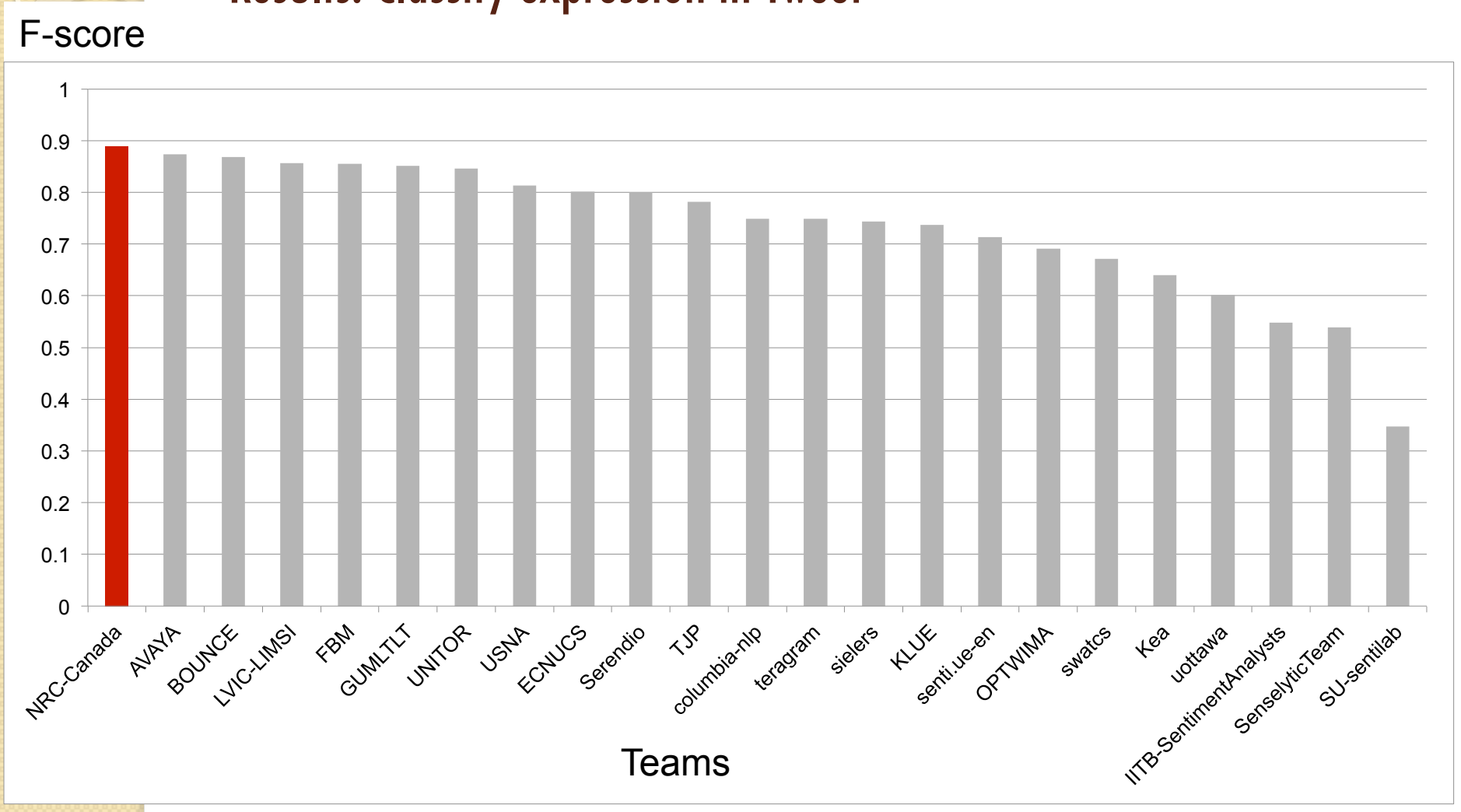
Sentiment Analysis Competition

Results: Classify SMS



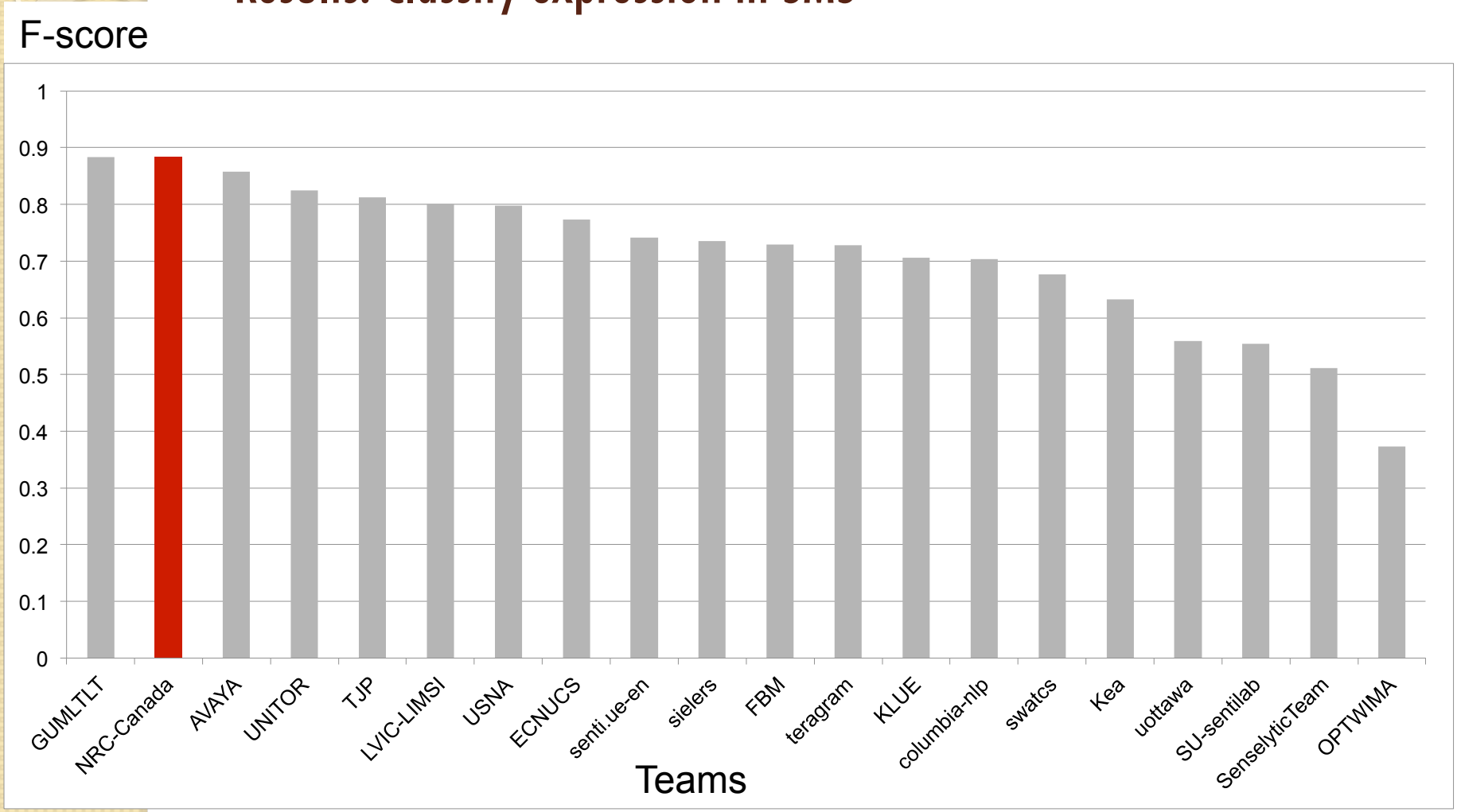
Sentiment Analysis Competition

Results: Classify expression in Tweet



Sentiment Analysis Competition

Results: Classify expression in SMS



Datasets for the Message-Level and Term-Level Tasks

Dataset	Positive	Negative	Neutral	Total
Tweets				
Message-level task:				
Train	3,045 (37%)	1,209 (15%)	4,004 (48%)	8,258
Dev	575 (35%)	340 (20%)	739 (45%)	1,654
Test	1,572 (41%)	601 (16%)	1,640 (43%)	3,813
Term-level task:				
Train	4,831 (62%)	2,540 (33%)	385 (5%)	7,756
Dev	648 (57%)	430 (38%)	57 (5%)	1,135
Test	2,734 (62%)	1,541 (35%)	160 (3%)	4,435
SMS				
Message-level task:				
Test	492 (23%)	394 (19%)	1,208 (58%)	2,094
Term-level task:				
Test	1,071 (46%)	1,104 (47%)	159 (7%)	2,334

Sentiment Lexicons

- Lists of word--sentiment pairs, with scores indicating the degree of association

spectacular positive 0.91	spectacular 0.91
okay positive 0.3	okay 0.3
lousy negative 0.84	lousy -0.84
unpredictable negative 0.17	unpredictable -0.17

- Manually created
  - NRC Emotion Lexicon (Mohammad and Turney, 2010): ~14,000 words
  - MPQA Lexicon (Wilson et al., 2005): ~8,000 words
  - Bing Liu Lexicon (Hu and Liu, 2004): ~6,800 words

Automatically Generated New Lexicons

- Hashtagged emotion words are good labels of emotions in tweets (Mohammad, 2012)

  That jerk stole my photo on Tumblr #grrrr #anger

- Created a list of seed sentiment words by looking up synonyms of *excellent, good, bad*, and *terrible*:
  - 32 positive words
  - 36 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

- A tweet is considered:
  - positive if it has a positive hashtag
  - negative if it has a negative hashtag

- For every word w in the set of 775,000 tweets, an association score is generated:

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

PMI = pointwise mutual information

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

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

NRC Hashtag Sentiment Lexicon

- w can be:
  - any unigram in the tweets: 54,129 entries
  - any bigram in the tweets: 316,531 entries
  - non-contiguous pairs (any two words) from the same tweet: 308,808 entries

- Multi-word entries incorporate context:

  unpredictable story 0.4

  unpredictable steering -0.7

Using sentiment lexicons was defined to be constrained.

Sentiment140 Lexicon

- Go et al. (2009) collected 1.6 million tweets with emoticons
  - Tweets with :) are considered positive
  - Tweets with :(are considered negative

- Created a sentiment lexicon from this corpus using the same PMI method
  - 62,648 unigrams
  - 677,698 bigrams
  - 480,010 non-contiguous pairs

NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon are available for download: www.purl.com/net/sentimentoftweets

MESSAGE-LEVEL TASK

(TASK B)

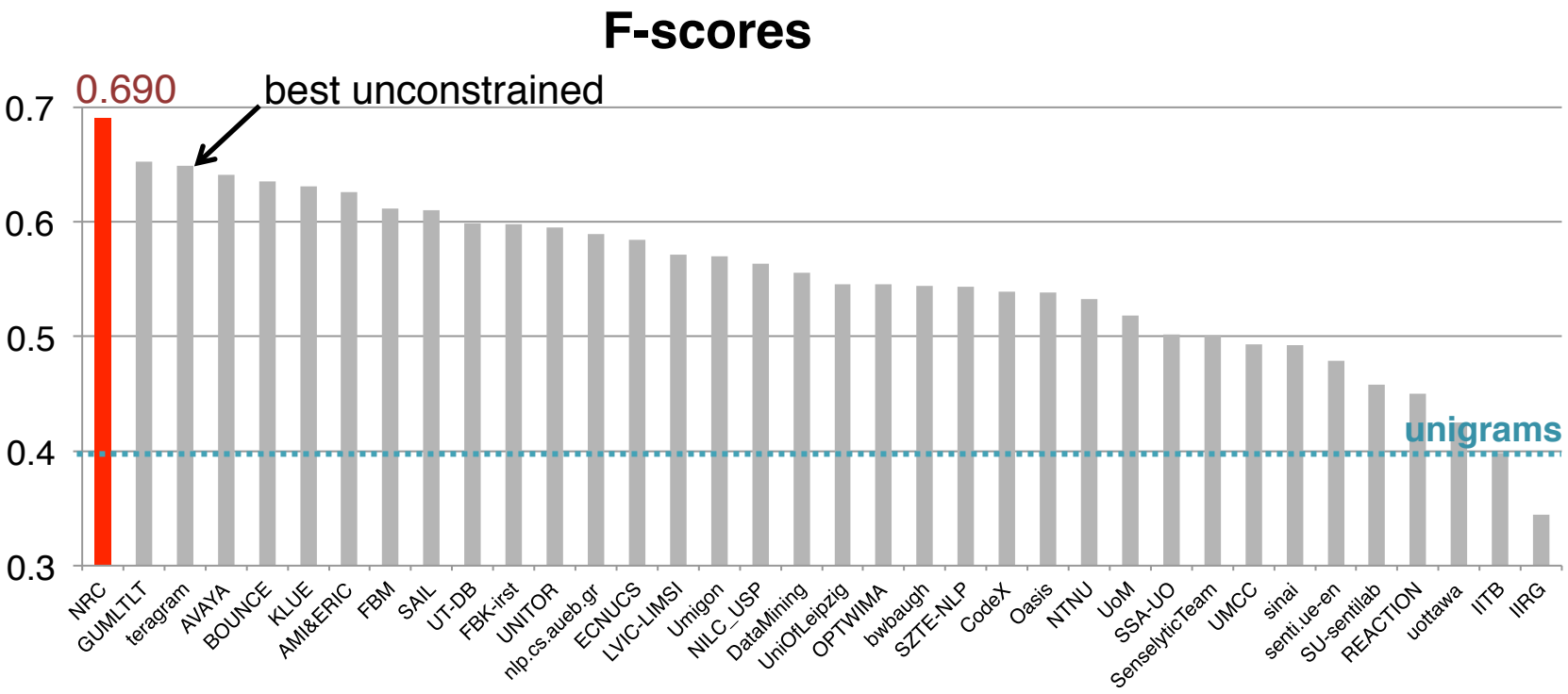
Message-Level Task

- **Pre-processing:**
  - URL -> http://someurl
  - UserID -> @someuser
  - Tokenization and part-of-speech (POS) tagging (CMU Twitter NLP tool)
- **Classifier:**
  - SVM with linear kernel
- **Evaluation:**
  - Macro-averaged F-pos and F-neg

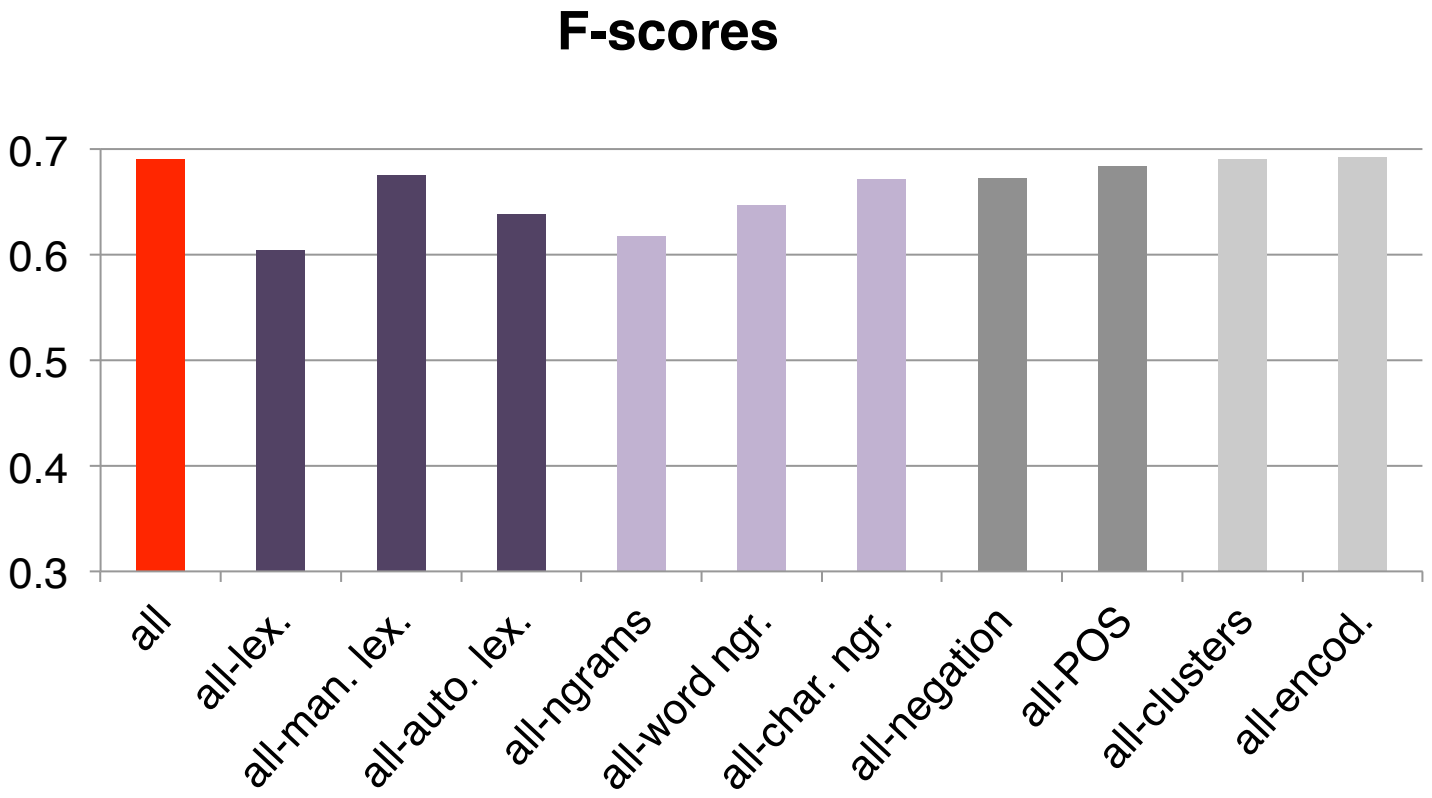
Features

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
word clusters	probably, definitely, probly
emoticons	:D, >:(
elongated words	soooo, yaayyy

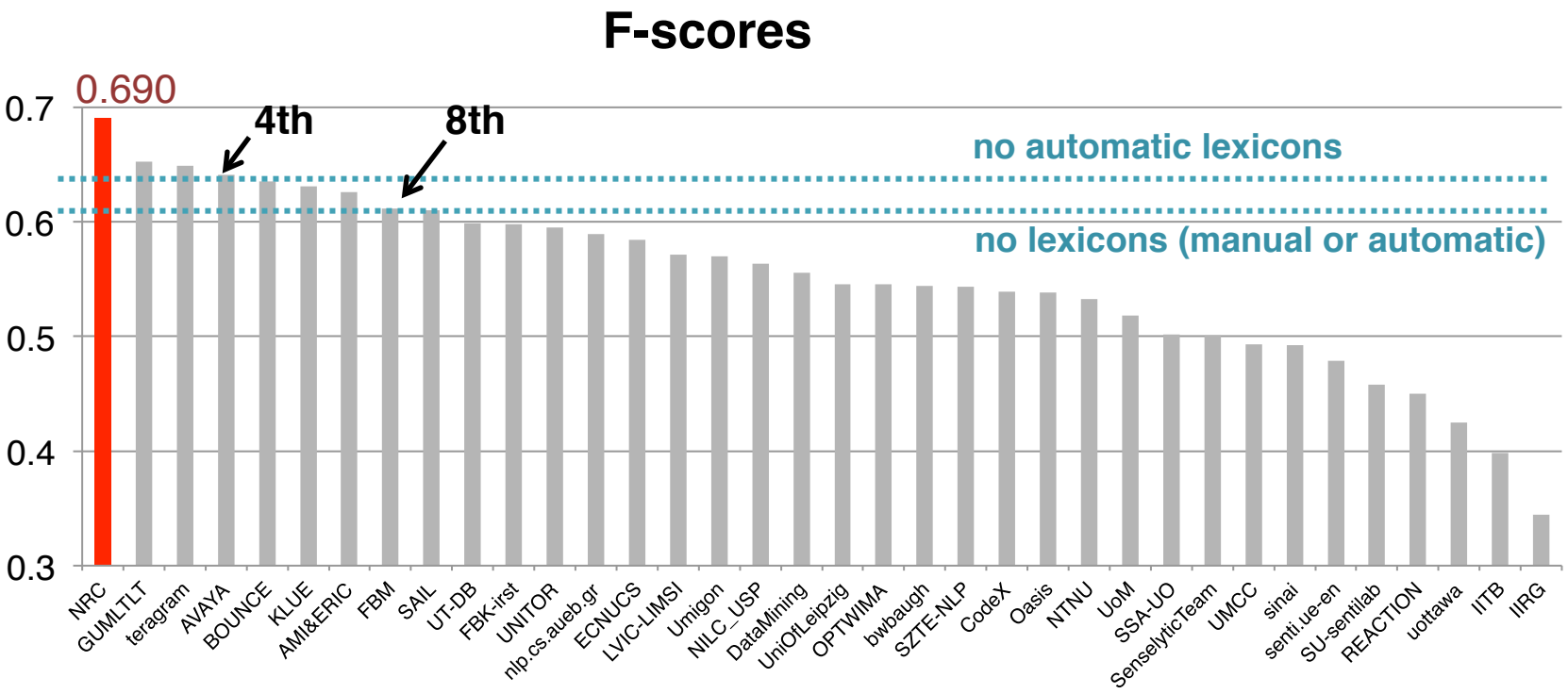
Results on Tweets



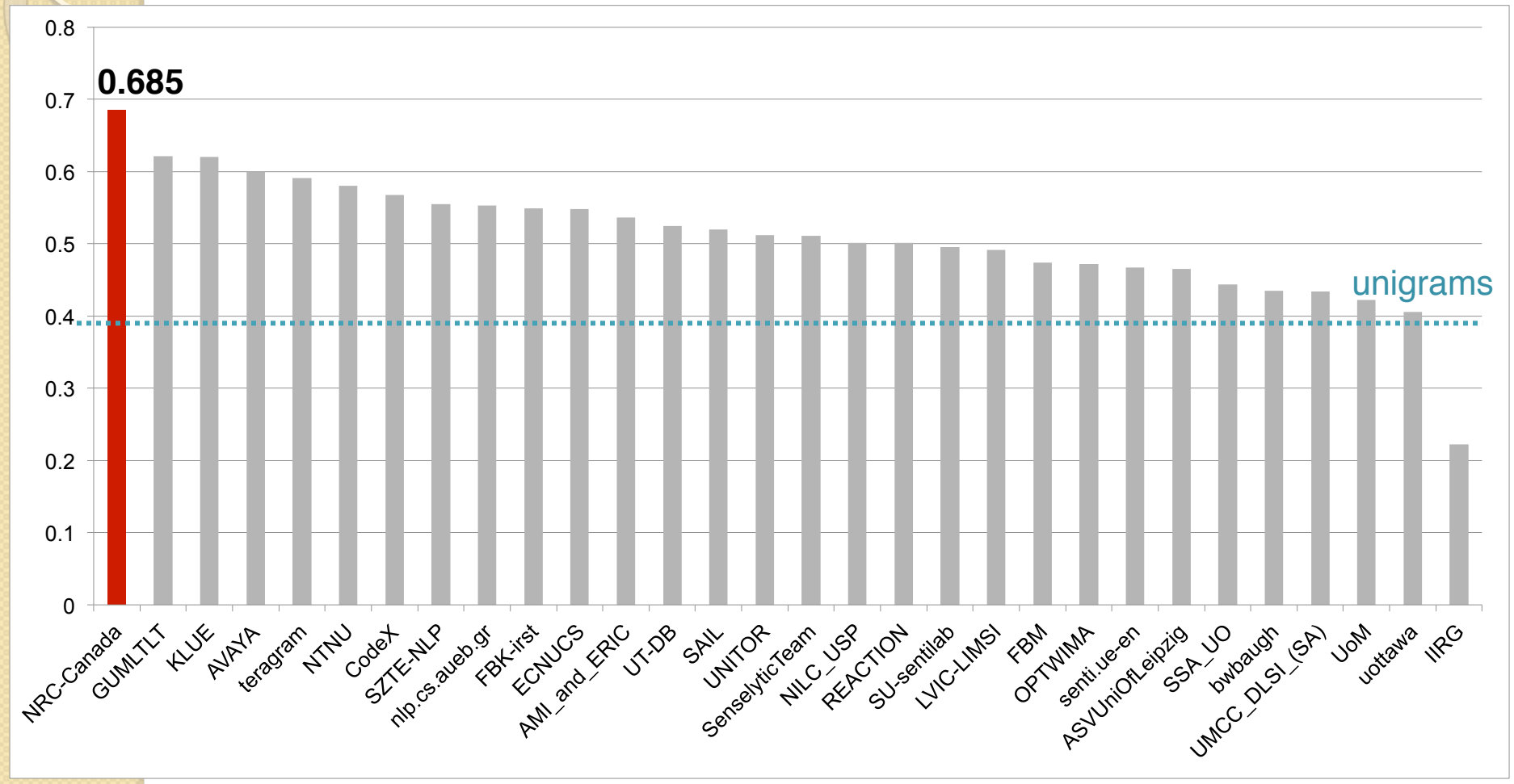
Feature Contributions (on Tweets)



Results on Tweets



Results on SMS



Error Analysis

- **Ambiguous messages** (~ 40%)

  Trey Burke has been suspended for the Northern Michigan game (exhibition) tomorrow. http://t.co/oefkAElW
  system's label: neutral

- **Human errors** (~ 40%)

  Going to Helsinki tomorrow or on the day after tomorrow, yay!
  system's label: positive

- **System's errors** (~20%)
  - Unseen words and expressions (~ 5%)

    2Day in 1999 Mario Lemieux's ownership grp takes over the Penguins. 1st player in modern era to buy the team he played 4. Le magnifique! system's label: neutral
  - Other (~15%)

    NJEA Teacher's Convention_ Nov 8th & 9th in Atlantic City has been cancelled for the 1st time in its 158-year history.
    system's label: neutral

TERM-LEVEL TASK

(TASK A)

A Recap of the Problem:

Examples at Term Level (Task A)

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

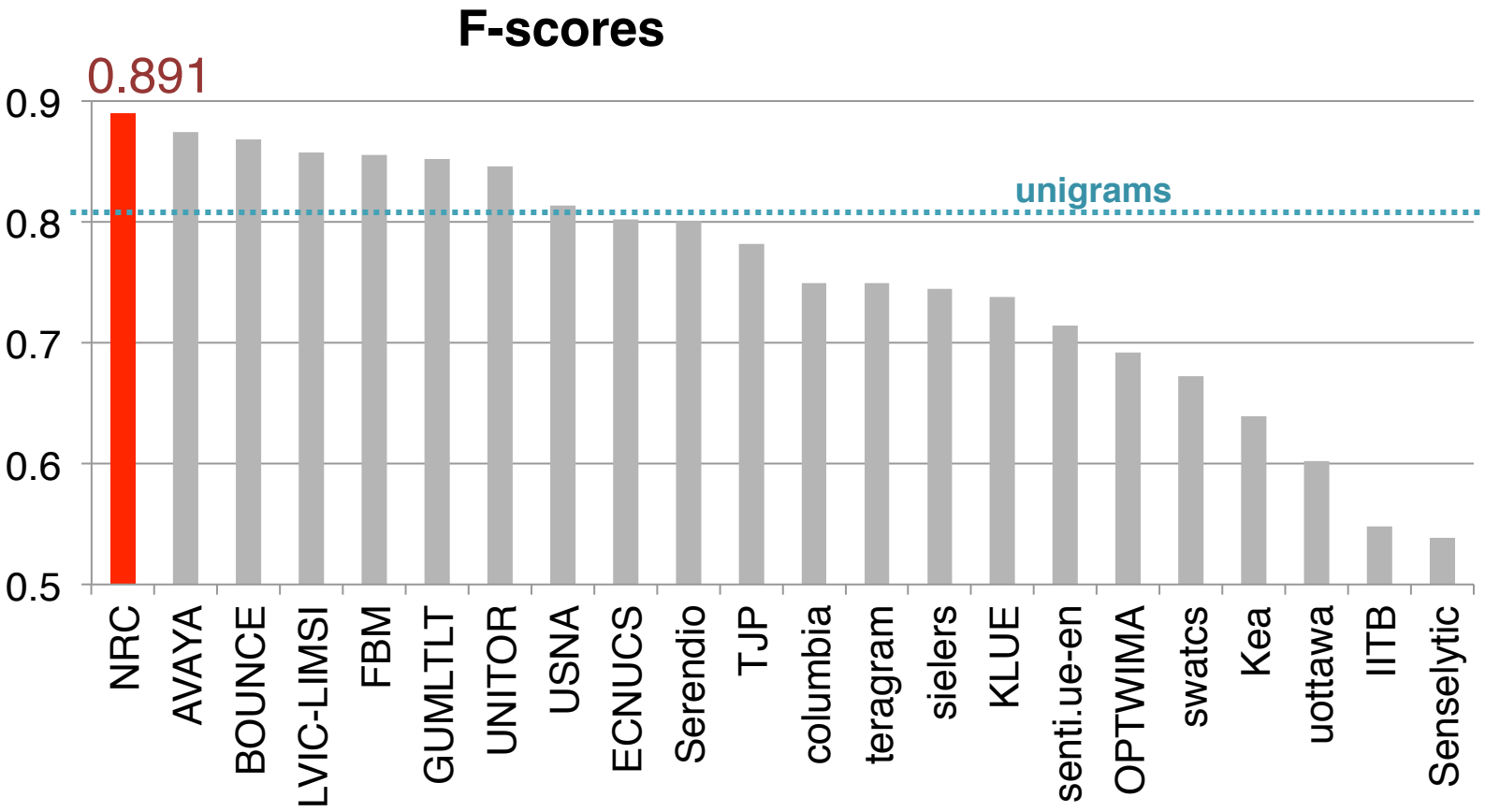
Tweet: The new Star Trek does not have much of a story, but it is visually spectacular.

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

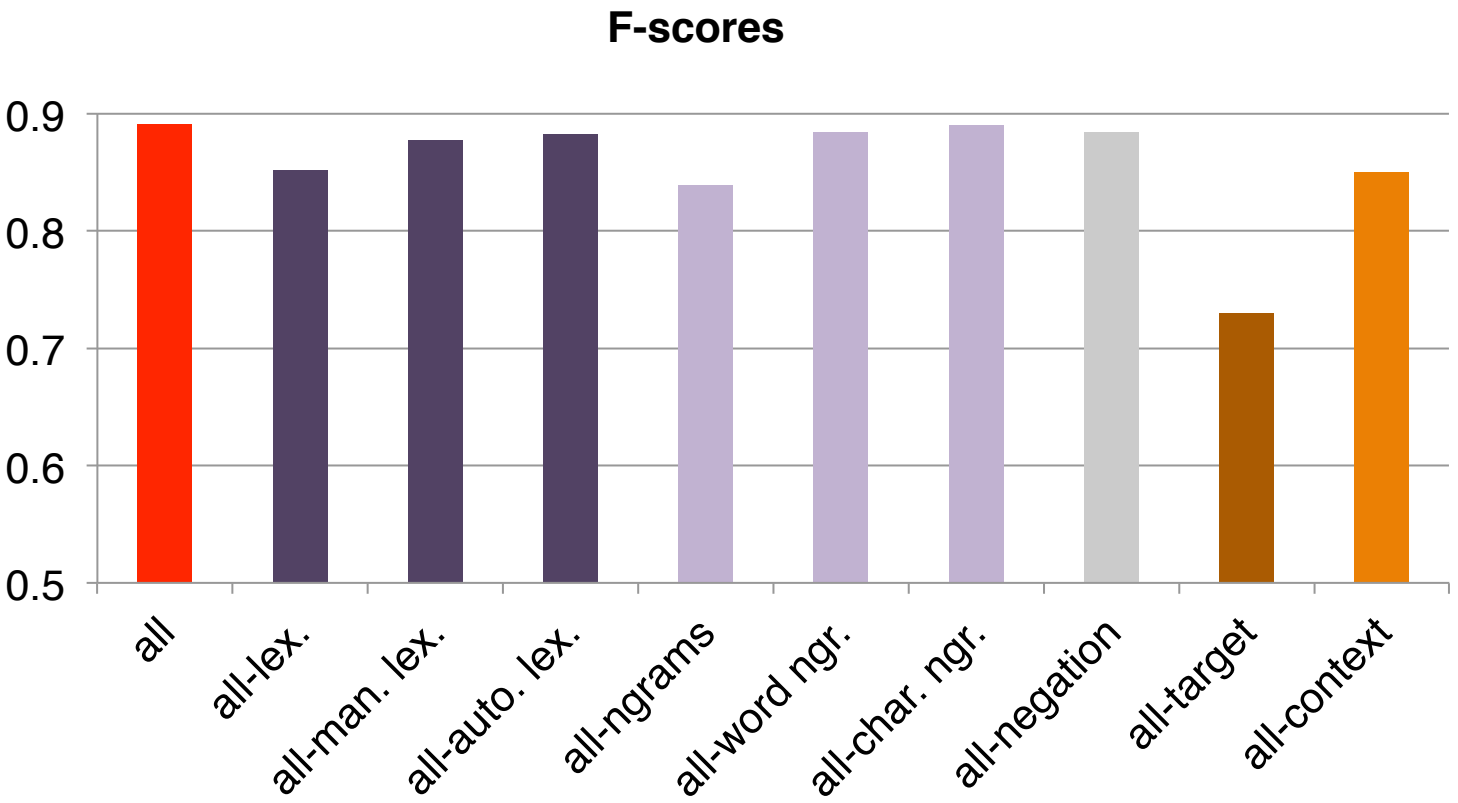
Our Features: From Another Viewpoint

Features	Description
term features	extracted from the target terms
context features	extracted from a window of words around a target term

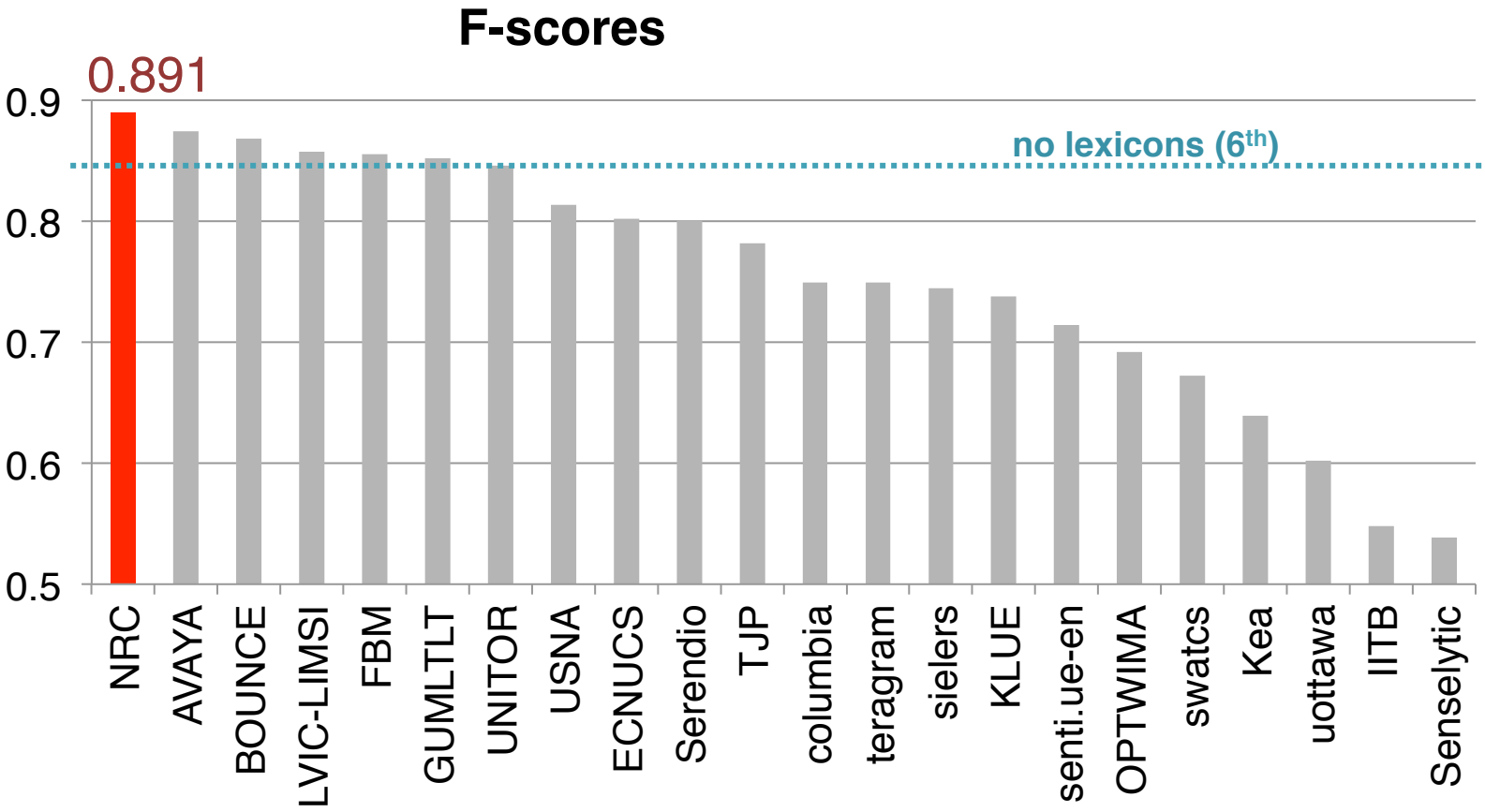
Results on Tweets



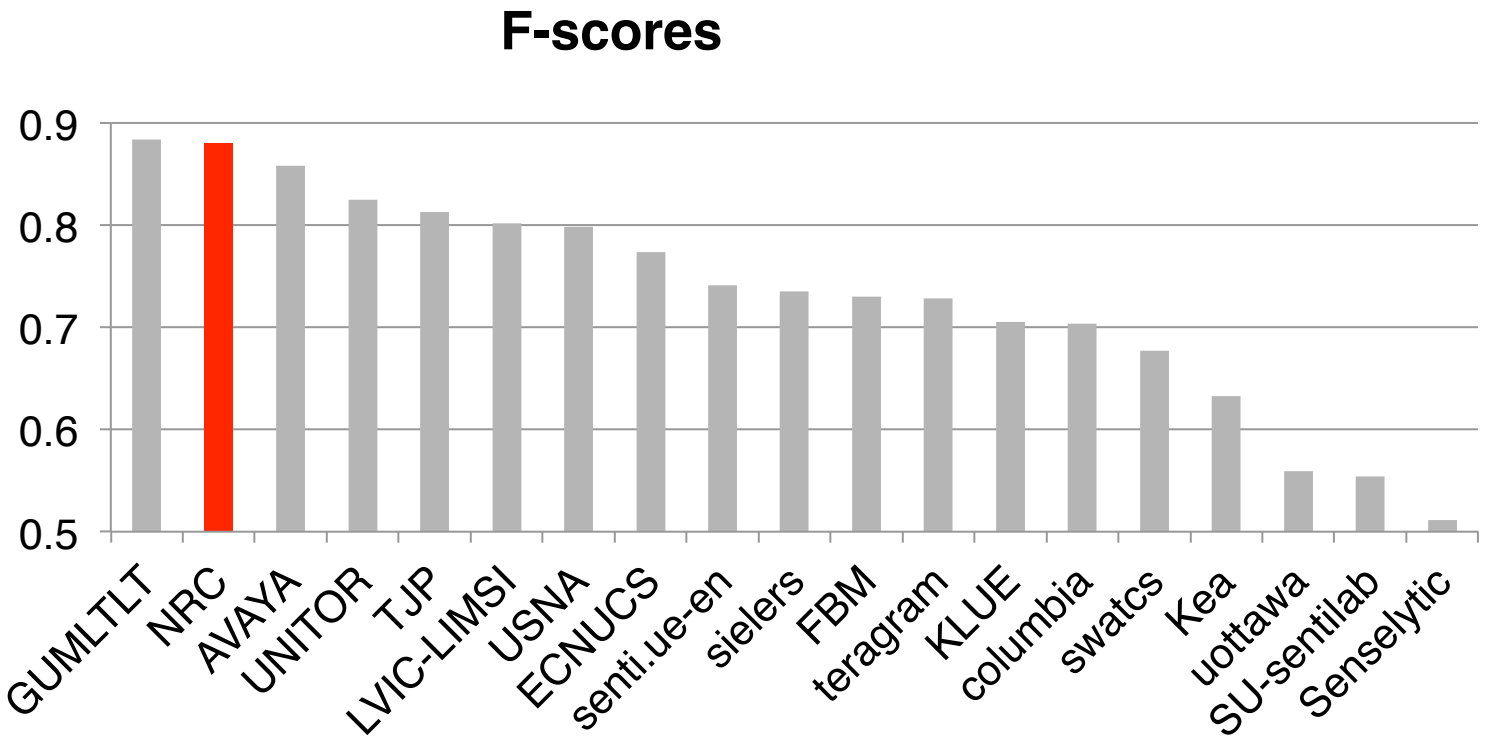
Ablation Experiments on Tweets



Results on Tweets



Results on SMS



Error Analysis

- **Pseudo errors: ~10-15%**
  - bad annotations by Turkers
- **Examples of errors made by our system**

  Blazer game on the 16th. The bulls [smash] the blazers.

  @Hannah_Sunder: The Walking Dead is just a great tv show its [bad ass] just started to watch the 2nd season to catch up with the 3rd

  [Holy shit] no class till Monday I love you iona!!!!!!

Discussion

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

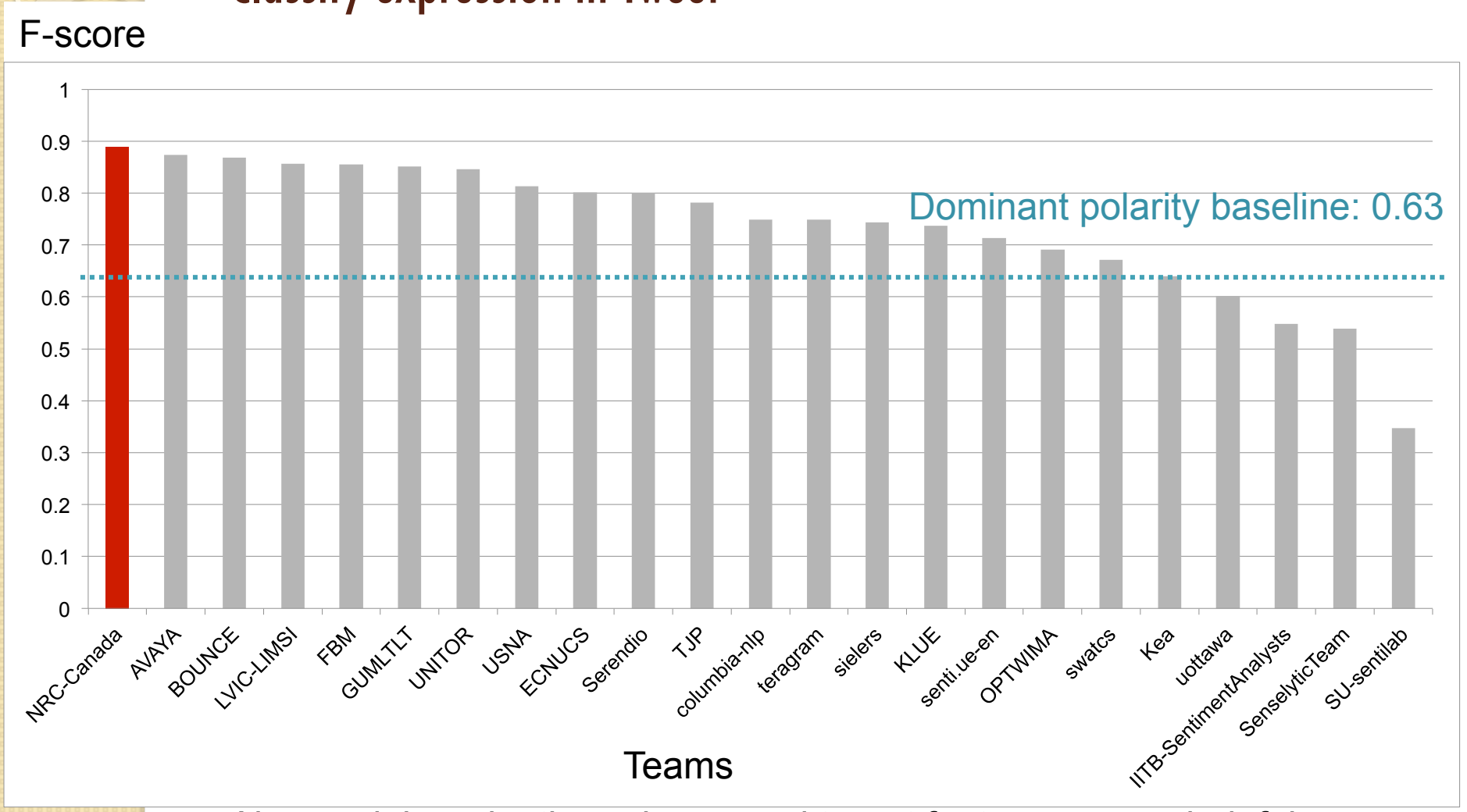
What does this mean?

- Is it harder for humans to determine sentiment of whole message?
- Does the task set-up favors the term-level task?
  - About 60% to 85% of the unigram, bigram, and other ngram target terms seen in training data
  - About 80% of the instances of a word in the training and test data have the same polarity

Simply guessing the dominant polarity of a term in the training data goes a long way.

Official, Competition Results

Classify expression in Tweet



Also explains why the unigram and target features are so helpful.

Summary

- Created SVM classifiers for term- and message-level sentiment detection
- Used an array of features
  - sentiment lexicons, word and character ngrams, pos, negation, punctuations, emoticons, spelling variations
  - generated sentiment lexicons from tweets using hashtags
    - two-, three-, and four-word entries incorporated context
- Official rankings:
  - message-level task: 1st on both tweets and SMS data
  - term-level task: 1st on tweets, 2nd on the SMS data

NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon are available for download: www.purl.com/net/sentimentoftweets