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

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

In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), June 2013, Atlanta, USA.

Four days of geo-tagged Tweets.

Sentiment Analysis of Term in Context: Task A

What is the polarity of the target: positive, negative, or neutral?

Tweet: The movie has no story, but it is visually spectacular.
target is positive

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

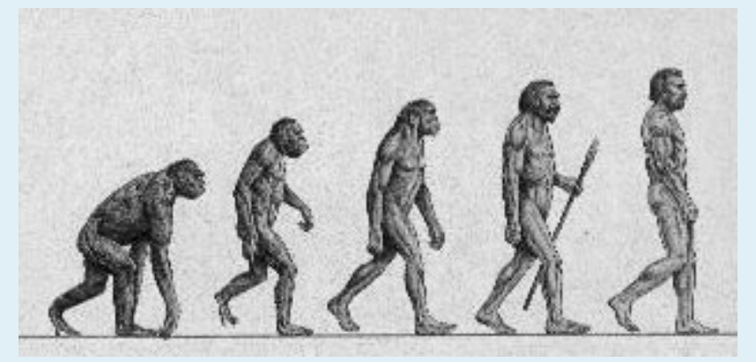
Tweet: The NatGeo documentary on early human evolution was fascinating.
target is neutral

Supervised Machine Learning Classifier

Sentiment Analysis of Message: Task B

What is the polarity of the message: positive, negative, or neutral?

Tweet: The movie is visually spectacular.
target is positive



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

Tweet: The NatGeo documentary on early human evolution was at 7pm.
target is neutral

Features

Features Used for Task A and B

sentiment lexicon	#positive: 3, scoreP: 2.2
word n-grams	spectacular, like documentary
char n-grams	spect, docu, visua
part of speech	#N: 5, #V: 2, #A:1
negation	#Neg: 1;
word clusters	probably, definitely, def
all-caps	YES, COOL
punctuation	#!+: 1, #?+: 0, #!?+: 0
emoticons	:D, >:(
hashtags	#excited, #NowPlaying
elongated words	soooo, yaayyy

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

Existing, Manual Sentiment Lexicons

- NRC Emotion Lexicon (Mohammad, 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
- Polled the Twitter API for tweets with hashtags
  - 32 positive and 36 negative seed words
  - A set of 775,000 tweets was compiled
- For every term t a score is generated:

  $$score(t) = PMI(t, positive) - PMI(t, negative)$$
  - If $score(t) > 0$, then w is positive
  - If $score(t) < 0$, then w word is negative

NRC Hashtag Sentiment Lexicon

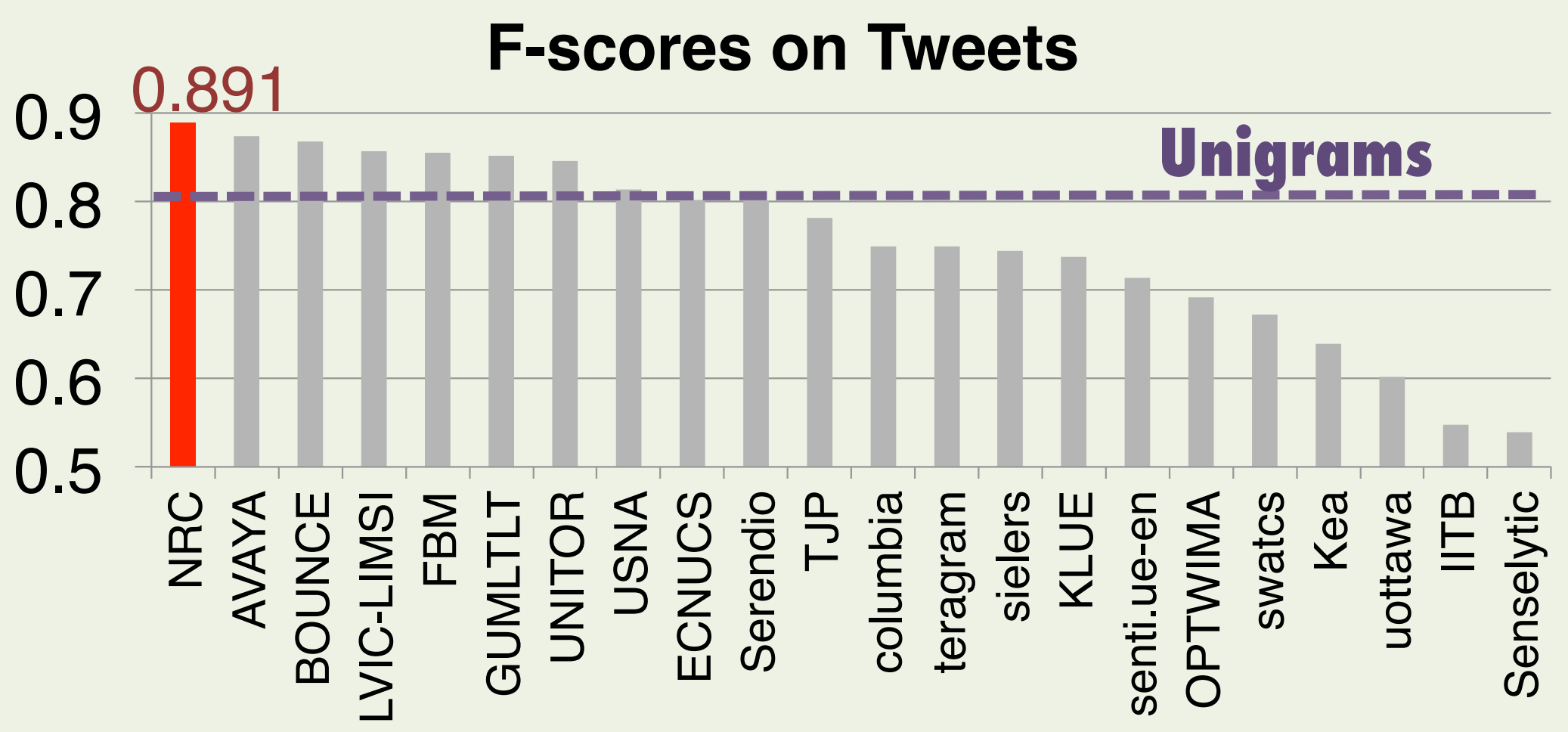
- 54,129 unigrams
- 316,531 bigrams
- 308,808 pairs of unigrams and bigrams

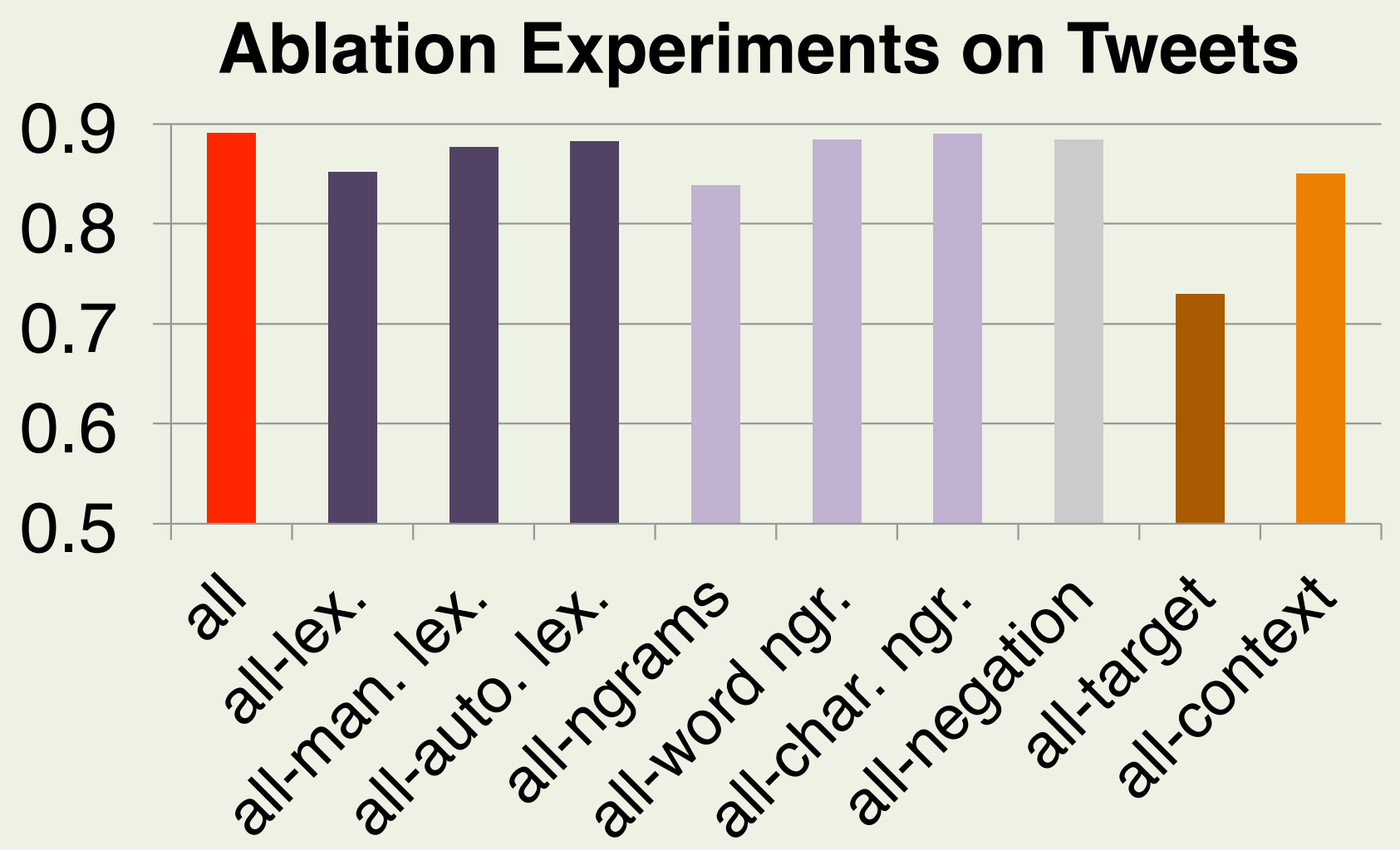
Sentiment140 Lexicon

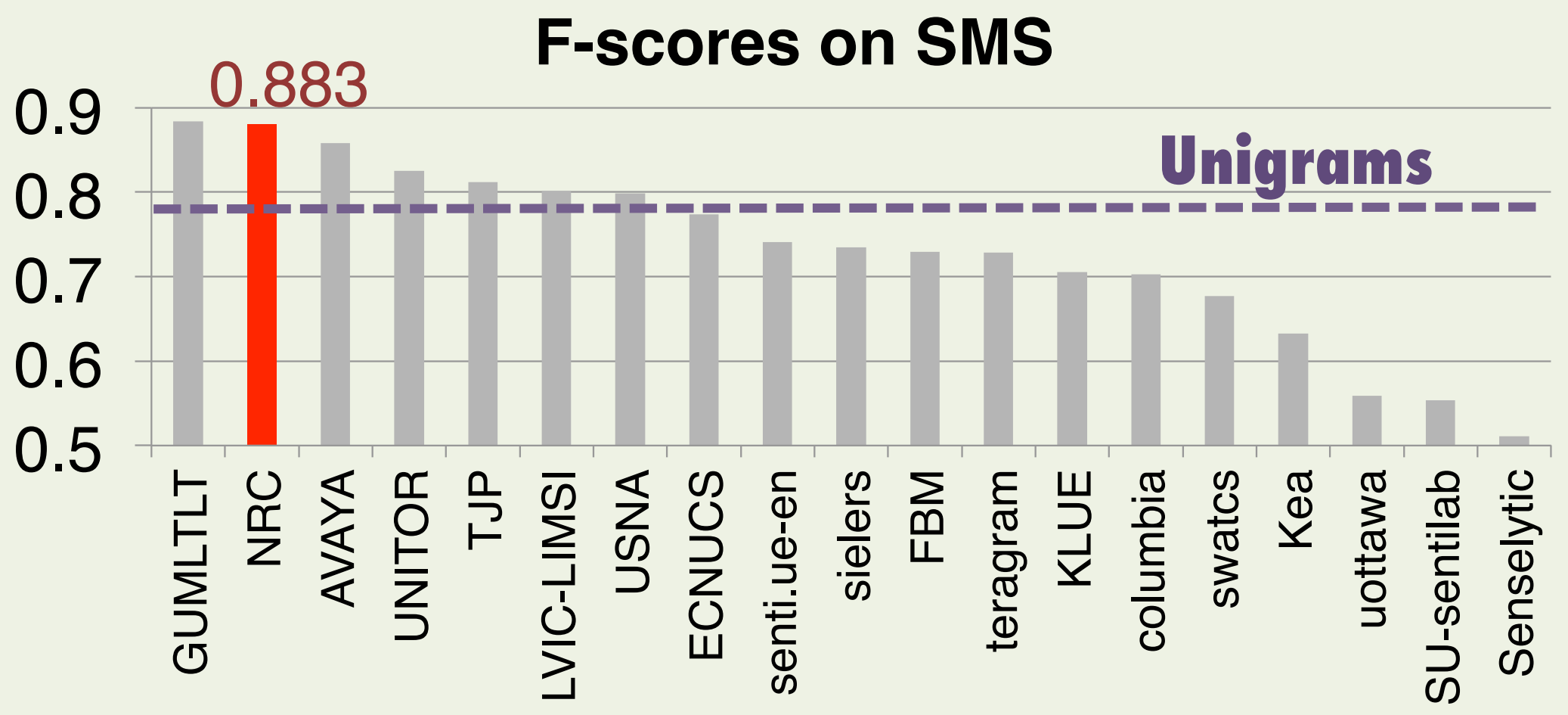
- 62,648 unigrams
- 677,698 bigrams
- 480,010 pairs of unigrams and bigrams

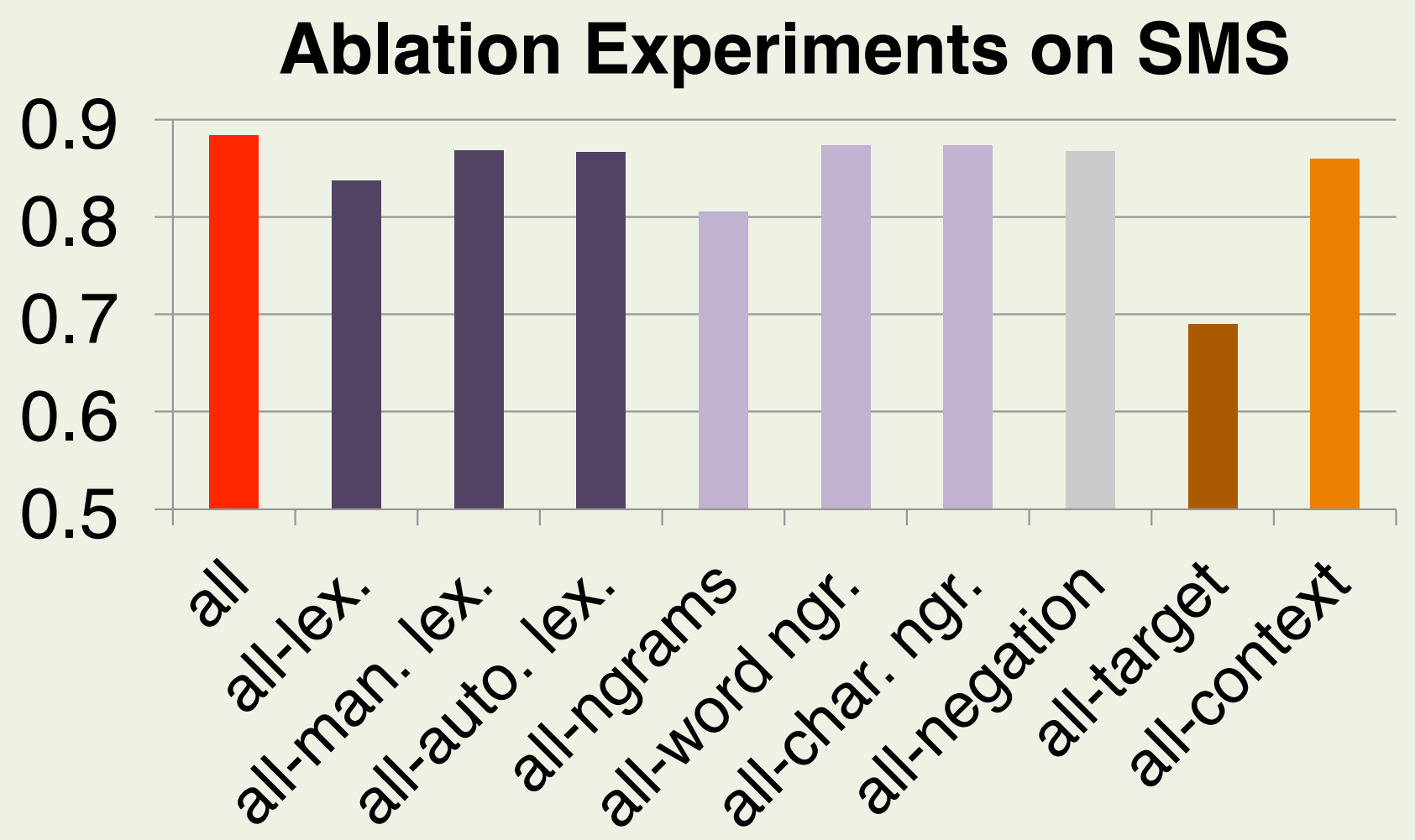
NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon available for download: www.purl.com/net/sentimentoftweets

Results: Task A





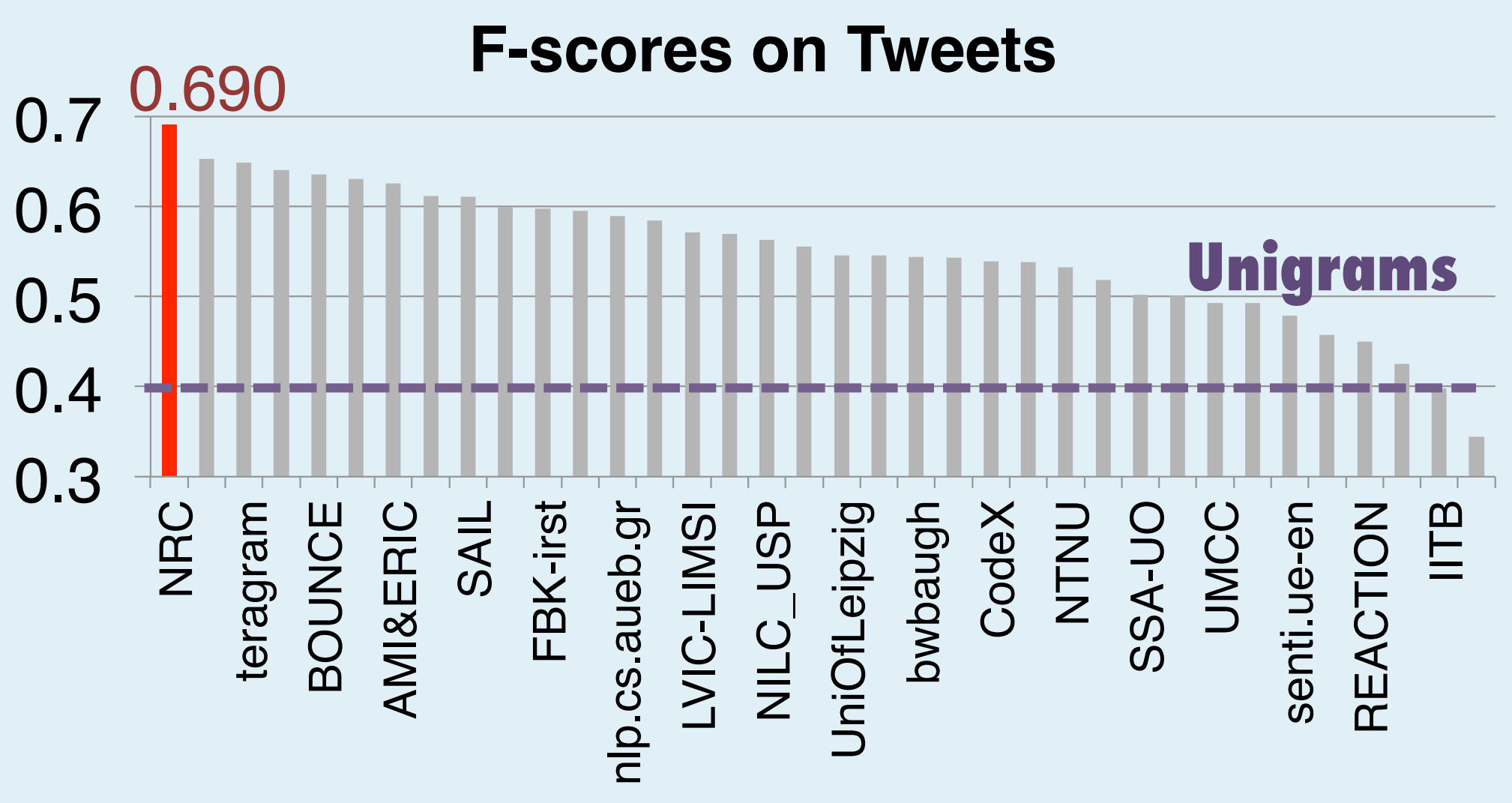


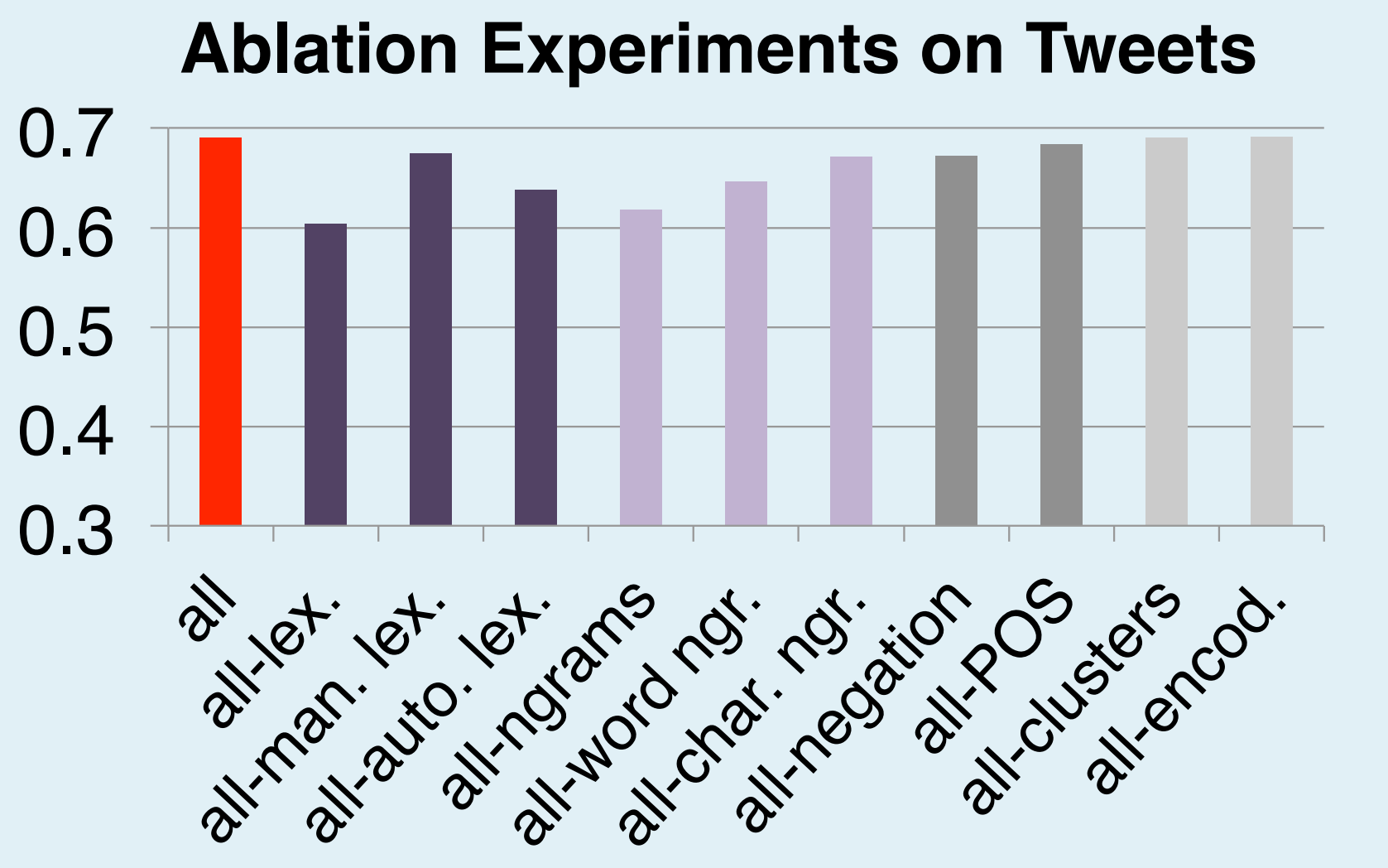


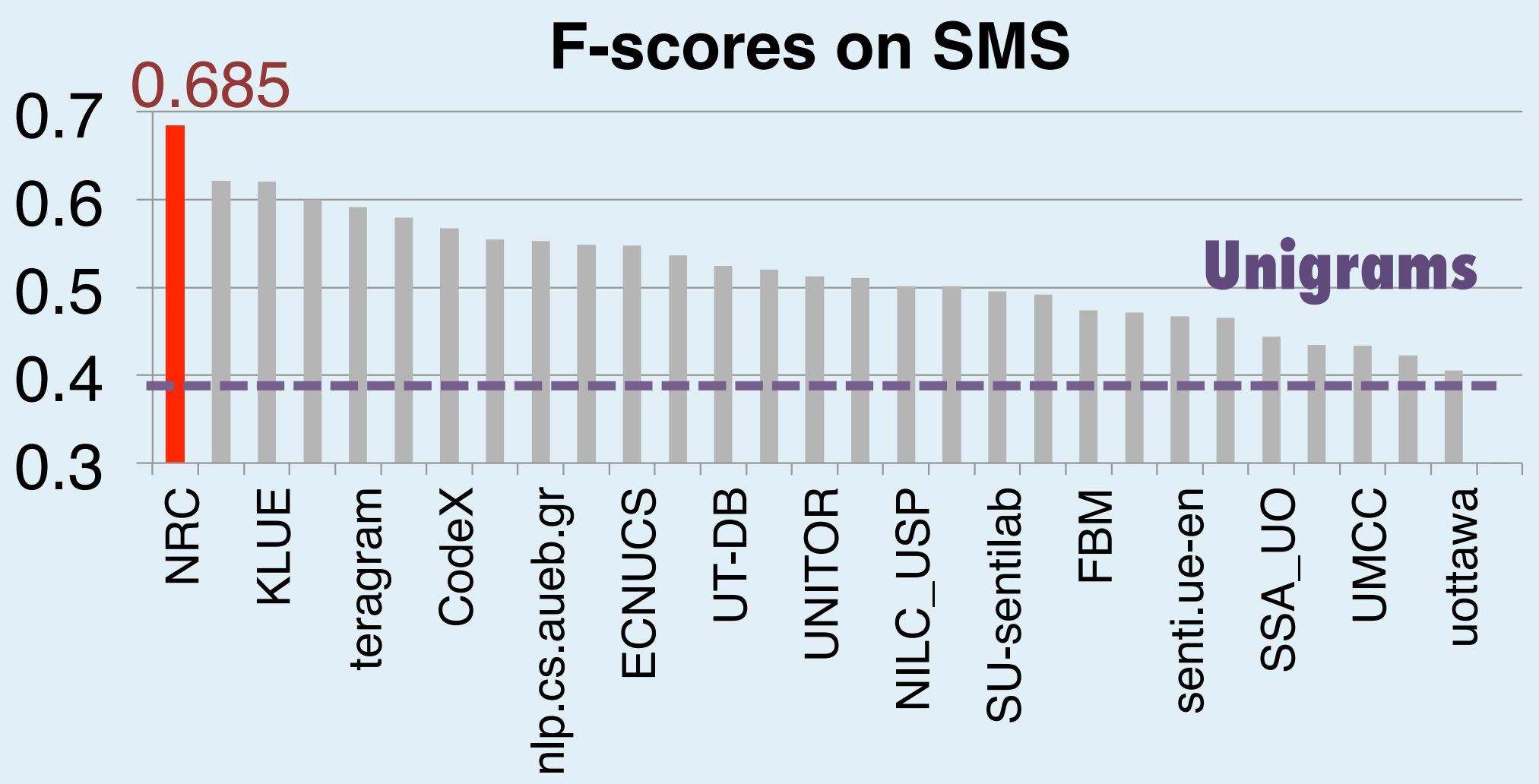
Conclusions

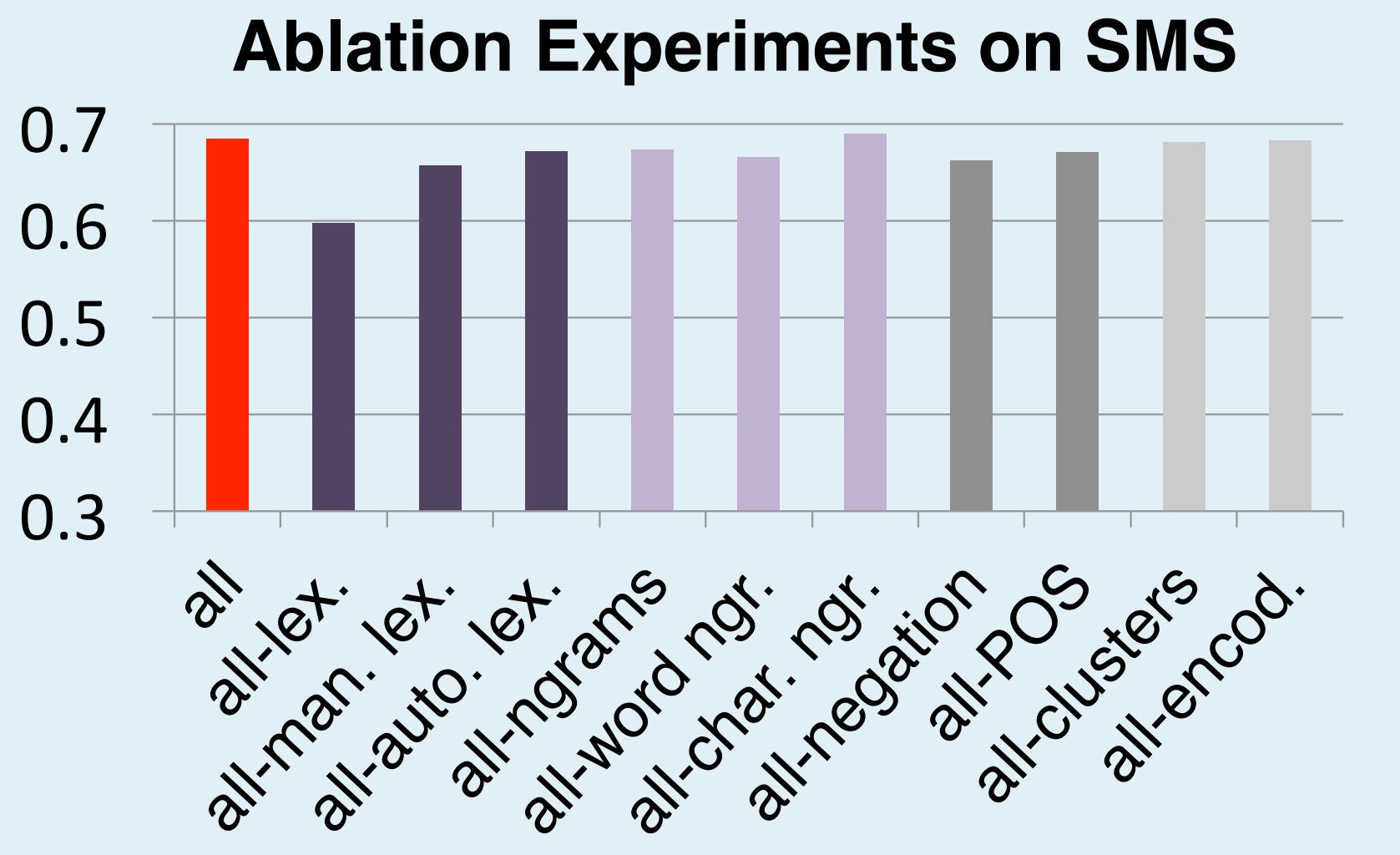
- Built state-of-the art sentiment analysis system using SVM and lexical features
- Generated sentiment lexicons from tweets using hashtags
  - two-, three-, and four-word entries incorporated context
- Most useful features
  - sentiment lexicons
  - ngrams
- SMS results similar to tweets

Results: Task B


F-scores on Tweets
Unigrams
NRC
teragram
BOUNCE
AMI&ERIC
SAIL
FBK-irst
nlp.cs.aueb.gr
LVIC-LIMSI
NILC_USP
UniOfLeipzig
bwbaugh
CodeX
NTNU
SSA-UO
UMCC
senti.ue-en
REACTION
IITB


Ablation Experiments on Tweets
all
all-lex.
all-man. lex.
all-auto. lex.
all-ngrams
all-word ngr.
all-char. ngr.
all-negation
all-POS
all-clusters
all-encod.


F-scores on SMS
Unigrams
NRC
KLUE
teragram
CodeX
nlp.cs.aueb.gr
ECNUCS
UT-DB
UNITOR
NILC_USP
SU-sentilab
FBM
senti.ue-en
SSA_UO
UMCC
uottawa


Ablation Experiments on SMS
all
all-lex.
all-man. lex.
all-auto. lex.
all-ngrams
all-word ngr.
all-char. ngr.
all-negation
all-POS
all-clusters
all-encod.