A Comparative Study of Different Sentiment Lexica for Sentiment Analysis of Tweets

Canberk Özdemir and Sabine Bergler
CLaC Labs, Concordia University
Montreal, Quebec, Canada

RANLP 2015

Sentiment

Concordia: *As one of the oldest and largest computer science and software engineering departments in Canada, we offer a vast array of options.*

McGill: *McGill University is one of the top research universities in Canada and is the only Canadian university to rank consistently among the top 25 universities in the world (based on Times and QS rankings).*

top in aFinn: +2 (scale: $[-3, 3]$)

$\Longrightarrow$ McGill (text) is more positive!

Background

	sentiment lexicon	**con-text**	**training**
bare baseline	one from literature		(majority)
	⋮		
NRC 2014	NRC-hashtag (bigger)	negation	SVM
CLaC SentiPipe	four from literature NRC-hashtag Gezi (biggest)	negation modality	SVM

	NRC	Gezi
Seeds	32 pos, 36 neg hashtags (Roget's)	35 pos, 34 neg hashtags (OAWT)
Corpus	775k tweets from 2012	2000k tweets from 2014
Filter		no duplicates or little-text tweets
Negation	*positive, negative*	*positive, negated positive, negative, negated negative*
POS	no	yes
Size	unigrams (54,128) bigrams (316,383) skip bigrams (308,791)	unigrams (376,863) bigrams (922,773) dependency triples (850,074)

Comparison of 5 Lexica

Lexicon	categories	strength	POS	size	creation
aFinn	pos, neg	yes	no	2,477	manual
Bing Liu	pos, neg	no	no	6,786	manual
MPQA	pos, neut, neg	no	yes	6,886	manual
NRC	pos, strongPos, neg, strongNeg	yes	no	25,721	auto
Gezi	pos, strongPos, neg, strongNeg	yes	yes	220,399	auto

Overlap and Agreement of 5 Lexica

Lex A	Lex B	$A \cap B$	Agree	$\frac{Agree}{A \cap B}$	Only A	Only B
aFinn	*Gezi*	1,911	1,624	0.850	565	218,488
aFinn	NRC	989	822	0.831	1,487	24,732
Liu	*Gezi*	4,028	3,386	0.841	2,758	216,371
Liu	NRC	1,840	1,488	0.809	4,946	23,881
MPQA	*Gezi*	4,105	2,993	0.729	2,781	216,294
MPQA	NRC	1,819	1,340	0.737	5,067	23,902

entry	positive	neutral	negative
provoking	NRC		aFinn, Gezi
terrified	MPQA		aFinn, Gezi, NRC
obey	MPQA		Gezi, NRC
fundamental	Gezi	MPQA	NRC

Performance of Individual Lexica

Neg/Mod enabled Baselines	Task 10B F measure 2015	2014	2013	Task 11 Cosine
aFinn	54.97	60.26	62.19	0.558
Gezi uni	54.65	60.81	57.86	0.554
Liu	53.88	53.90	57.20	0.555
MPQA	52.22	51.42	53.39	0.548
NRC uni	49.83	52.39	50.90	0.609
aFinn, Gezi uni	58.28	65.48	65.40	0.576
aFinn, NRC uni	57.33	63.01	64.15	0.617
SemEval Winner	64.84	70.96	69.02	0.758

$\Longrightarrow$ any one lexicon is a reasonable baseline, aFinn is best, all lexica interoperate efficiently

Pipeline for SemEval 2015 Participation

- **Tokens** → **Sentences** → **POS tags**
  - Tokenizing, sentence splitting, POS tagging with CMU
  - CMU tokens for tweet-specific POS
  - Annie tokens otherwise
  - Ritter's NER for fused named-entity tokens
  - remove non-content tokens (beginning and end of tweet)
- **Stanford Parser** constituent trees and dependencies

  parsing is feasible and effective
- **Negator** scope detection for *negation* and *modality*

  linguistic notions are effective even in shallow treatment
- **Feature Encoding**

Linguitic Context Encodings

(1) *Negation flips polarity and dampens strength*

*X * -0.5*

a *I'm hurt. -2*
b *I'm not hurt. +1*

(2) *Modality dampens strength*

*X * 0.5*

a *He's hurt. -2*
b *He may be hurt. -1*

Primary Feature Encoding

(3) *Working for 5 hours tonight* —
with no help *positive-aFinn-negated: -1*
should be joyful *positive-aFinn-mod: 1.5*

lexicon	trigger	sent.		score	feature
aFinn	*help*	+2	negated	-1	positive-aFinn-negated
aFinn	*joyful*	+3	mod	1.5	positive-aFinn-mod

Secondary Feature Encoding

Observed ad-hoc features:

- POS counts and scores by type
- First and last two tokens' POS type and sentiment score
- Highest and lowest sentiment scores within a tweet
- Frequencies of emoticons, negation/modality triggers, contrastive discourse markers, selected idioms, named entities

Ablation

Primary Feat. Subsets		
f_1	aFinn	9
f_2	MPQA	12
f_3	BingLiu	8
f_4	NRC unigrams	17
f_5	NRC bigrams	17
f_6	Gezi unigrams	17
f_7	Gezi bigrams	17

Primary Feat. cont'd		
f_8	dependency scores	13
f_9	dependency counts	8

Secondary Feat. Subsets		
f_{10}	POS based sc. & freq.	9
f_{11}	specific annotation freq.	12
f_{12}	position and min/max sc.	6

feature bundles	Task 10B F1 measures 2015	2014	2013	Task 11 Cosine
$f_{1,6,7,8,9,10,11,12}$	61.31	68.63	70.06	**0.768**
$f_{1,2,3,6,7,8,9,10,11,12}$	62.38	69.90	**70.85**	0.767
$f_{1,2,3,4,5,6,7,8,9,10,11,12}$	62.18	69.98	70.81	0.765
$f_{1,3,6,7,8,9,10,11,12}$	61.88	68.97	70.03	0.765
$f_{1,3,5,6,7,8,9,10,11,12}$	**62.64**	69.57	70.61	0.763
$f_{3,6,7,8,9,10,11,12}$	60.77	67.80	68.37	0.763
$f_{1,2,3,6,8,9,10,11,12}$	*62.00*	**70.16**	70.42	0.761
$f_{1,2,3,6,7,10,11,12}$	60.17	65.80	66.91	*0.758*
$f_{1,3,4,5,7,8,9,10,11,12}$	61.25	67.96	70.36	0.757
$f_{1,4}$	57.33	63.01	64.15	0.617
$f_{1,6}$	58.28	65.48	65.40	0.576

SemEval 2015 Results for Task 10B

40 Teams	#	**Tw15**	Sarc15	Tw14	Sarc14	Live-J	Tw13	SMS
Webis	1	**64.84**	53.59	70.86	49.33	71.64	68.49	63.92
unitn	2	64.59	55.01	73.60	55.44	72.48	72.79	68.37
Splusplus	5	63.73	60.99	**74.42**	42.86	**75.34**	**72.80**	67.16
IOA	7	62.62	**65.77**	71.86	51.48	74.52	71.32	68.14
CLaC	9	62.00	58.55	70.16	51.43	73.59	70.42	63.05
Grad-Ana	16	60.62	56.45	66.87	**59.11**	72.63	65.29	61.97
ECNU	18	59.72	52.67	66.37	45.87	74.40	65.25	**68.49**

F-measure

SemEval 2015 Results for Task 11

15 Teams	#	Overall	Sarcasm	Irony	Metaphor	Other
Cosine measure						
CLaC	1	**0.758**	0.892	0.904	**0.655**	0.584
UPF	2	0.711	0.903	0.873	0.520	0.486
LLT PolyU	3	0.687	0.896	**0.918**	0.535	0.290
elirf	5	0.658	**0.904**	0.905	0.411	0.247
Mean Squared Error						
CLaC	1	**2.117**	1.023	0.779	**3.155**	**3.411**
UPF	2	2.458	**0.934**	1.041	4.186	3.772
LLT PolyU	3	2.600	1.018	**0.673**	3.917	4.587
elirf	8	3.096	1.349	1.034	4.565	5.235

Error Analysis on Task 10B

	neutral	positive	negative
gold-neutral	**596**	244	58
gold-positive	184	**659**	34
gold-negative	207	135	**273**

- Tweets with contrastive discourse markers:
  I may not like the walking dead but Norman reedus is pretty attractive. (gold:positive, predicted:negative)
- Tweets with comparatives
  Three hours of sleep would be a lot worse if it was Monday. It's a Sugar Free Red Bull breakfast morning.# TGIF
  (gold:negative, predicted:positive)
- Adjective scope over sentiment-laden terms
  I find it very `difficult` *not to be happy with him*
  (gold:positive, predicted:negative)

Conclusions

- negation and modality trigger and scope detection in tweets is feasible with proper preprocessing and effective with current techniques
- the NRC technique can be replicated and extended to create even larger sentiment lexica
- size increase and use of multiple sentiment lexica is effective but leads to ever smaller increase in performance
- aFinn is an example of a manually compiled lexicon that adequately covers the sentiment core with a 5% lead over the NRC hashtag lexicon 10 times its size
- general, linguistically inspired features outperformed task-biased systems on deriving sentiment values of figurative language