Sentiment analysis on Twitter

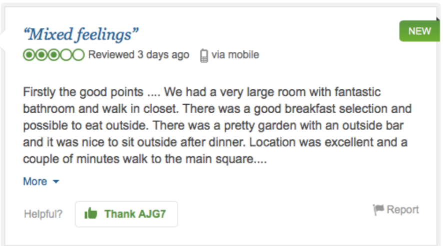
Manfred Stede, Univ of Potsdam

ESSLLI 2016

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- Twitter sentiment
- An ensemble approach to classifying tweets
- Building a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm

One Bolzano hotel @ tripadvisor.com



Near-Synonyms ?

- Opinion mining
- Sentiment analysis
- Subjectivity analysis

- I don't like this wine.
- There is a cat on the mat.
- I'm dizzy.
- Peter adores Barack Obama.
- I don't think that Trump can win the election.
- Last night I met this really nice musician.
- Hooray!
- That's probably a dromedar, not a camel.

Subjectivity

- The linguistic expression of somebody's **opinions**, **sentiments**, emotions, evaluations, beliefs, speculations (Wilson/Wiebe: MPQA guidelines)
- Private state: state that is not open to objective observation or verification

  Quirk, Greenbaum, Leech, Svartvik (1985). *A Comprehensive Grammar of the English Language*.

- Automatic subjectivity analysis classifies content as objective or subjective

Subjectivity

- **Sentiment:** an attitude or feeling
  (not necessarily directed toward sth)
- **Opinion:** an evaluation *of* something
  (necessarily directed)
- => Sentiment analysis and Opinion mining have a large overlap, but there can be sentiment analysis that does not mine opinions (e.g., capture the general mood in the newspapers)
- **In practice**, most automated systems reduce evaluation to polarity

Text-level sentiment analysis

- Firstly the good points .... We had a very large room with fantastic bathroom and walk in closet. There was a good breakfast selection and possible to eat outside. There was a pretty garden with an outside bar and it was nice to sit outside after dinner. Location was excellent and a couple of minutes walk to the main square.
- => positive

Text-level sentiment analysis

- Firstly the good points .... We had a very large room with fantastic bathroom and walk in closet. There was a good breakfast selection and possible to eat outside. There was a pretty garden with an outside bar and it was nice to sit outside after dinner. Location was excellent and a couple of minutes walk to the main square.
- => positive

Units for sentiment analysis

- **Text**
  - assume it has one topic and one overall orientation
- **Paragraph**
  - likewise, but can compute text orientation afterward
- **Sentence**
  - likewise, but can compute para orientation afterward
- **Phrase**
  - can capture things like
    While the breakfast was good, I couldn't stand dinner

Extension: „Polar facts"

- Firstly the good points .... We had a very large room with fantastic bathroom and walk in closet. There was a good breakfast selection and possible to eat outside. There was a pretty garden with an outside bar and it was nice to sit outside after dinner. Location was excellent and a couple of minutes walk to the main square.

Aspect-based sentiment analysis

- Firstly the good points .... We had a very large room with fantastic bathroom and walk in closet. There was a good breakfast selection and possible to eat outside. There was a pretty garden with an outside bar and it was nice to sit outside after dinner. Location was excellent and a couple of minutes walk to the main square.

Aspect-based sentiment analysis

- Firstly the good points .... We had a very large room with fantastic bathroom and walk in closet. There was a good breakfast selection and possible to eat outside. There was a pretty garden with an outside bar and it was nice to sit outside after dinner. Location was excellent and a couple of minutes walk to the main square.

Fine-grained sentiment analysis

- Sentiment words
- Intensifiers/diminishers
- Source
- Target

(slides from Jan Wiebe: MPQA annotation)

(Writer)

(Writer Xirao-Nima) (Writer Xirao-Nima)

"The report is full of absurdities," Xirao-Nima said the next day.

"The report is full of absurdities," Xirao-Nima said the next day.

Objective speech event
anchor: *the entire sentence*
source: <writer>
implicit: true

Direct subjective
anchor: *said*
source: <writer, Xirao-Nima>
intensity: high
expression intensity: neutral
attitude type: negative
target: report

Expressive subjective element
anchor: *full of absurdities*
source: <writer, Xirao-Nima>
intensity: high
attitude type: negative

A current trend: Even finer grain

What is good or bad for whom?

Roger Federer won the match against Nadar, who had been fervently supported by the audience.

=> „Sentiment flow" between entities in the text

Automatic (simple) polarity analysis

- **Start:** *prior polarity* of words
  - excellent, fantastic, good, nice, ...
  - boring, terrible, uncool, ugly, ...
- from lexicon (next)
- via statistics (last part of the lecture)

Some resources (needs updating)

- Lexicons
  - General Inquirer (Stone et al., 1966)
  - OpinionFinder lexicon (Wiebe & Riloff, 2005)
  - SentiWordNet (Esuli & Sebastiani, 2006)
  - German: SentiWS (Remus et al., 2010)

- Annotated corpora
  - Movie reviews (Hu & Liu 2004, Pang & Lee 2004)
  - MPQA corpus (Wiebe et al., 2005)
  - German: MLSA (Clematide et al., 2012)

- Tools
  - OpinionFinder (Wiebe et al., 2005)
  - SentiStrength (Thelwall et al., 2010)
  - SoCal (Taboada et al., 2011)

I can't say that I enjoyed my stay at the Belvedere Hotel. Other reviewers said it's a great place, but my impression was otherwise. Neither was the food particularly good, nor did we consider the location very convenient. Just a standard place to live for a day, that's it.

Contextual polarity

I can't say that I enjoyed my stay at the Belvedere Hotel. Other reviewers said it's a great place, but my impression was otherwise. Neither was the food particularly good, nor did we consider the location very convenient. Just a standard place to live for a day, that's it.

Overview

- Sentiment analysis: Introduction, Terminology
- **One system: SO-CAL**
- Twitter sentiment
- An ensemble approach to classifying tweets
- Building a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm

SO-CAL

- **S**emantic **O**rientation **CAL**culator
- Selling points:
  - Use crowdsourcing in building a lexicon
  - Rule-based approach to contextual polarity (with some new ideas)
  - Achieves good level of domain-neutrality

M. Taboada / J. Brooke / M. Tofiloski / K. Voll / M. Stede: Lexical Methods for Sentiment Analysis. *Computational Linguistics* 37(2), 2011

Size
2252 adjectives
1142 nouns
903 verbs
745 adverbs

Words collected from 500 movie and product reviews
(8 categories, balanced for pos and neg)
Extended with General Inquirer dict. (Stone et al 1966)

Manually ranked on -5 .. 5 scale: prior polarity and strength (reviewed by three native speakers; later by crowds)

Word	SO Value
monstrosity	−5
hate (noun and verb)	−4
disgust	−3
sham	−3
fabricate	−2
delay (noun and verb)	−1
determination	1
inspire	2
inspiration	2
endear	3
relish (verb)	4
masterpiece	5

Ambiguity

- Sense ambiguity: sometimes resolved via PoS
  - *plot*: neutral noun, negative verb
  - *novel*: positive adjective, neutral noun
- Connotation ambiguity: „resolved" by averaging
  - *The teacher inspired her students to pursue their dreams.*
  - *This movie was inspired by true events.*

Derivations

- Some nouns derived automatically from verb dictionary, but strength can change
  - *exaggerate*: -1
  - *exaggeration*: -2
  - also: *complicate / complication*, etc
  - (hypothesis: general trend?)

Derivations (2)

- Adverb dictionary built from adjectives, by *–ly* matching
- sometimes value needs to be corrected: *essential / essentially*

Word	SO Value
excruciatingly	-5
inexcusably	-3
foolishly	-2
satisfactorily	1
purposefully	2
hilariously	4

Coverage of the lexicon?

- **How to measure it?**
- Wilson et al 05:
  8000-word list of subjectivity clues
  BUT: many neutral, repeated, related entries
- Maybe the best argument for coverage is **performance in new domains**

Intensification

- **amplifiers** (*very*) / **downtoners** (*slightly*)
- Polanyi/Zaenen 06, Kennedy/Inkpen 06: **add** and **subtract** values
- **BUT:** degree of intensification should also depend on the word intensified – nonlinear
- In total, 177 intensifiers in the lexicon

Intensifier	Modifier %
slightly	-50%
somewhat	-30%
pretty	-10%
really	+15%
very	+25%
extraordinarily	+50%
(the) most	+100%

really very good$_3$
3 x (100 + 25%) x (100 + 15%)
= 4.3

Negation

- *The acting was not very good.*
- Some negators appear at long distance
  - *Nobody gives a good performance in this movie.*
- Strategy: Look backwards until a clause boundary (punctuation or connective) is reached
  - *I don't think this will be a problem.*

Negation: value change

- One approach: polarity flipping (e.g., Choi/Cardie 08)
- Problems
  - *excellent*: +5
  - *not excellent*: -5 ??
  - *atrocious*: -5
- => Use polarity shift (+/-4) rather than flip
  - *She's not terrific* (5 – 4 = 1) *but not terrible* (-5 + 4 = -1) *either*.
  - *It's not a spectacular* (5 - 4 = 1) *film*.

Irrealis blocking

- *For kids, this movie* ***could*** *be one of the* ***best*** *of the holiday season.*
- *I thought this movie* ***would*** *be as* ***good*** *as the Grinch, but unfortunately it wasn't.*
- **Implementation:** ignore SO words in the scope of an irrealis marker (scope via heuristic rule)
  - modals
  - conditionality
  - NPIs (any, anything, ..)
  - questions
  - material in quotes
  - certain verbs (doubt, expect, ...)
- *This* ***should*** *have been a* ***great*** *movie.* (3 -> 0)

Text-level features

- Negativity is „marked" and deserves more cognitive weight (+ 50%)
- Decrease the weight of repeated words
  - *I saw great acting, a great plot, and a great ending.*
  - ***n***th occurrence receives **1/*n*** of its full SO value

Evaluation: Lexicon complexity

- Use not/recommended value of the review: **>0** / **<0**
- 3 variants of the approach
  - simple: only 2/-2 values and 1/-1 intensification (Polanyi/Zaenen 06)
  - only-adj: use only adjectives
  - one-word: no multi-word exprs

Dictionary	Percent correct by corpus				
	Epinions 1	Epinions 2	Movie	Camera	Overall
Simple	76.75	76.50	69.79*	78.71	75.11*
Only-Adj	72.25*	74.50	76.63	71.98*	73.93*
One-Word	80.75	80.00	75.68	79.54	78.23
Full	80.25	80.00	76.37	80.16	78.74

*Statistically significant using the chi-square test, p < 0.05.

Evaluation

Subcorpus	Epinions 1			Epinions 2		
	Pos-F	Neg-F	Accuracy	Pos-F	Neg-F	Accuracy
Books	0.69	0.74	0.72	0.69	0.77	0.74
Cars	0.90	0.89	0.90	0.80	0.75	0.78
Computers	0.94	0.94	0.94	0.90	0.89	0.90
Cookware	0.74	0.58	0.68	0.79	0.76	0.78
Hotels	0.76	0.67	0.72	0.80	0.70	0.76
Movies	0.84	0.84	0.84	0.76	0.79	0.78
Music	0.82	0.82	0.82	0.83	0.81	0.82
Phones	0.81	0.78	0.80	0.85	0.83	0.84
Total	0.81	0.79	0.80	0.81	0.79	0.80

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- **Twitter sentiment**
- An ensemble approach to classifying tweets
- Building a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm

Sentiment analysis on Twitter

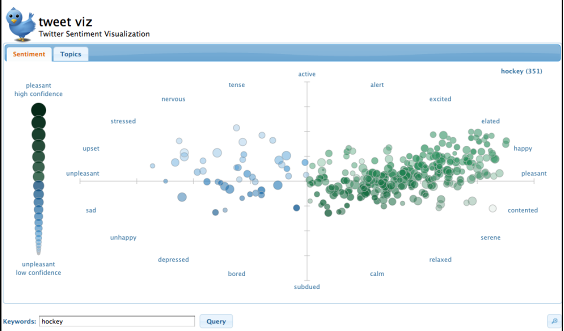
- **Very popular!**
- **SemEval shared task** since 2013
  - Tweet => **pos** / **neg** / **neut**
  - **Pos**:
    Gas by my house hit $3.39!!!! I'm going to Chapel Hill on Sat. :)
  - **Neg**:
    Dream High 2 sucks compared to the 1st one.
  - **Neut**:
    Battle for the 17th banner: Royal Rumble basketball edition

Some English corpora

- Go et al. (09): **1.6 million** tweets containing emoticons, mapped automatically to polarity classes
- Davidov et al. (10): **65.000** tweets with (1 of 50) „emotional" hashtags or (1 of 15) emoticons
- Barbosa/Feng (10): **200.000** tweets labelled by publicly-available sentiment classifiers
- Nakov et al. (13): **15.000** tweets manually annotated for SemEval shared task

- **Online tools**: sentiment140.com, ...

A. Joshi, P. Bhattacharyya, M. Carman: Automatic Sarcasm Detection: A Survey. arXiv preprint arXiv:1602.03426



Some applications

- identify the **general public's mood** on given events from media, politics, culture, economics
- evaluation of politicians' **TV debate performance**
- identifying the **employees' mood** in a company
- **opinion on products** or events
- identifying **product aspects** that are important to the users

A. Joshi, P. Bhattacharyya, M. Carman: Automatic Sarcasm Detection: A Survey. arXiv preprint arXiv:1602.03426

Some Twitter-specific features

- **Emoticons**
- **Abbreviations:** IMHO, LOL, ...
- **Emphasis markers**
  - UPPER CASE
  - word leeeengthening
  - multiple punctuation marks ???
  - many many many repeated words
  - ...
- **=> need to adjust the tweet pre-processing**

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- Twitter sentiment
- **An ensemble approach to classifying tweets**
- Building a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm

An *ensemble classifier* combining three tweet sentiment analysers

Output System 1
Output System 2
Output System 3
} **Ensemble classifier** } **Output**

M. Hagen, M. Potthast, M. Büchner, B. Stein: Twitter Sentiment Detection via Ensemble Classification. In: *Advances in Information Retrieval*. 37th European Conference on IR Research (ECIR 15)

(This team won the SemEval ST, subtask B, in 2015)

Implementation 1: NRC-Canada (rank 1)

- token n-grams (1 .. 4) (no weighting)
- character n-grams (3 .. 5) (no weighting)
- POS frequencies
- polarity dictionaries
  - existing: MPQA, ...
  - own: use #good etc. (70 tags) to harvest polarity terms
- number of >1 punctuation marks
- emoticons, their polarity, and final position
- Brown cluster IDs
- number of negated segments

S. M. Mohammad, S. Kiritchenko, and X. Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. of SemEval 2013*, pp. 321–327.

Implementation 2: GU-MLT-LT (rank 2)

- work with three versions of the tweet
  - raw
  - lowercased
  - collapsed (de-lengthened)
- unigrams (no weighting)
- Porter stems
- Brown cluster IDs
- polarity dictionary: SentiWordNet
- negated collapsed tokens and stems

T. Günther and L. Furrer. GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In *Proc. of SemEval 2013*, pp. 328–332.

Implementation 3: KLUE (rank 5)

- unigrams and bigrams, frequency-weighted
- length: number of tokens
- polarity dictionary: AFINN-111 (2500 words, twitter-specific)
- emoticons and colloquial abbreviations (list of 212/95, manually categorized)
- negation: polarity scores of 4 tokens following the neg-op are adjusted

T. Proisl, P. Greiner, S. Evert, and B. Kabashi. Klue: Simple and robust methods for polarity classification. In *Proc. of SemEval 2013*, pp. 395–401.

Ensemble

- Hagen et al re-implemented the three approaches

Classifier	Original SemEval 2013	Reimplemented Version
NRC	69.02	69.44
GU-MLT-LT	65.27	67.27
KLUE	63.06	67.05

- Observation: Each approach is correct on many instances where the others fail
- Observation: Simple **majority voting** performs worse than NRC alone
- Observation: **Confidences** provided by the classifiers give a good hint on uncertainties
  (e.g.: two narrowly prefer A; one clearly prefers B; => B)

Ensemble (2)

- **train** each re-implementation on SemEval-13 training set
  - 9,728 tweets (3,662 pos, 1,466 neg, 4,600 neut)
    - crawled for various entities (Gaddafi, Steve Jobs, ...), products (Kindle, ...), events (earthquake, NHL playoffs, ...)
- **test:**
  - use no weighting
  - simple **averages of the three classifiers' confidence scores** for each category
  - ignore their overall predictions

Evaluation

- Datasets:
  - 2013: 3,813 tweets (1,572 pos, 601 neg, 1,640 neut)
  - 2014: 1,853 tweets (982 pos, 202 neg, 669 neut)

Ranking SemEval 2013		Ranking SemEval 2014		
Team	F1-score	Team	F1 on 2014	F1 on 2013
Our ensemble	**71.09**	TeamX	70.96	72.12
NRC-Canada	69.02	coooolll	70.14	70.40
GU-MLT-LT	65.27	RTRGO	69.95	69.10
teragram	64.86	NRC-Canada	69.85	70.75
BOUNCE	63.53	**Our ensemble**	**69.79**	**71.09**
KLUE	63.06	TUGAS	69.00	65.64
AMI&ERIC	62.55	CISUC KIS	67.95	67.56
FBM	61.17	SAIL	67.77	66.80
AVAYA	60.84	SWISS-CHOCOLATE	67.54	64.81
SAIL	60.14	Synalp-Empathic	67.43	63.65
Average	53.70	Average	60.41	59.78

Evaluation (2)

Ensemble	F1-score	$prec_{pos}$	rec_{pos}	$prec_{neg}$	rec_{neg}
All	71.09	72.61	79.60	65.73	66.72
All - GU-MLT-LT	70.67 (-0.42)	72.83	78.78	67.31	64.06
All - KLUE	70.56 (-0.53)	73.39	78.59	65.22	65.22
All - NRC	68.80 (-2.29)	69.15	78.71	57.97	71.38

Error analysis

		Computed classification of our ensemble					
		SemEval 2013			SemEval 2014		
		positive	neutral	negative	positive	neutral	negative
Gold label	positive	1,249	234	86	768	184	30
	neutral	394	1,123	123	177	436	56
	negative	77	123	401	41	33	128

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- Twitter sentiment
- An ensemble approach to classifying tweets
- **Building a sentiment-annotated corpus of tweets**
- Complication: Detecting irony and sarcasm

German Twitter Sentiment Corpus

- **Tracking with keywords** (several dozen) for these categories:
  - federal election
  - papal conclave
  - general political issues
  - casual everyday conversations (from Scheffler 2014)
- Plus: set of unfiltered tweets
- => **27 million** tweets

Uladzimir Sidarenka: PotTS: The Potsdam Twitter Sentiment Corpus. Proc. of LREC 2016
https://github.com/WladimirSidorenko/PotTS

Corpus creation

- Goal: build a representative excerpt of the set, such that it includes a fair amount of sentiment
- For each category, divide tweets into three bins
  - Tweets containing >=1 „polar" words (SentiWS)
  - Tweets containing no polar words but emoticons or exclamation marks
  - All others
- Randomly select 666 tweets from each bin
  => **7.992 tweets**

Annotation scheme

- evaluative expressions: words/phrases with an *inherent* evaluative meaning
  – polarity: positive, negative, comparative
  – intensity: weak, mdedium, strong
  – irony/sarcasm: +/-
- intensifiers
  – degree
- diminishers
  – degree
- negations

- targets
- sources
- sentiments: minimal units in which evaluative expressions and targets appear together
  – irony/sarcasm: +/-

Example

[[Diese Milliardeneinnahmen]$_{target}$ sind selbst [Schäuble]$_{source}$ [peinlich]$_{emo\text{-}expression}$]$_{sentiment}$

[[*These billions of revenues*]$_{target}$ *are* [*embarrassing*]$_{emo\text{-}expression}$ *even for* [*Schäuble*]$_{source}$]$_{sentiment}$

Technicalities of annotation

- **Tool:** MMAX2 (Müller, Strube 2004)
  – markables: possibly discontinuous tokens
  – multiple annotations on the same markable
  – relations among markables
  – standoff XML
- **Tokenization** via a (slightly) adapted version of Potts's tokenizer
- **Data:** 80 project files of roughly 100 tweets each, single topic, equal share of bins

Computing annotator agreement

- **Cohen kappa** over token annotations
- **Complication:**

  [My [sister hates [[this nice book]]]]

- **binary** kappa:
  – tokens counted multiple times
  – spans agree when overlapping
- **proportional** kappa:
  – tokens counted only once
  – spans have to be identical

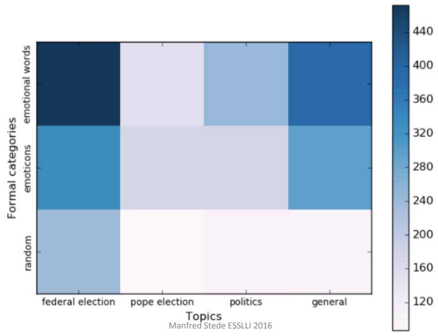
Annotation procedure

- Both annotators labeled half of the corpus, after only minimal training
  => binary kappa for sentiments: 0.38
  => binary kappa for evaluative exprs: 0.64
- Compute differences of annotations, highlight, let annotators re-consider individually (consulting with the author)
  => binary kappa for sentiments: 0.68
- Annotate the rest of the data (79% / 100%)

Annotator agreement in the three stages

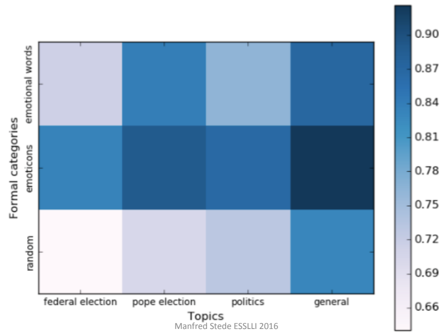
Element	Initial annotation					Adjudication step					Final corpus				
	M_1	A_1	M_2	A_2	κ	M_1	A_1	M_2	A_2	κ	M_1	A_1	M_2	A_2	κ
Binary Kappa															
Sentiment	4,215	7,070	3,484	9,827	**38.05**	8,198	8,530	8,260	14,034	**67.92**	14,748	15,929	14,969	26,047	**65.03**
Target	1,103	1,943	1,217	4,162	**35.48**	3,088	3,407	2,814	5,303	**65.66**	5,765	6,629	5,292	9,852	**64.76**
Source	159	445	156	456	**34.53**	573	690	545	837	**72.91**	966	1,207	910	1,619	**65.99**
EExpression	1,951	2,854	2,029	3,188	**64.29**	3,164	3,298	3,261	4,134	**85.68**	5,574	5,989	5,659	7,419	**82.83**
Intensifier	57	101	59	123	**51.71**	111	219	113	180	**56.01**	192	432	194	338	**49.97**
Diminisher	3	10	3	8	**33.32**	9	16	10	16	**59.37**	16	30	17	34	**51.55**
Negation	21	63	21	83	**28.69**	68	84	67	140	**60.21**	111	132	110	243	**58.87**
Proportional Kappa															
Sentiment	3,269	6,812	3,269	9,796	**31.21**	7,435	8,243	7,435	13,714	**61.94**	13,316	15,375	13,316	25,352	**58.82**
Target	898	1,905	898	4,148	**26.85**	2,554	3,326	2,554	5,212	**57.27**	4,789	6,462	4,789	9,659	**56.61**
Source	153	439	153	456	**33.75**	539	676	539	833	**71.12**	898	1,180	898	1,604	**64.1**
EExpression	1,902	2,851	1,902	3,180	**61.36**	3,097	3,290	3,097	4,121	**82.64**	5,441	5,977	5,441	7,395	**80.29**
Intensifier	57	101	57	123	**50.81**	111	219	111	180	**55.51**	192	432	192	338	**49.71**
Diminisher	3	10	3	8	**33.32**	9	16	9	15	**58.05**	16	30	16	33	**50.78**
Negation	21	63	21	83	**28.69**	67	83	67	140	**60.03**	110	131	110	242	**58.92**

Sentiment in the categories



Manfred Stede ESSLLI 2016

Agreement on eval.exprs in categories



Manfred Stede ESSLLI 2016

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- Twitter sentiment
- An ensemble approach to classifying tweets
- Buidling a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm

Manfred Stede ESSLLI 2016

Irony and Sarcasm

- Webster: Sarcasm: the use of words that mean the opposite of what you really want to say; especially in order to insult someone, to show irritation, or to be funny
- Webster: Irony: the use of words that mean the opposite of what you really think; especially in order to be funny

http://www.merriam-webster.com

Manfred Stede ESSLLI 2016

Irony and Sarcasm

- Standard case:
  Surface:positive & Depth:negative
  A: *How did you like the movie?*
  B: (yawns ostensibly) *Totally exciting.*

- Non-standard case:
  Surface:negative & Depth:positive
  X's child always gets good grades in school, Today brought home another „A“
  X: *Ouch, one of these terrible results!*

Manfred Stede ESSLLI 2016

Irony and Sarcasm

- Fowler: *Modern English Usage* (1926):
- In terms of motive and aim, sarcasm aims to **inflict pain**, while irony aims for **exclusiveness**.
- For the audience sarcasm is perceived by the **victim and bystanders**, while irony is intended for **inner circle**.
- For province, ...

Manfred Stede ESSLLI 2016

Computational approaches

- **Classification**
  - Two-way: +/- sarcastic
  - Variant: Sense disambiguation (Ghosh et al 15) – words can have an additional sarcastic sense
  - Three-way: saracasm / irony / humour
- **Sequence labeling** (Wang et al. 15)

A. Joshi, P. Bhattacharyya, M. Carman: Automatic Sarcasm Detection: A Survey. arXiv preprint arXiv:1602.03426

Manfred Stede ESSLLI 2016

Twitter datasets

- Manual annotation
  - Riloff et al. (13): +/- sarcastic
  - Maynard/Greenwood (14): 600 tweets with subjectivity, sentiment, sarcasm annotation
- Hashtag-based supervision
  - Claim: the only way to get reliable data
  - #sarcasm, #sarcastic, #not, ...
  - e.g., Reyes et al. (13): 40.000 Tweets
  - Gonzales et al. (11) eliminate syntactically-integrated tags: *#sarcasm is popular in india*

Manfred Stede ESSLLI 2016

One approach: Riloff et al. (13)

- **Observation:** In Twitter, sarcasm often comes with a characteristic structure: positive/negative contrast between a sentiment and a situation

- Oh how I love being ignored. #sarcasm
- Thoroughly enjoyed shoveling the driveway today! :) #sarcasm
- Absolutely adore it when my bus is late #sarcasm

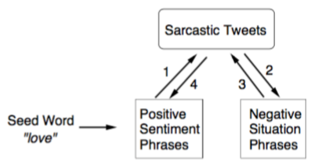
Manfred Stede ESSLLI 2016

Find it automatically

- **Positive sentiment** words: relatively easy (lexicon)
- **Negative situations**: difficult (no resource)
- **Idea**: bootstrapping approach to learn both parts from lots of tweets

Manfred Stede ESSLLI 2016

Bootstrapping approach



Manfred Stede ESSLLI 2016

Bootstrapping approach

- **Underlying assumption:** If you find a pos-sentiment or a neg-situation in a sarcastic tweet, you have found (part of) the source of the sarcasm
- Exploit syntactic structure to extract phrases
  - pos-sentiment in verb phrase or predicative expr
  - negative activities/states as verb complements
- Avoid parsing; approximate via POS + proximity

Manfred Stede ESSLLI 2016

Bootstrapping approach

- Thoroughly enjoyed *shoveling the driveway* today!
- [+ verb-phrase] [- situation-phrase]

- given the +VP, harvest n-grams to the right of it, score them, add to the pool
- given the –sit, harvest n-grams to the left of it, score them, add to the pool
- (add a little machinery for predicative constructions)

Data

- Collect 35.000 tweets with #sarcasm or #sarcastic
- Collect 140.000 random tweets, remove those with #sarcasm, consider the rest to be non-sarcastic

- Use Twitter-specific POS tagger (Owoputi et al. 13)

Learn –sit phrases

- Given a +VP, take the subsequent 1-gram, 2-gram 3-gram
- *I love waiting forever for the doctor*
  => waiting / waiting forever / waiting forever for
- apply POS-pattern filtering, using pre-defined lists (V+V, V+ADV, ...)
  => waiting / waiting forever

Learn –sit phrases

- Score each candidate phrase:

$$\frac{|\text{follows}(-\text{candidate}, +\text{sentiment}) \ \& \ \text{sarcastic}|}{|\text{follows}(-\text{candidate}, +\text{sentiment})|}$$

- discard candidates with freq < 3
- rank the candidates according to scores
- add top-20 candidates with score > 0.8 to pool
- remove existing phrases that are subsumed by new ones (e.g., *waiting* removes *waiting forever*)

Learn +verb phrases

- For „standard“ VPs, same procedure as above
- For predicative constructions,
  - use a list of 24 copular verbs
  - devise patterns of 1-grams and 2-grams
  - for scoring, replace adjacency with proximity

Experimental results

- 26 +VPs
- 20 +Pred-exprs
- 239 –sit phrases

Positive Verb Phrases (26): missed, loves, enjoy, cant wait, excited, wanted, can't wait, get, appreciate, decided, loving, really like, looooove, just keeps, loveee, ...

Positive Predicative Expressions (20): great, so much fun, good, so happy, better, my favorite thing, cool, funny, nice, always fun, fun, awesome, the best feeling, amazing, happy, ...

Negative Situations (239): being ignored, being sick, waiting, feeling, waking up early, being woken, fighting, staying, writing, being home, cleaning, not getting, crying, sitting at home, being stuck, starting, being told, being left, getting ignored, being treated, doing homework, learning, getting up early, going to

Evaluation

- **Create gold standard**
  - collect 1600 tweets with #sarcasm + 1600 without
  - remove #sarcasm tag
  - manually annotate for +/-sarc (in any way)
  - Cohen kappa: 0.8
  - 742/3200 tweets judged as +sarc (23%)
  - only 713 of the 1600 with hashtag were judged as +sarc
    - sarcasm can be invisible when context is missing
    - sarcasm can arise from a URL rather than from the tweet
  - 29 of the 1600 without hashtag were judged +sarc (1.8%)

Evaluation

- **Baselines**
  - SVM classifiers for unigrams and unigrams+bigrams
    => F-score 0.46 / 0.48
  - Existing sentiment lexicons (Liu 05, MPQA 05, AFINN11), in various configurations (pos and/or neg sent, unordered versus ordered)
    => F-score up to 0.47

Evaluation

Our Bootstrapped Lexicons			
Positive VPs	.28	.45	.35
Negative Situations	.29	.38	.33
Contrast(+VPs, –Situations), Unordered	.11	.56	.18
Contrast(+VPs, –Situations), Ordered	.09	**.70**	.15
& Contrast(+Preds, –Situations)	.13	.63	.22
Our Bootstrapped Lexicons ∪ SVM Classifier			
Contrast(+VPs, –Situations), Ordered	.42	.63	.50
& Contrast(+Preds, –Situations)	.44	.62	**.51**

The hybrid approach: label a tweet as +sarc if *either* the bootstrapped-lexicon classifier *or* the unigram/bigram SVM classifier predicts +sarc

Riloff et al. 13: Conclusions

- Focus was on **one type of sarcasm constructions**
- **Bootstrapping** with pattern-based recognition yields **good precision** (but **low recall**)
- **Combining** the method with standard **word-based classification** works quite well
- **Ordering information** is important: „our“ [+VP] [-sit] construction is in fact characteristic

The last word on sarcasm: Context!

- Author's history: (Rajadesingan et al. (15)) compute features from previous posts
  - familiarity with twitter (in terms of use of hashtags),
  - familiarity with language (in terms of words and structures)
  - familiarity with sarcasm
- Topic history: Is the topic likely to evoke sarcasm?
- Conversation history:
  - „it is not always easy to identify sarcasm in tweets because sarcasm often depends on conversational context that spans more than a single tweet.“ (Riloff et al. 13)
  - „Using tweets in an ongoing conversation in order to predict sarcasm has not been explored yet.“ (Joshi et al. 15)

Overview

- Sentiment analysis: Introduction, Terminology
- One system: SO-CAL
- Twitter sentiment
- An ensemble approach to classifying tweets
- Buidling a sentiment-annotated corpus of tweets
- Complication: Detecting irony and sarcasm