SENTIMENT ANALYSIS

CS 562/662: Natural Language Processing

2015-01-27

Just what is *sentiment*, exactly?

Three ways to classify text sentiment:

- Using a lexical knowledge base
- Inducing a task-specific lexical knowledge base
- Using generalized classification algorithms

Nearly everything here also applies to generic text/ document classification problems.

DEFINITIONS

sentiment, noun: “a view or attitude towards a situation or event; an opinion” (NOA)

These sentiments are *subjective* (extracting information from objective sources is Information Extraction)

These sentiments are present in text (no audio, gesture, etc.)

Selecting from a set of nominal labels is sentiment *classification*

Predicting a ordinal or continuous score (like a user’s rating) is sentiment *regression*

AFFECT TYPOLOGY

Emotions (angry, happy, sad, proud, ashamed)

Mood (cheerful/optimistic, depressed/hopeless)

Attitudes (love, hate, value, desire)

Interpersonal stance (supportive, flirtatious, distant)

Personality traits (Big 5: open, conscientious, extraverted, agreeable, neurotic)

[Source: Scherer 1984]

Sentiment analysis attempts to detect *who* holds $what_1$ attitude towards $what_2$.

Common simplifying assumption: we know the *who* and $what_2$ from context.

Well as usual Keanu Reeves is nothing *special*, but surprisingly, the *very talented* Laurence Fishburne is not so *good* either.

This movie doesn't *care* about *cleverness*, *wit*, or any other kind of *intelligent humor*.

If you are reading this because it is your *darling* fragrance, please wear it at home *exclusively*, and tape the windows shut.

Could *anyone but Uwe Boll* make such a *bad* movie?

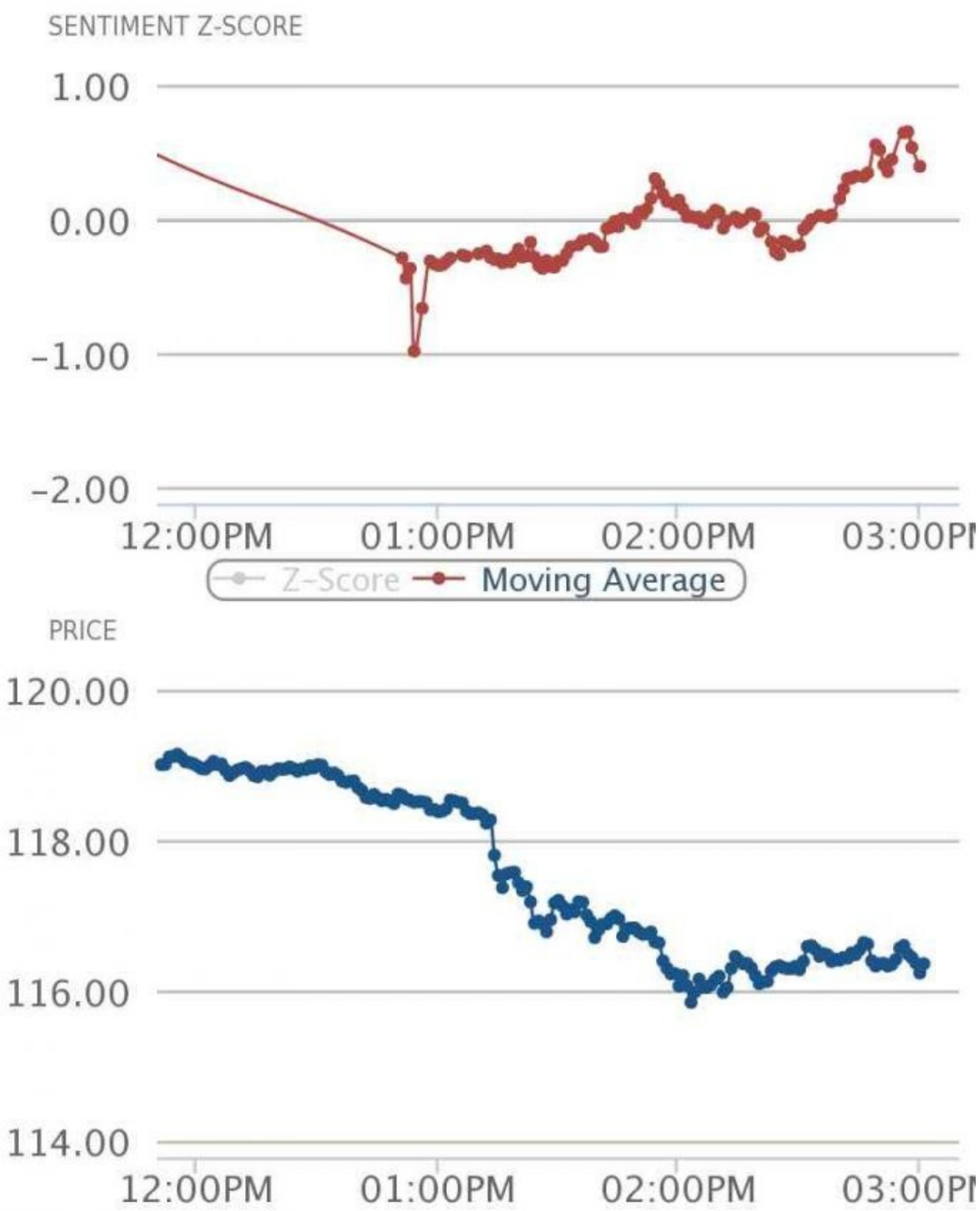
APPLICATIONS (RIPPED FROM THE HEADLINES)

How To Manage Your Brand On Social Media

World's Largest Hedge Fund Uses Twitter For Real-Time Economic Modeling

Investigator: Herman Cain Innocent Of Sexual Advances

Rove: Lee's Bill To Defund Obamacare Gives Obama "A Gigantic Stick With Which To Beat Us"



[Source: Market Prophit]

KNOWLEDGE-BASED SENTIMENT ANALYSIS

HARVARD GENERAL INQUIRER

Major categories:

`Pstv` (*awareness*) vs. `Ngtv` (*abolish*)

`Strong` (*antagonism*) vs. `Weak` (*brittle*)

`Active` (*accelerate*) vs. `Passive` (*ail*)

[Source: Stone et al. 1966]

LINGUISTIC INQUIRY & WORD COUNT

LIWC “dimensions”:

Social words

Positive emotions

Negative emotions

[Source: Pennebaker et al. 2007]

INDUCING SENTIMENT LEXICA

OPINION LEXICON

Approx. 6,800 words rated *positive* or *negative*, extracted from WordNet:

If *X* and *Y* are antonyms, assume one is positive and one is negative; use a short list of seed words and an iterative algorithm following the synset chain to figure out which is which.

[Source: Hu & Liu 2004]

COLLOCATION HEURISTICS

Co-occurrence with *excellent* vs. *poor*

$$\text{Polarity}(t_i) = \text{PMI}(t_i, \text{“excellent”}) - \text{PMI}(t_i, \text{“poor”})$$

Co-occurrence with :) and :(

[Sources: Turney 2002, Hogenboom et al. 2013]

CONJUNCTION HEURISTIC

And-conjoined adjectives have the same polarity

Kind and *caring*
?*Kind* and *brutal*

But-conjoined adjectives have an opposite polarity

Kind but *incompetent*
?*Kind* but *caring*

[Source: Hatzivassiloglou and McKeown 1997]

FULLY SUPERVISED SENTIMENT ANALYSIS

TOKEN-CLASS LIKELIHOOD

How likely is each token *t* to appear in each sentiment class *c*?

Likelihood:

$$P(t \mid c) = \frac{f(t,c)}{\sum_{t \in T, c \in C} f(t,c)}$$

Can also be scaled by $P(t)$

[Source: Potts 2011]

TOKENIZATION AND FEATURE EXTRACTION

- Tokenize, removing markup (`@ZizekOnNFL`), numbers, dates, but preserving ideograms like `:)`

- Case-fold (except, perhaps, words in all caps, e.g., `BAD`)

- Map “lengthened” tokens onto a canonical lengthened token (`Coooooooooooooollllllllllllll` → `COOL_LENGTHENED`)

- Optional: filter by term frequency or document frequency, remove stopwords, stem/lemmatize, etc.

- Extract binary (boolean) token presence/absence features (i.e., we “clip” counts at 1)

[Sources: Pang et al. 2002, Brody & Diakopoulous 2011]

MULTINOMIAL BINARY NAÏVE BAYES CLASSIFIER

Decision rule:

$$\text{argmax}_c \, P(c) \prod_{c \in C} P(t_i \mid c)$$

(Laplace smoothing would be a good idea here.)

Experiments on predicting the polarity of IMDB movie reviews by Pang et al. (2002) find more elaborate classifiers (tuned SVMs—even using a linear kernel—and MaxEnt classifiers) perform better, with binary accuracies around 83%.

OPEN PROBLEMS

NEGATION

Many ways to express negation:

This movie does*n't* care about…humor.

No one thinks this is a good movie.

This is not a good movie.

Ce *n'*est *pas* un bon film.

C'est le *pire* film que j'ai *jamais* vu.

HANDLING NEGATION

- Use n-gram (or “skip-gram”) features instead of word features
- Mark words between negation and following punctuation with NOT_ (Das & Chen 2001):
  - I didn’t like *The Godfather Part III*, but I loved *Lost In Translation*.

    ```
    I did n’t NOT_like NOT_The NOT_Godfather
    NOT_Part NOT_3 , ...
    ```

- Use sequence-based or parse-based heuristics (e.g., Jia et al. 2009):
  - E.g., assume the negation of a double-object verb scopes the direct object but not the indirect object

SUBJECTIVITY DETECTION

This movie is directed by *horrible* director *Paul Verhoeven*, **but I loved it anyways**.

DOUBLE ENTENDRE IDENTIFICATION

TWSS jokes ‘consist of saying “that’s what she said” after someone else utters a statement in a non-sexual context that could also have been used in a sexual context.’

Kiddon & Brun (2011) use cosine similarity to a corpus of erotica to estimate *token sexiness* features, and combine them into a decision rule with an bagged SVM.