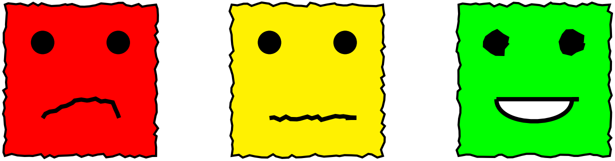
Sentiment Analysis

Subrata Pal

Iowa State University

Nov 19, 2019

Sentiment analysis:



Sentiment analysis and opinion mining (introduced in the early 2000s) is the method to **understand and analyze opinions and feelings**.

Overall view:

- *Human readers* use their *knowledge of language* to *determine the feelings behind a text*. This applies to individual words as well as their meaning within the document's context.
  A computer, unfortunately, needs to be told which words or phrases change the polarity of a text

Overall view:

- *Human readers* use their *knowledge of language* to *determine the feelings behind a text*. This applies to individual words as well as their meaning within the document's context.

  A computer, unfortunately, needs to be told which words or phrases change the polarity of a text
- Two common approaches to sentiment analysis:

Overall view:

- *Human readers* use their *knowledge of language* to *determine the feelings behind a text*. This applies to individual words as well as their meaning within the document's context.
  A computer, unfortunately, needs to be told which words or phrases change the polarity of a text
- Two common approaches to sentiment analysis:
  1. **Dictionary, or lexicon:** The lexicon approach assigns a polarity to words from a previously created dictionary.

Overall view:

- *Human readers* use their *knowledge of language* to *determine the feelings behind a text*. This applies to individual words as well as their meaning within the document's context.

  A computer, unfortunately, needs to be told which words or phrases change the polarity of a text
- Two common approaches to sentiment analysis:
  1. **Dictionary, or lexicon:** The lexicon approach assigns a polarity to words from a previously created dictionary.
  2. **Corpus, or learning:** It builds a sentiment classifier for a document set previously annotated with sentiments. From there, a classifier is trained that can be applied to new, unseen data.

Lexicon Approach:

- The lexicon approach uses **previously scored words and word phrases to assign a sentiment value** to a new text. Each word or phrase that matches the corresponding word or phrase in the lexicon is given that value. For the full text, the values are then summed up.
- Numerous scored lexicons exist for use in sentiment analysis; e.g.: OpinionFinder (Wilson et al. 2005), General Inquirer (Stone et al. 1966), SentiWordNet (Baccianella et al. 2010), and AFINN (Nielsen 2011).

Some ‘positive’ and ‘negative’ words:

+	-
accept	abandon
courageous	aggravate
enjoy	dismal
fondness	disturb
forgive	harsh
humor	inadequate
improvement	obliterate
luckily	obstinate
optimism	rash
outstanding	scorn
virtuous	unhappy
worthy	worthless

Sample of positive and negative words that coincide and are consistent across some lexicons

Sneak peek

- Let's see a mixed review:
  The price is good, but I was disappointed in the food; it was bland.
- It is unitized, Tokenized, Cleaned, then stop-word removed and Lemmatized to get:
  [price][good][disappoint][food][bland]
- good -> 3; disappoint -> (-2)
  They are summed up as 1 to get the final score for the document.
- For lexicon with +/- only, the final score for any documents is just +/-.

Questions and concerns

- Why it is not normalized?

Questions and concerns

- Why it is not normalized?
- What about sarcasm or metaphor or similes?

Questions and concerns

- Why it is not normalized?
- What about sarcasm or metaphor or similes?
- (Extreme case scenario) what about the cases in which a document does not have any terms which is common to any scored lexicons?

Questions and concerns

- Why it is not normalized?
- What about sarcasm or metaphor or similes?
- (Extreme case scenario) what about the cases in which a document does not have any terms which is common to any scored lexicons?

- To evaluate the results, we can use measures of external validity. Additionally, we can check them against human-coded sentiment. An agreement rate of at least 80% is considered good (Mullich 2013).

Questions and concerns

- Why it is not normalized?
- What about sarcasm or metaphor or similes?
- (Extreme case scenario) what about the cases in which a document does not have any terms which is common to any scored lexicons?

- To evaluate the results, we can use measures of external validity. Additionally, we can check them against human-coded sentiment. An agreement rate of at least 80% is considered good (Mullich 2013).
- To evaluate a text, a random subset of documents of managable size is chosen. Next, have people manually read and score each document's polarity and try to have $\sim$ 80% agreement.

(Machine) Learning Approach

The machine learning approach to sentiment analysis builds a classifier on a dataset with labeled sentiments.

- Naive Bayes: We perform a NB analysis on the unweighted DTM
- We can do SVM also using the tfidf-weighted DTM.
- We can do Logistic regression also. (In the book, it has been said that this process can be used in Binary sentiments. But we all know, this can be easily extended to non-binary cases also).