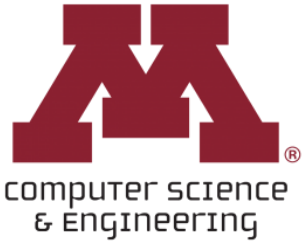
CSCI 5541: Natural Language Processing

Lecture 3: Text Classification

Dongyeop Kang (DK), University of Minnesota

dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io



UNIVERSITY OF MINNESOTA
Driven to Discover®

Outline

- Applications of text classification
- Applications of sentiment classifier in other fields
- Why is sentiment analysis difficult?
- How can we build a sentiment classifier?
- Tutorial on building text classifier using Scikit-Learn (25m) (Zae)

Movie review

Eternals is far from perfect, but it pushes the MCU into promising new territory....it feels like an amalgam of what Marvel does best - splendidly chaotic fight scenes, dazzling special effects, and stories that speak to who we are as human beings.

December 17, 2021 | Full Review...

Michael Blackmon
BuzzFeed News
★ TOP CRITIC

Erick V

½

Really bad story, uninteresting and boring.

Product review

Customer reviews & ratings

4.3 out of **5**

★★★★½ (3074 reviews)

Most helpful positive review

★★★★★ Verified Purchaser

GREAT CONSOLE

One of the greatest consoles ever! Granted, there's not many PS5 exclusives out but I have had every console from Playstation, from PS1-PS3, and all the fatboys and slim models of those consoles and I enjoyed every console but I never purchased a Playstation 4, so that makes the Playstation 5 an even bigger purchase because I have a HUGE catalog to enjoy. Right now, I'm enjoying Spiderman Remastered and I recently beat God Of War. I can't wait for the experience of what Sony will bring with the Playstation 5 in the next 7 years. Thank You Walmart for keeping me up to date with the status of my purchase and thank you for the early delivery!!

Jesse

Most helpful negative review

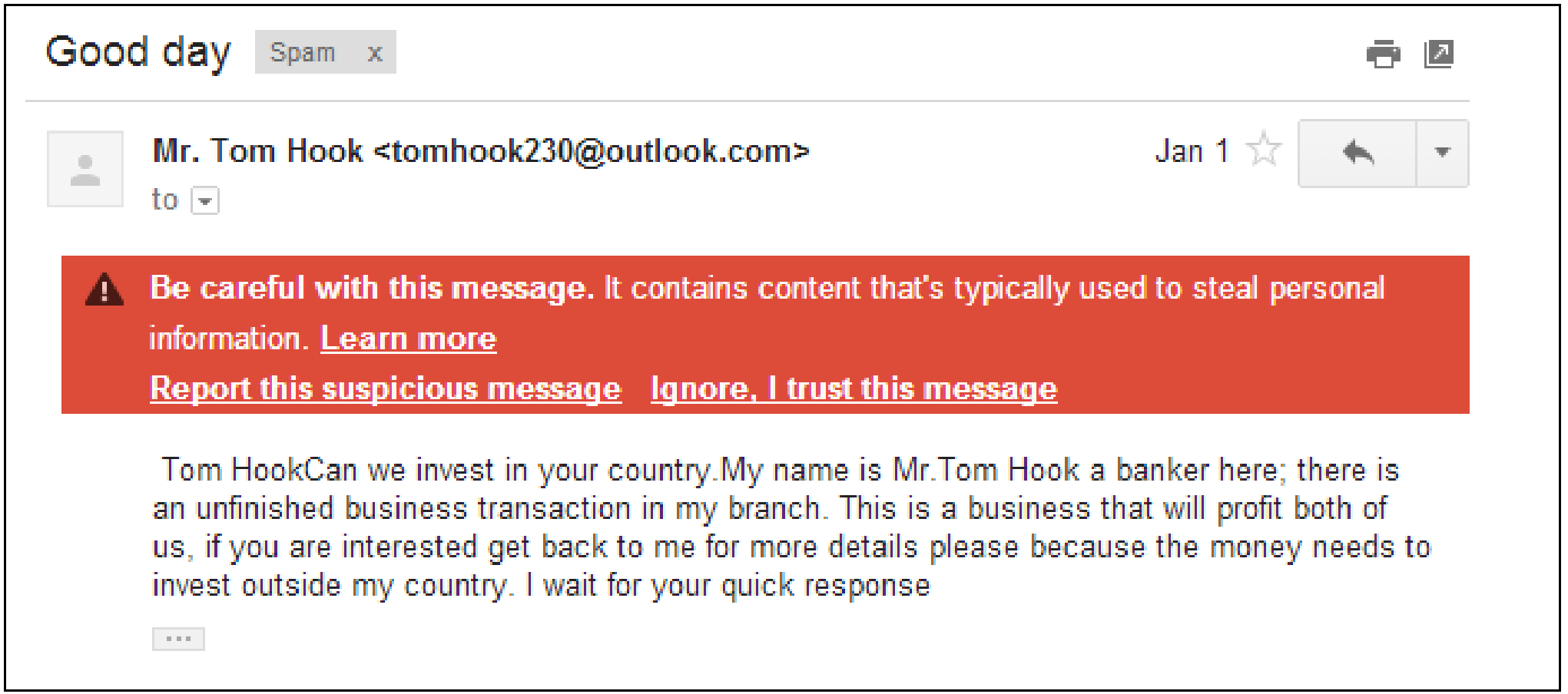
★☆☆☆☆ Verified Purchaser

a little passing would be nice

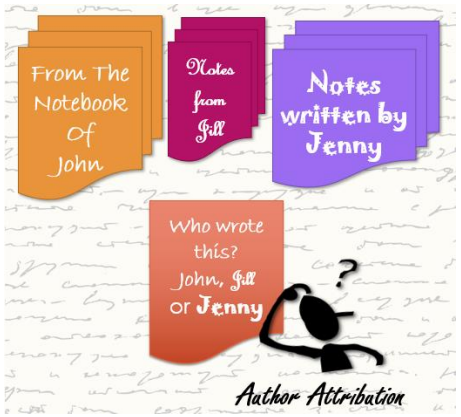
Ordered a $500 gaming console. It was shipped in a box twice the size needed and there was NO packing in the box!

Ron

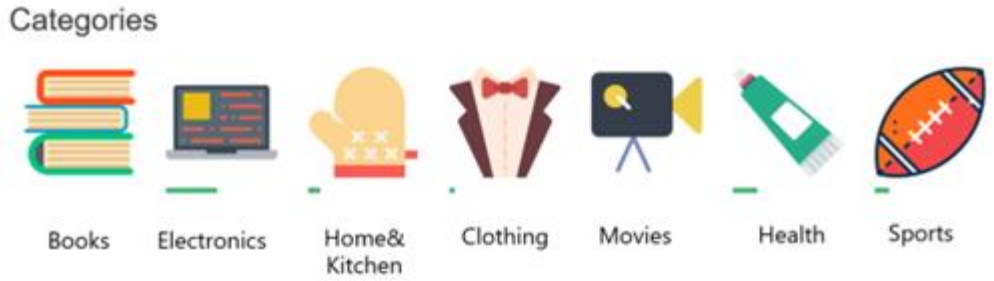
Spam detection



Language Identification



Authorship Identification

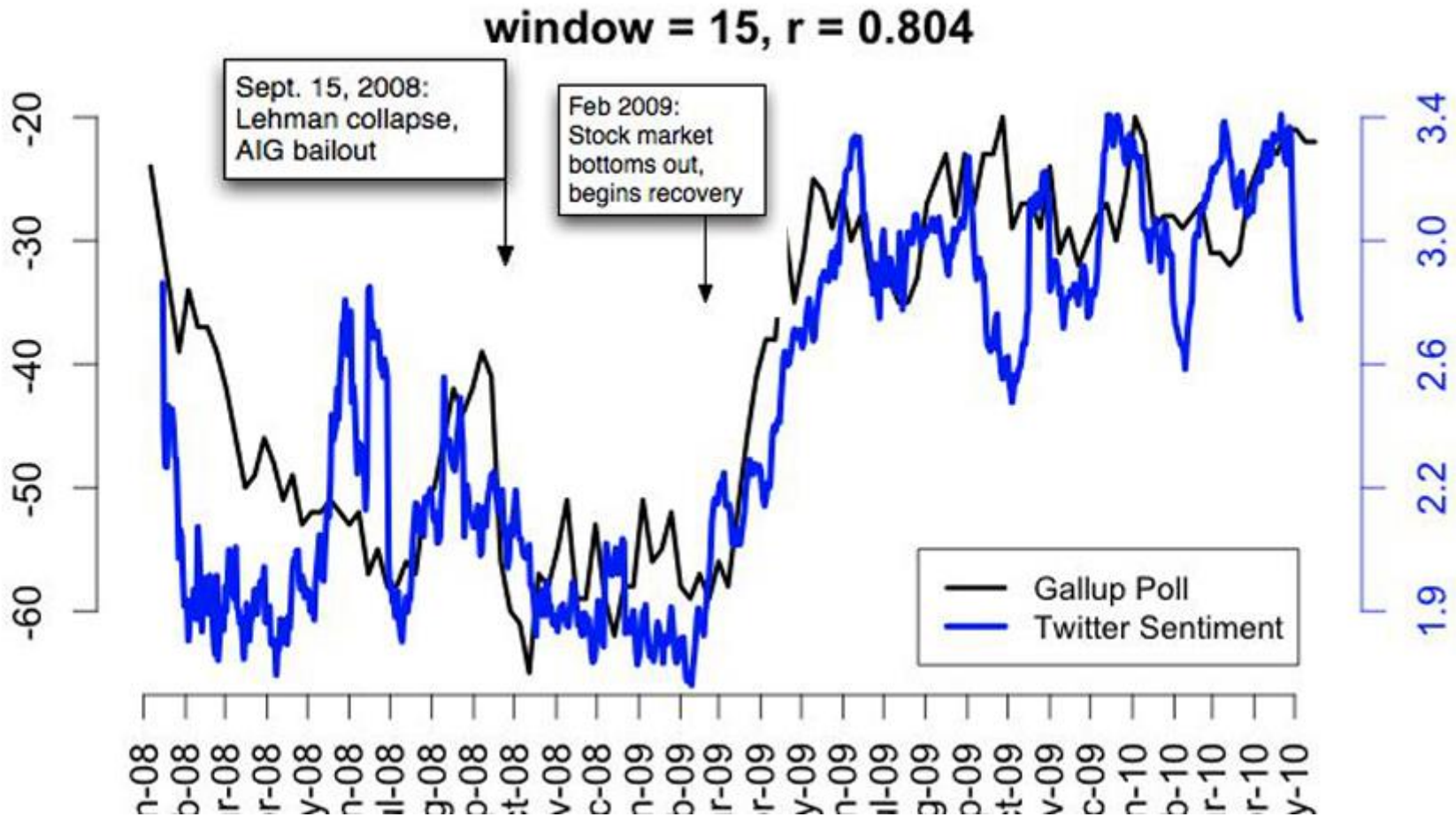


Topic/Genre Assignment

(as an example of text classification)

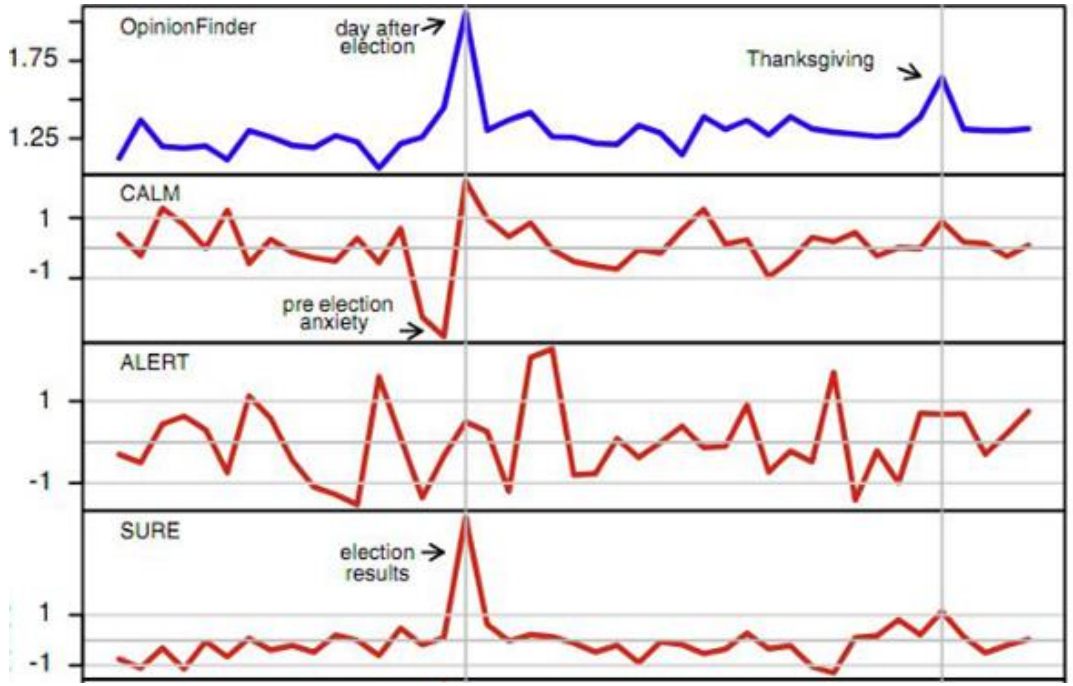
Applications of sentiment classifier in other fields

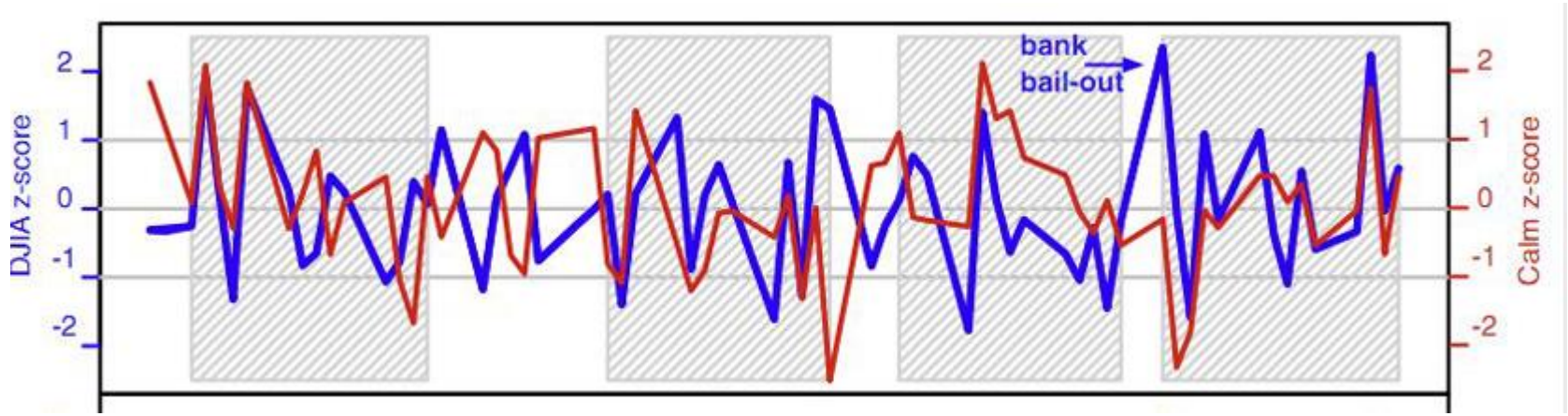
Twitter Sentiment vs Gallup Poll



O'Connor et al, From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, ICWSM 2010

Public Sentiment to Events and Stock Price

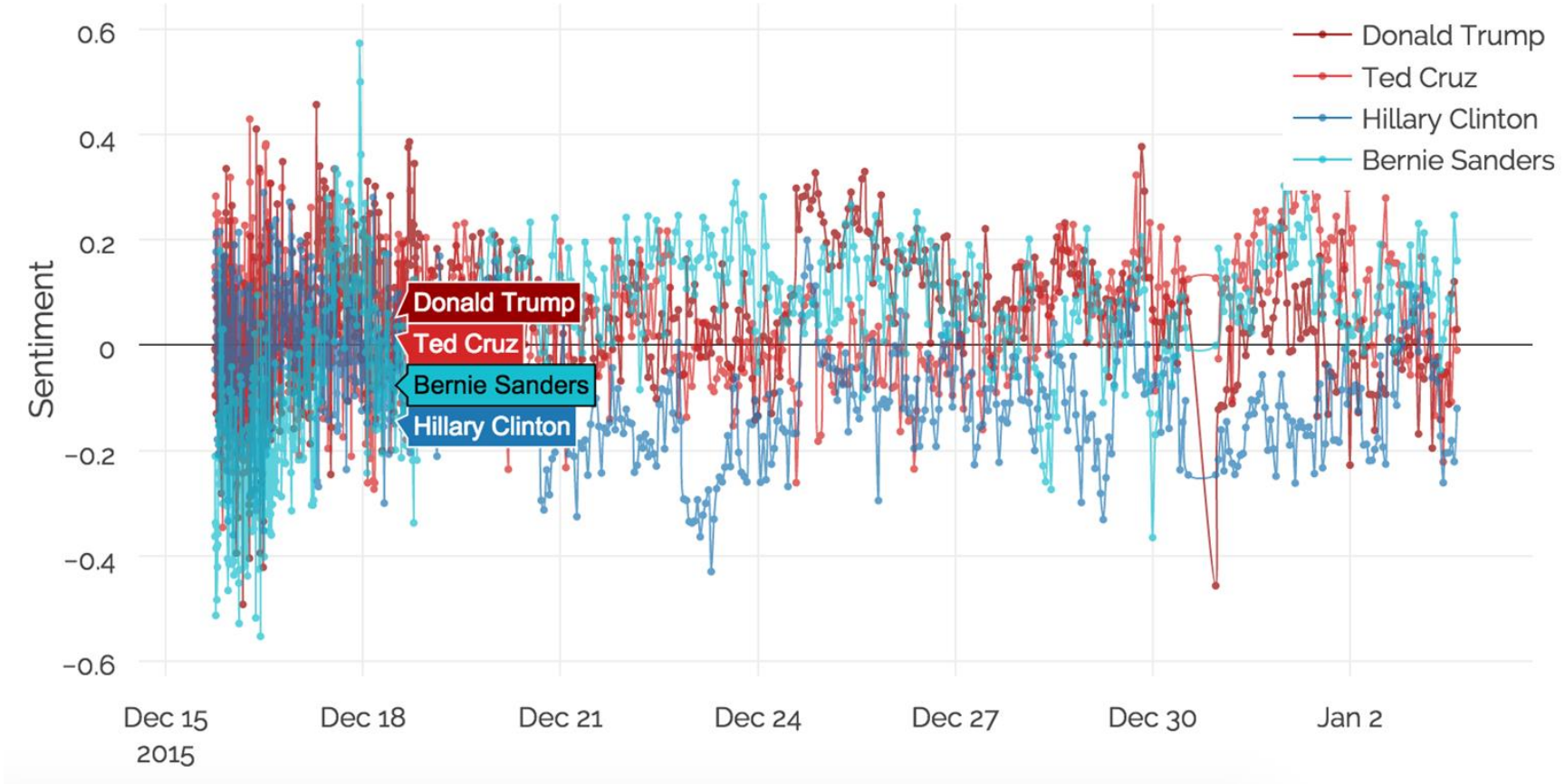




Dow Jones
CALM sentiment

Bollen et al, Twitter mood predicts the stock market, Journal of Computational Science, 2010

Twitter Sentiment vs 2016 Presidential Election



https://rampages.us/giny/2017/03/05/twitter-research

Why is sentiment analysis difficult?

There was an earthquake in California

The team failed to complete physical challenge. (We win/lose!)

They said it would be great.

They said it would be great, and they were great.

They said it would be great, and they were wrong.

Oh, you're terrible!

Long-suffering fans, bittersweet memories, hilariously embarrassing moments

Examples from Chris Potts

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous

Difficulty of task

- Simplest task:
  - Is the attitude of this text positive or negative (or neutral)?
- More complex:
  - Rank the attitude of this text from 1 to 5
- Advanced:
  - Detect the target (stance detection)
  - Detect source
  - ..

What makes reviews hard to classify?

Subtlety

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

Perfume review in *Perfumes: the Guide*

What makes reviews hard to classify?

Thwarted expectations and ordering effects

"This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.."

"Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised."

Why is sentiment analysis hard ?

- Sentiment is a measure of a speaker's private state, which is *unobservable*.
- Sentiment is contextual;
  - Words are a good indicator of sentiment (love, hate, terrible); but many times it requires deep world + contextual knowledge

> "*Valentine's Day* is being marketed as a Date Movie. I think it's more of a First-Date Movie. If your date likes it, do not date that person again. And if you like it, there may not be a second date."
>
> Roger Ebert, Valentine's Day

- Deep understanding of language behaviors (e.g., politeness)

Related Tasks

- Subjectivity (Pang & Lee 2008)
- Stance (Anand et al., 2011)
- Hate-speech (Nobata et al., 2016)
- Sarcasm (Khodak et al., 2017)
- Deception and betrayal (Niculae et al., 2015)
- Online trolls (Cheng et al., 2017)
- Politeness (Danescu-Niculescu-Mizil et al., 2013)
- …

How can we build a sentiment classifier?

Supervised Learning

- Given training data in the form of <x, y> pairs, learn **f(x)**

X	Y
I loved it!	Positive
Terrible movie.	Negative
Not too shabby	Positive
Such a lovely movie!	Positive

Learning f(x)

Two components:

- The formal structure of the learning method:
  - How x and y are mapped
  - Logistic regression, Naïve Bayes, RNN, CNN, etc
- The representation of the data (x)

Representation of data (x)

- Only positive/negative words in sentiment dictionaries
- Only words in isolation
- Conjunctions of words
- Linguistic structures
- ..

Sentiment Dictionaries

- General Inquirer (1996)
- MPQA subjectivity lexicon (Wilson et al., 2005)
  - http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- LIWC (Pennebaker 2015)
- AFINN (Nielsen 2011)
- NRC Word-Emotion Association Lexicon (EmoLex) (Mohmmad and Turney, 2013)

Positive	Negative
unlimited	lag
prudent	contortions
superb	fright
closeness	lonely
impeccably	tenuously
fast-paced	plebeian
treat	mortification
destined	outrage
blessing	allegations

Dictionary Counting

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

Positive	Negative
unlimited	lag
prudent	contortions
superb	fright
closeness	lonely
impeccably	tenuously
fast-paced	plebeian
treat	mortification
destined	outrage
blessing	allegations

happy	1
love	2
recommend	2

lonely	0
outrage	0
not	2

Limitation?

$f($

happy	1
love	2
recommend	2
lonely	0
outrage	0
not	2

$) = y$

Representation of data (x)

- Only positive/negative words in sentiment dictionaries
- Only words in isolation (bag-of-words)
  - E.g., good, bad
- Conjunctions of words (sequential, high-order n-grams, skip n-grams, etc)
  - E.g., “not good”, “not bad”
- Linguistic structures (Part-of-speech, etc)
- ..

Bag of words

Representation of text only as the counts of words that it contains

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

$f($

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

$)$ $= y$

Representation of data (x)

- Only positive/negative words in sentiment dictionaries
- Only words in isolation (bag-of-words)
- Conjunctions of words (sequential, high-order n-grams, skip n-grams, etc)
- Linguistic structures (Part-of-speech, etc)
- ..

$$f\left(\begin{array}{|c|c|}\hline \text{NP} & 5 \\ \hline \text{VP} & 2 \\ \hline \text{Parse depth} & 5 \\ \hline \end{array}\right) = y$$

How to implement f(x)=y using Python?

Two components:

- The formal structure of the learning method:
  - How x and y are mapped
  - Logistic regression, Naïve Bayes, RNN, CNN, etc
- The representation of the data (x)

Tutorial on building text classifier using Scikit-Learn