Business Analytics Using Python

Sentiment Analytics

Cyrus Lentin

What is Sentiment?

- Sentiment = Feelings

What is Sentiment?

- Sentiment = Feelings

- Attitudes
- Emotions
- Opinions

What is Sentiment?

- Sentiment = Feelings

- Attitudes
- Emotions
- Opinions

- Subjective
- No Rational
- Will Differ From Person To Person
- Not Facts

What is Sentiment?

- Sentiment = Feelings

- Attitudes
- Emotions
- Opinions

- Subjective
- No Rational
- Will Differ From Person To Person
- Not Facts

- Sentiment Analysis Are Machine Learning Methods To Extract, Identify, Or Otherwise Characterize The Sentiment Content Of A Text Unit

What is Sentiment?

- Sentiment = Feelings

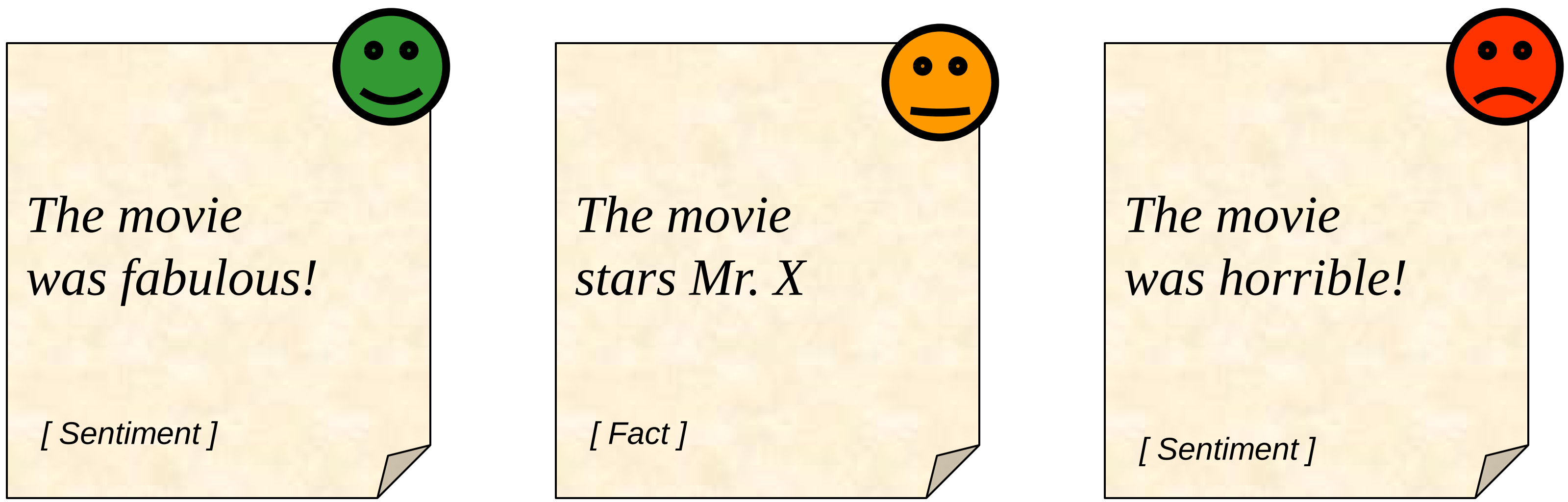
- Attitudes
- Emotions
- Opinions

- Subjective
- No Rational
- Will Differ From Person To Person
- Not Facts

- Sentiment Analysis Are Machine Learning Methods To Extract, Identify, Or Otherwise Characterize The Sentiment Content Of A Text Unit
- Sometimes Also Referred To As Opinion Mining, Which Is Computational Study Of Opinions (Sentiments, Emotions) Expressed In Text

What is Sentiment?

- In Others Words Determine If A Sentence Or A Document Expresses Positive, Negative, Neutral Sentiment Towards Some Object?



Why Opinion Mining Now? Because The Web Contains Huge Volumes Of Opinionated Text

Applications

- Product Acceptance
- Brand Perception
- Reputation Management
- Customer Satisfaction
- Flame Detection (Bad Rants)
- Influencers
- Child-suitability Identification
- News Classification
- (In)appropriate Content Identification

Challenges

- How does a machine define subjectivity & objectivity?
- How does a machine analyze polarity (negative / positive)?
- How does a machine know about polarity intensity?
- How does a machine analyze emotions (happy / sad / etc…)?

Language Is Ambiguous – 1

- The Watch Isn't Water Resistant
  [In A Product Review This Could Be Negative]
- Hit The Nail On The Head
  [Use Of Phrases Have Difference Meaning]
- Low Price / Low Quality
  [Sentiment Changes With Accompanying Word]
- The Canon Camera Is Better Than The Fisher Price One
  [Comparisons Are Hard To Classify]
- The Ice Cream Is Luuuvvvvveeeely
  [Slangs]
- IMHO …. / LOL / FO
  [Abbreviations]
- That Won't Do You No Good
  [Double Negative]
- Not Good / Not Bad
  [Flipped Sentiment]
- Got Up And Walked Out
  [No Sentiment From Word]

Language Is Ambiguous – 2

- I Do Not Dislike Cabin Cruisers
  [Negation Handling]
- Disliking Watercraft Is Not Really My Thing
  [Negation, Inverted Word Order]
- Sometimes I Really Hate Ribs
  [Adverbial Modifies The Sentiment]
- I'd Really Truly Love Going Out In This Weather!
  [Possibly Sarcastic]
- Chris Craft Is Better Looking Than Limestone
  [Two Brand Names, Identifying Target Of Attitude Is Difficult]
- Chris Craft Is Better Looking Than Limestone, But Limestone Projects Are Seaworthy And Reliable
  [Two Attitudes, Two Brand Names]
- The Movie Is Surprising With Plenty Of Unsettling Plot Twists
  [Negative Term Used In A Positive Sense In Certain Domains]
- I Love My Mobile But Would Not Recommend It To Any Of My Colleagues
  [Qualified Positive Sentiment, Difficult To Categorize]
- Next Week's Gig Will Be Right Koide9!
  [Newly Coined Terms Can Be Strong In Polarity But Out Of Known Vocabulary]

Classification Methods

- Baseline Classification
- Bayes Classification
- Entropy Classification
- Ngram Classification
- Support Vector Machine

Baseline Classification

- Baseline Approach Is To Use A Dictionary (List) Of Positive And Negative Keywords
- You May Either Use A List Of Keywords, Which Is Publicly Available Or Create Your Own
- For Each Record / Line / Unit Of Words, We Count The Number Of Negative Keywords And Positive Keywords That Appear
- The Classifier Returns The Polarity With The Higher Count
- If There Is A Tie, Then Neutral Polarity (The Majority Class) Is Returned

Enhancements

- Rather Than Just Checking Count Of Words, We Assign Weights To Each Word Based On Past Training
- The Classifier Returns The Polarity With The Higher Weighted Score
- If There Is A Tie, Then Neutral Polarity (The Majority Class) Is Returned

Baseline Classification – Simple

- Pos Dict: good, better, best, wonderful
- Neg Dict: bad, worse, worst, horrible

- Today is a good day

  Pos Dict Score: 1

  Neg Dict Score: 0

  Positive
- Today is a bad day

  Pos Dict Score: 0

  Neg Dict Score: 1

  Negative
- Today is a Monday

  Pos Dict Score: 0

  Neg Dict Score: 0

  Neutral

Baseline Classification – Weighted

- Pos Dict: good(1), better (3), best (5), wonderful (5)
- Neg Dict: bad(1), worse(3), worst(5), horrible (5)

- Today was a wonderful day

  Pos Dict Score: 5

  Neg Dict Score: 0

  Positive
- Today is a horrible day

  Pos Dict Score: 0

  Neg Dict Score: 5

  Negative
- The movie was bad but the acting was wonderful

  Pos Dict Score: 5

  Neg Dict Score: 1

  Positive

Today Is Monday

Pos Dict Score: 0

Neg Dict Score: 0

Neutral

Bayes Classification

- Statistical method for classification
- Supervised Learning Method
- Assumes an underlying probabilistic model, the Bayes theorem
- Can solve problems involving both categorical and continuous valued attributes
- Named after Thomas Bayes, who proposed the Bayes Theorem
- Bayes approach is to use a dictionary (list) of positive and negative keywords
- You may either use a list of keywords, which is publicly available or create your own
- For each record / line / unit of words, we compute the probability of the words-in-the-text appearing in the negative keyword list and positive keyword list
- The classifier returns the polarity with the higher probability

Bayes Theorem

Given a hypothesis h and data D, the following is are the probabilities:

- P(h): prior probability
  independent probability of hypothesis h
- P(D): independent probability
  independent probability of data D
- P(D|h): likelihood
  the statistical probability of the data D for given hypothesis h
- P(h|D): posterior probability
  the statistical probability that a hypothesis h is true calculated in the light of relevant data D

The following is formula to calculate the posterior probability:

$$P(h \mid D) = \frac{P(D \mid h)\, P(h)}{P(D)}$$

Bayes Classification – How It Works – A Language Model

- In Order To Understand The Process, We Will Use An Example With Small Number Of Posts

Type	Post Text	Class
Training	good happy good	Positive
Training	good good service	Positive
Training	good friendly	Positive
Training	lousy good cheat	Negative
Test	good good good cheat lousy	???

What Was The Problem / Question?

- We Are Trying To Determine Whether The Class For Last Post Is Positive Or Negative.
- So In Effect, We Want To Compute:
- P(Pos|good good good cheat lousy)
- By Bayes Theorem, This Is Equal To:

$$\frac{P(\text{Pos}) * P(\text{good good good cheat lousy} \mid \text{Pos})}{P(\text{good good good lousy cheat})}$$

Bayes Classification – How It Works – A Language Model

- By Bayes Theorem, This Is Equal To:

$$\frac{P(Pos) * P(good\ good\ good\ cheat\ lousy\ |Pos)}{P(good\ good\ good\ lousy\ cheat)} \qquad \frac{P(Pos) * P(good|Pos)^3 * P(cheat|Pos) * P(lousy\ |Pos)}{P(good\ good\ good\ lousy\ cheat)}$$

- P(Pos) = 3/4
- P(Neg) = 1/4

- P(good|Pos) = (5)/(8) = 5/8
- P(cheat|Pos) = (0)/(8) = 0
- P(lousy|Pos) = (0)/(8) = 0
- P(good|Neg) = (1)/(3) = 1/3
- P(cheat|Neg) = (1)/(3) = 1/3
- P(lousy|Neg) = (1)/(3) = 1/3

Type	Document Text	Class
Training	good happy good	Positive
Training	good good service	Positive
Training	good friendly	Positive
Training	lousy good cheat	Negative
Test	good good good cheat lousy	???

- However, this would break as soon as we encounter a word that isn't in our training set?
- For example, if “goood” is not in our training set, and occurs in our test set, then since
- P(goood|Pos) = 0, so our product is zero for all classes

Bayes Classification – How It Works – A Language Model

- By Bayes Theorem, This Is Equal To:

$$\frac{P(Pos) * P(good\ good\ good\ cheat\ lousy\ |Pos)}{P(good\ good\ good\ lousy\ cheat)} \qquad \frac{P(Pos) * P(good|Pos)^3 * P(cheat|Pos) * P(lousy\ |Pos)}{P(good\ good\ good\ lousy\ cheat)}$$

- We need nonzero probabilities for all words, even words that don't exist
- **Introducing +1 Smoothing**
- Just count every word one time more than it actually occurs
- Since we are only concerned with relative probabilities, this slight inaccuracy should not be a problem

$$P(word|C) = \frac{count(word|C) + 1}{count(C) + V}$$

- Where V is the total vocabulary, so that our probabilities sum to 1

Bayes Classification – How It Works – A Language Model

Type	Document Text	Class
Training	good happy good	Positive
Training	good good service	Positive
Training	good friendly	Positive
Training	lousy good cheat	Negative
Test	good good good cheat lousy	???

- P(Pos) = 3/4
- P(Neg) = 1/4

- P(good|Pos) = (5+1)/(8+5+1) = 3/7
- P(cheat|Pos) = (0+1)/(8+5+1)= 1/14
- P(lousy|Pos) = (0+1)/(8+5+1) = 1/14
- P(good|Neg) = (1+1)/(3+5+1) = 2/9
- P(cheat|Neg) = (1+1)/(3+5+1) = 2/9
- P(lousy|Neg) = (1+1)/(3+5+1) = 2/9

- P(Pos|D5) = P(Pos) * P(good|Pos)^3 * P(cheat|Pos) * P(lousy |Pos)
- P(Pos|D5) = 3/4 * (3/7)^3 * (1/14) * (1/14) = 0.0003
- P(Neg|D5) = P(Neg) * P(good| Neg)^3 * P(cheat| Neg) * P(lousy | Neg)
- P(Neg|D5) = 1/4 * (2/9)^3 * (2/9) * (2/9) = 0.0001
- Prediction: Positive

Issues – Tokenization

- Use Only Whitespace To Tokenize?

  "Food", "Food.", Food," And "Food!" All Different.
- Use Whitespace And Punctuation To Tokenize?

  "Won't" Tokenized To "Won" And "T"
- What About Emails? Urls? Phone Numbers?
- What About The Things We Haven't Thought About Yet?
- Don't Re-Invent The Wheel; Use A Library!

Tokenization Strategies

- Stop Words
- Sparse Words
- Profanity
- Remove Punctuations
- Consistent Case
- Stemming

Issues – Arithmetic

- What Happens When You Multiply A Large Amount Of Small Numbers?

  Very Small Number
- To Prevent Underflow, Use Sums Of Logs Instead Of Products Of True Probabilities.
- Key Properties Of Log:

  Log(ab) = Log(a) + Log(b)

  X > Y => Log(x) > Log(y)
- Turns Very Small Numbers Into Manageable Negative Numbers

Training The Classifier

- Do We Need To Build The Vocabulary Of All Distinct Words That Appear In All The Documents Of The Training Set.
- Do We Need To Build A Bag Of Words That Appear In All The Documents Of The Training Set.

Training The Classifier

- Do We Need To Build The Vocabulary Of All Distinct Words That Appear In All The Documents Of The Training Set.
- Do We Need To Build A Bag Of Words That Appear In All The Documents Of The Training Set.

NO

- We Use Ready Made Functions From NLTK & TextBlob Library
- We Need To Interpret The Results

Sentiment Analytics Tasks

- Document Pre-processing
  - Tokenization (Word Or Sentenance)
- Document Cleaning
  - Special Texts
  - Punctuations
  - Digits
  - Stemming / Lemmetization
- Document Cleaning
  - Stop Words
  - Profanities
  - Sparse Words
- Classification
  - Polarity
  - Subjectivity
  - Emotions
- Visualization
  - Polarity Frequency Distribution
  - Emotions Frequency Distribution

NLTK Results Interpretation – Polarity

```
*** Test 1 - Positive Polarity ***
Today I am very happy
{'neg': 0.0, 'neu': 0.429, 'pos': 0.571, 'compound': 0.6115}

*** Test 2 - Negative Polarity ***
Today is a bad day
{'neg': 0.538, 'neu': 0.462, 'pos': 0.0, 'compound': -0.5423}

*** Test 2 - Neutral Polarity ***
The board is clean
{'neg': 0.0, 'neu': 0.526, 'pos': 0.474, 'compound': 0.4019}
```

- How To Classify The Result?
- Polarity Options
  - Positive
  - Negative
  - Neutral

TextBlob Results Interpretation – Polarity

```
*** Test 1 - Positive Polarity ***
Today I am very happy
Sentiment(polarity=1.0, subjectivity=1.0)

*** Test 2 - Negative Polarity ***
Today is a bad day
Sentiment(polarity=-0.6999999999999998, subjectivity=0.6666666666666666)

*** Test 2 - Neutral Polarity ***
The board is clean
Sentiment(polarity=0.3666666666666667, subjectivity=0.7000000000000001)
```

- How To Classify The Result?
- Polarity Options
  - Positive
  - Negative
  - Neutral

TextBlob Results Interpretation – Subjectivity

```
*** Test 1 - Positive Polarity ***
Today I am very happy
Sentiment(polarity=1.0, subjectivity=1.0)

*** Test 2 - Negative Polarity ***
Today is a bad day
Sentiment(polarity=-0.6999999999999998, subjectivity=0.6666666666666666)

*** Test 2 - Neutral Polarity ***
The board is clean
Sentiment(polarity=0.3666666666666667, subjectivity=0.7000000000000001)
```

- How To Classify The Result?
- Subjectivity Options
  - Objective – fact
  - Subjective – opinion
  - Neutral – undecided / not clear

Text2Emotions Results Interpretation – Emotions

```
*** Test 1 - Happy Emotion ***
Today I am very happy
{'Happy': 1.0, 'Angry': 0.0, 'Surprise': 0.0, 'Sad': 0.0, 'Fear': 0.0}

*** Test 2 - Sad Emotion ***
Today I am very sad
{'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, 'Sad': 1.0, 'Fear': 0.0}

*** Test 3 - Neutral Emotion ***
The board is clean
{'Happy': 0, 'Angry': 0, 'Surprise': 0, 'Sad': 0, 'Fear': 0}
```

- How To Classify The Result?
- Emotions Options
  - Happy
  - Angry
  - Surprise
  - Sad
  - Fear
  - Neutral – undecided / not clear

Wind Up

- Sentiment analysis is a difficult task
- The difficulty increases with the nuance and complexity of opinions expressed
- Product reviews, etc are relatively easy
- Books, movies, art, music are more difficult
- Policy discussions, indirect expressions of opinion more difficult still
- Non-binary sentiment (political leanings etc) is extremely difficult
- Patterns of alliance and opposition between individuals become central

Thank you!

Contact:

Cyrus Lentin
cyrus@lentins.co.in
+91-98200-94236