

CS 6501 Natural Language Processing

Text Classification

Yangfeng Ji

September 3, 2018

Department of Computer Science
University of Virginia



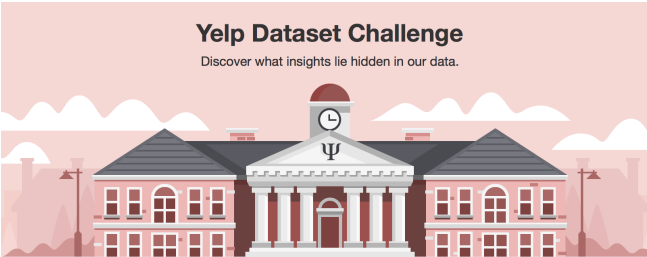
ENGINEERING

Case I: Like a business?



[Pang et al., 2002]

Potential Group Project



Yelp Dataset Challenge

Discover what insights lie hidden in our data.

What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the [past winners](#) and [hundreds of academic papers written](#) using the dataset.

Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Case II: Topic Classification



Example topics

- ▶ Business
- ▶ Arts
- ▶ Technology
- ▶ Sports
- ▶ ...

Google News Category

Top stories

For you

Favorites

Saved searches

U.S.

World

Local

Business

Technology

Entertainment

Sports

Science

Health

Language & region
English | United States

Headlines

Why saying Roe v. Wade is 'settled' isn't saying much

CNN • 2 hours ago

- Democratic Senator Says Brett Kavanaugh's Confirmation Process Is 'Not Normal'
HuffPost • today
- Trump admin withholds 100000-plus pages of Kavanaugh docs
CNN • today
- Trump couldn't care less about Kavanaugh's judicial philosophy
The Washington Post • today • Opinion
- Kavanaugh is not anti-women – Democrats, Planned Parenthood are insulting Americans' intelligence
Fox News • today

View full coverage



Special Report: How Myanmar punished two reporters for uncovering an atrocity

Reuters • 2 hours ago

- Myanmar: Reuters journalists investigating Rohingya killings sentenced to 7 years in prison
CNN • 3 hours ago

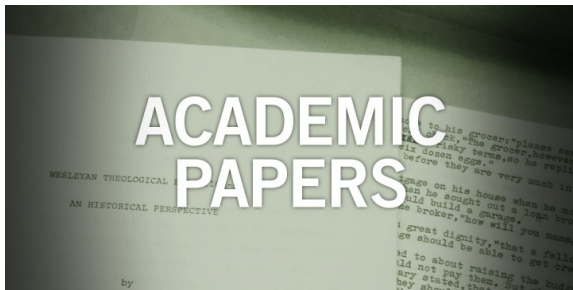
View more

More Headlines



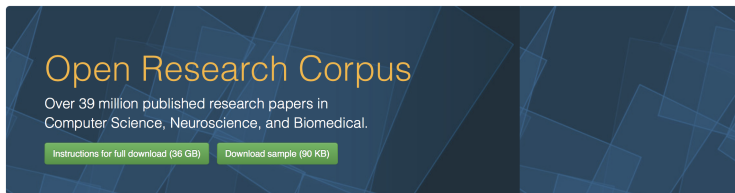
4

Case III: Scientific Literature Analysis



- ▶ Claim
- ▶ Arguments
- ▶ Good writing?

Open Research Corpus



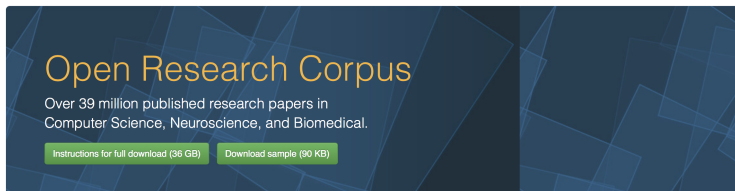
<https://labs.semanticscholar.org/corpus/>

Data:

- ▶ ~ 39M papers
- ▶ ~ 12M authors

[Ammar et al., 2018]

Potential Group Projects



<https://labs.semanticscholar.org/corpus/>

- ▶ Citation recommendation based on paper abstract
- ▶ Relation analysis on different machine learning methods
- ▶ Influence analysis of machine/deep learning on biomedical domain

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Classification

- ▶ Input: a text x
- ▶ Output: $y \in \mathcal{Y}$, where \mathcal{Y} is the predefined category set



Mathematical Formulation

- ▶ Input: a **numeric representation** x of text d
- ▶ Output: **scores** $\Psi(x, y; \theta) \in \mathbb{R}, \forall y \in \mathcal{Y}$

Mathematical Formulation

- ▶ Input: a **numeric representation** x of text d
- ▶ Output: **scores** $\Psi(x, y; \theta) \in \mathbb{R}, \forall y \in \mathcal{Y}$

Classification

$$\hat{y} = \arg \max_{y' \in \mathcal{Y}} \Psi(x, y'; \theta) \quad (1)$$

Key Questions

$$\hat{y} = \arg \max_{y' \in \mathcal{Y}} \Psi(x, y'; \theta) \quad (2)$$

1. How to represent a text as x ?
2. How to formulate score function $\Psi(x, y; \theta)$

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Sentiment Analysis

Task: predicting user sentiment polarity (POSITIVE or NEGATIVE) based on a review



The screenshot shows a Yelp interface with a red header containing the Yelp logo. Below the header, the title "Recommended Reviews" is displayed in red. To the right of the title is a search bar labeled "Search reviews" and a red magnifying glass icon. Below the title, there are filters: "Yelp Sort", "Date", "Rating", "Elites", and "English 16". The main content area shows a review by "Jenn P." from "San Francisco, CA". The reviewer's profile picture is a small square image of a person. To the right of the name, it says "1 friend" and "22 reviews". The review itself is dated "10/17/2013" and has a rating of 5 stars, represented by five red star icons. The text of the review is: "Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place."

yelp

Recommended Reviews 

Yelp Sort Date Rating Elites English 16

 **Jenn P.**
San Francisco, CA
1 friend
22 reviews

★★★★★ 10/17/2013

Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place.

Twitter Sentiment Analysis



Donald J. Trump ✓

@realDonaldTrump

Follow



Happy Labor Day! Our country is doing better than ever before with unemployment setting record lows. The U.S. has tremendous upside potential as we go about fixing some of the worst Trade Deals ever made by any country in the world. Big progress being made!

4:28 AM - 3 Sep 2018

A Simple Predictor

Example 1: POSITIVE

*Super quick and really **friendly** staff. I **like** starting off my mornings at this store!!*

- ▶ SentiWordnet: a publicly available word sentiment polarity dictionary.

Another Example

Example II: POSITIVE

Din Tai Fung, every time I go eat at anyone of the locations around the King County area, I keep being reminded on why I have to keep coming back to this restaurant.

...

- ▶ No signal word

Counting Words Can Be Risky



Donald J. Trump ✓

@realDonaldTrump

Follow



Happy Labor Day! Our country is doing better than ever before with unemployment setting record lows. The U.S. has tremendous upside potential as we go about fixing some of the worst Trade Deals ever made by any country in the world. Big progress being made!

4:28 AM - 3 Sep 2018

Data Driven Approach

★★★★★ 4/25/2018

⚙️ 1 check-in

Din Tai Fung, every time I go eat at anyone of the locations around the King County area I keep being reminded on why I have to keep coming back. I planned an outing for my sister and I so I can take her to some place to eat she hasn't been to before. I wasn't sure where but DTF popped in my head immediately and BAM. We ended up here and so satisfied.

- ▶ Discover the relationship between **words** and **sentiment polarity** from data
- ▶ Need a collection of texts and their category labels $\{(x^{(i)}, y^{(i)})\}$

Basic Idea of Statistical Machine Learning

Given $\{(\mathbf{x}^{(i)}, y^{(i)})\}$

- ▶ Principle: Discover the **statistical** relationship between patterns and categories from training data.
- ▶ Goal: To make better decisions for **unseen** data points in a *test* set (generalization power)

Standard Setup of Statistical Machine Learning

- ▶ Training set $\mathcal{T} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$
- ▶ Test set $\mathcal{U} = \{(\mathbf{x}^{(l)}, y^{(l)})\}_{l=1}^L$

Standard Setup of Statistical Machine Learning

- ▶ Training set $\mathcal{T} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$
- ▶ Development set $\mathcal{D} = \{(\mathbf{x}^{(j)}, y^{(j)})\}_{j=1}^M$
- ▶ Test set $\mathcal{U} = \{(\mathbf{x}^{(l)}, y^{(l)})\}_{l=1}^L$

Simple Framework

$$\hat{y} = \arg \max_{y' \in \mathcal{Y}} \Psi(x, y'; \theta) \quad (3)$$

Score function

$$\Psi(x, y; \theta) = \theta^\top f(x, y) \quad (4)$$

- ▶ $f(x, y)$: feature function
- ▶ θ : classification weights

Questions

Score function

$$\Psi(x, y; \theta) = \theta^\top f(x, y) \quad (5)$$

- ▶ How to design a feature function $f(x, y)$?
 - ▶ Bag-of-words representation
- ▶ How to learn θ ?
 - ▶ Perceptron algorithm

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Bag-of-Words Representation

From texts to bag of words:

*Super quick and really
friendly staff. I like starting off
my mornings at this store!!* \Rightarrow SUPER,QUICK,AND, REALLY,
FRIENDLY, \dots , STORE

NLTK function

```
nltk.tokenize.wordpunct_tokenize
```

Given the texts from training set, build a vocab first

$$\left\{ \begin{array}{c} \text{SUPER} \\ \dots \\ \text{QUICK} \\ \text{FOOD} \\ \text{FRIENDLY} \\ \text{EAT} \\ \dots \\ \text{STAFF} \end{array} \right\} \quad (6)$$

Feature Function

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

$$\left\{ \begin{array}{l} \langle \text{SUPER}, -1 \rangle \\ \langle \text{SUPER}, 1 \rangle \\ \dots \\ \langle \text{FRIENDLY}, -1 \rangle \\ \langle \text{FRIENDLY}, 1 \rangle \\ \langle \text{EAT}, -1 \rangle \\ \langle \text{EAT}, 1 \rangle \\ \dots \\ \langle \text{DELICIOUS}, -1 \rangle \\ \langle \text{DELICIOUS}, 1 \rangle \end{array} \right\}$$

Feature Function

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

$$\left\{ \begin{array}{c} \langle \text{SUPER}, -1 \rangle \\ \langle \text{SUPER}, 1 \rangle \\ \dots \\ \langle \text{FRIENDLY}, -1 \rangle \\ \langle \text{FRIENDLY}, 1 \rangle \\ \langle \text{EAT}, -1 \rangle \\ \langle \text{EAT}, 1 \rangle \\ \dots \\ \langle \text{DELICIOUS}, -1 \rangle \\ \langle \text{DELICIOUS}, 1 \rangle \end{array} \right\} \quad \left[\begin{array}{c} 0 \\ 1 \\ \dots \\ 0 \\ 1 \\ 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{array} \right] = f(x, y)$$

Preprocessing for Building Vocab

1. convert all characters to lowercase

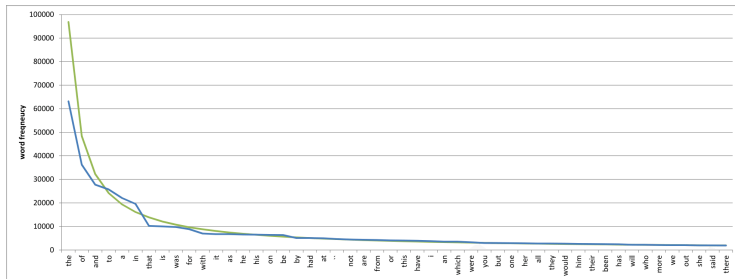
UVa, UVA \rightarrow uva

Preprocessing for Building Vocab

1. convert all characters to lowercase

$$UVa, UVA \rightarrow uva$$

2. map low frequency words to a special token UNK



$$\text{Zipf's law: } f(w_t) \propto 1/r_t$$

Information Embedded in BoW Representation

- ▶ Lose:
 - ▶ word order
 - ▶ sentence boundary
 - ▶ paragraph boundary
 - ▶ ...
- ▶ Keep: words in texts

Alternative Formulation

Example I: POSITIVE

Super quick and really friendly staff. I like starting off my mornings at this store!!

Vocab

$\left(\begin{array}{c} \text{SUPER} \\ \dots \\ \text{QUICK} \\ \text{FOOD} \\ \text{FRIENDLY} \\ \text{EAT} \\ \dots \\ \text{STAFF} \end{array} \right)$

$f(x)$

θ_y

Alternative Formulation (Cont.)

$$\left\{ \begin{array}{c} \langle \text{SUPER}, -1 \rangle \\ \langle \text{SUPER}, 1 \rangle \\ \dots \\ \langle \text{FRIENDLY}, -1 \rangle \\ \langle \text{FRIENDLY}, 1 \rangle \\ \langle \text{EAT}, -1 \rangle \\ \langle \text{EAT}, 1 \rangle \\ \dots \\ \langle \text{DELICIOUS}, -1 \rangle \\ \langle \text{DELICIOUS}, 1 \rangle \end{array} \right\}$$

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Question

How to learn θ from training examples?

(Online) Supervised Learning

Given a training example (x, y)

- ▶ Predict \hat{y} as

$$\hat{y} = \arg \max_{y'} \theta^\top f(x, y')$$

- ▶ If $y \neq \hat{y}$, **update** θ

Perceptron Algorithm: Updating rule

If $\hat{y} \neq y$:

$$\begin{aligned} \theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) &\rightsquigarrow \text{ground truth} \\ - f(x, \hat{y}) &\rightsquigarrow \text{predicted label} \end{aligned} \quad (7)$$

[Eisenstein, 2018, Sec. 2.2.1]

Justification

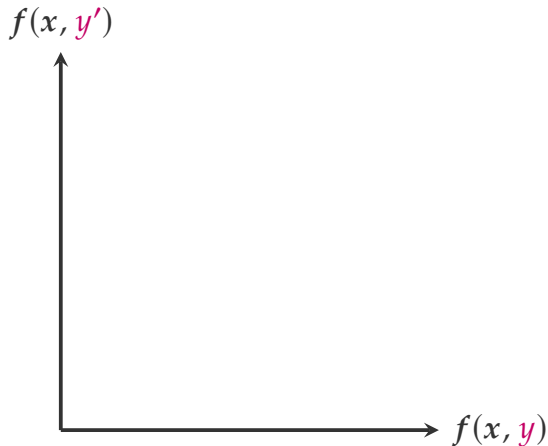
$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) - f(x, \hat{y}) \quad (8)$$

$$\forall y' \in \mathcal{Y}$$

$$\begin{aligned} (\theta^{(\text{new})})^\top f(x, y') &= (\theta^{(\text{old})})^\top f(x, y') \\ &+ (f(x, y))^\top f(x, y') \\ &- (f(x, \hat{y}))^\top f(x, y') \end{aligned}$$

Geometric Interpretation

$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) - f(x, \hat{y}) \quad (9)$$



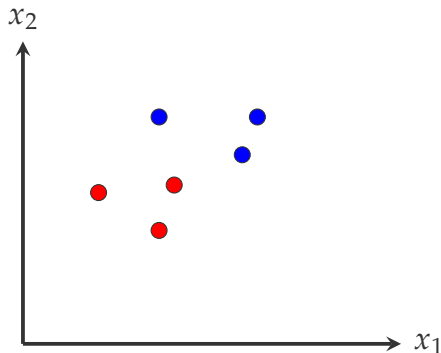
Algorithm Overview

Algorithm 3 Perceptron learning algorithm

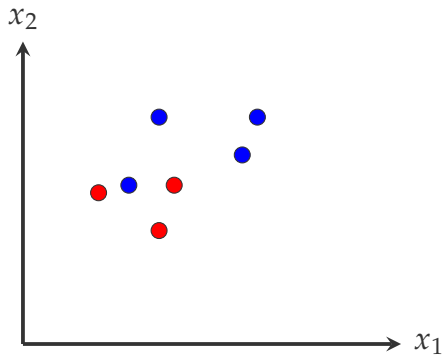
```
1: procedure PERCEPTRON( $\mathbf{x}^{(1:N)}, y^{(1:N)}$ )
2:    $t \leftarrow 0$ 
3:    $\boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}$ 
4:   repeat
5:      $t \leftarrow t + 1$ 
6:     Select an instance  $i$ 
7:      $\hat{y} \leftarrow \operatorname{argmax}_y \boldsymbol{\theta}^{(t-1)} \cdot \mathbf{f}(\mathbf{x}^{(i)}, y)$ 
8:     if  $\hat{y} \neq y^{(i)}$  then
9:        $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \mathbf{f}(\mathbf{x}^{(i)}, y^{(i)}) - \mathbf{f}(\mathbf{x}^{(i)}, \hat{y})$ 
10:    else
11:       $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$ 
12:  until tired
13:  return  $\boldsymbol{\theta}^{(t)}$ 
```

Convergence: Separable Case

The algorithm will converge, if the training examples are well separated w.r.t. labels



Non-Separable Case



Averaged Perceptron

In practice, average all the classification weights θ_t across all the iterations

$$\bar{\theta} = \frac{1}{T} \sum_t \theta^{(t)} \quad (10)$$

Then,

$$\Psi(x, y) = \bar{\theta}^\top f(x, y) \quad (11)$$

[Eisenstein, 2018, Sec. 2.2.2]

Overview

1. Problem Definition
2. Case Study: Sentiment Analysis
3. Bag-of-Words Representation
4. Perceptron Algorithm
5. Classification Evaluation

Standard Setup of Statistical Machine Learning

- ▶ Training set $\mathcal{T} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$
- ▶ Development set $\mathcal{D} = \{(\mathbf{x}^{(j)}, y^{(j)})\}_{j=1}^M$
- ▶ Test set $\mathcal{U} = \{(\mathbf{x}^{(l)}, y^{(l)})\}_{l=1}^L$

Evaluation Measurements

- ▶ Accuracy
- ▶ Precision, recall, and F-measure

[Eisenstein, 2018, Sec 4.4]

$$\text{ACC}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \delta(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)}) \quad (12)$$

δ function

$$\delta(y, \hat{y}) = \begin{cases} 1 & y = \hat{y} \\ 0 & y \neq \hat{y} \end{cases} \quad (13)$$

The loss function of perceptron algorithm

True/False Positive/Negative

For a particular category/class, e.g., user reviews with rating 1 on Yelp

		Ground truth	
		POSITIVE	NEGATIVE
Prediction	POSITIVE	True Positive (TP)	False Positive (FP)
	NEGATIVE	False Negative (FN)	True Negative (TN)

Recall, Precision and F Measure

		Ground truth	
		POSITIVE	NEGATIVE
Prediction	POSITIVE	True Positive (TP)	False Positive (FP)
	NEGATIVE	False Negative (FN)	True Negative (TN)

Recall:

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \frac{TP}{TP + FN} \quad (14)$$

Precision:

$$p(\mathbf{y}, \hat{\mathbf{y}}) = \frac{TP}{TP + FP} \quad (15)$$

F measure:

$$F(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2 \cdot p \cdot r}{p + r} \quad (16)$$

Summary

- ▶ Bag-of-words representations
- ▶ Perceptron updating rule

$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + f(x, y) - f(x, \hat{y}) \quad (17)$$

- ▶ Evaluation measurements
 - ▶ Accuracy
 - ▶ Recall, Precision and F-Measure

Reference



Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.-H., Peters, M. E., Power, J., Skjonsberg, S., Wang, L. L., Wilhelm, C., Yuan, Z., van Zuylen, M., and Etzioni, O. (2018). Construction of the literature graph in semantic scholar.
In *NAACL-HTL*.



Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.



Pang, B., Lee, L., and Vaithyanathan, S. (2002).
Thumbs up?: sentiment classification using machine learning techniques.
In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.