CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Business Analytics and Data-Driven Decision Making

Text Analytics - 01: Natural Language Processing and Text Classification

Raghava Mukkamala

Associate Professor & Director,
Centre for Business Data Science
Copenhagen Business School, Denmark
Email: rrm.digi@cbs.dk, Centre: https://cbsbda.github.io/

Many of the slides have been taken and adapted from:
Speech and Language Processing by Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED CEMS PIM

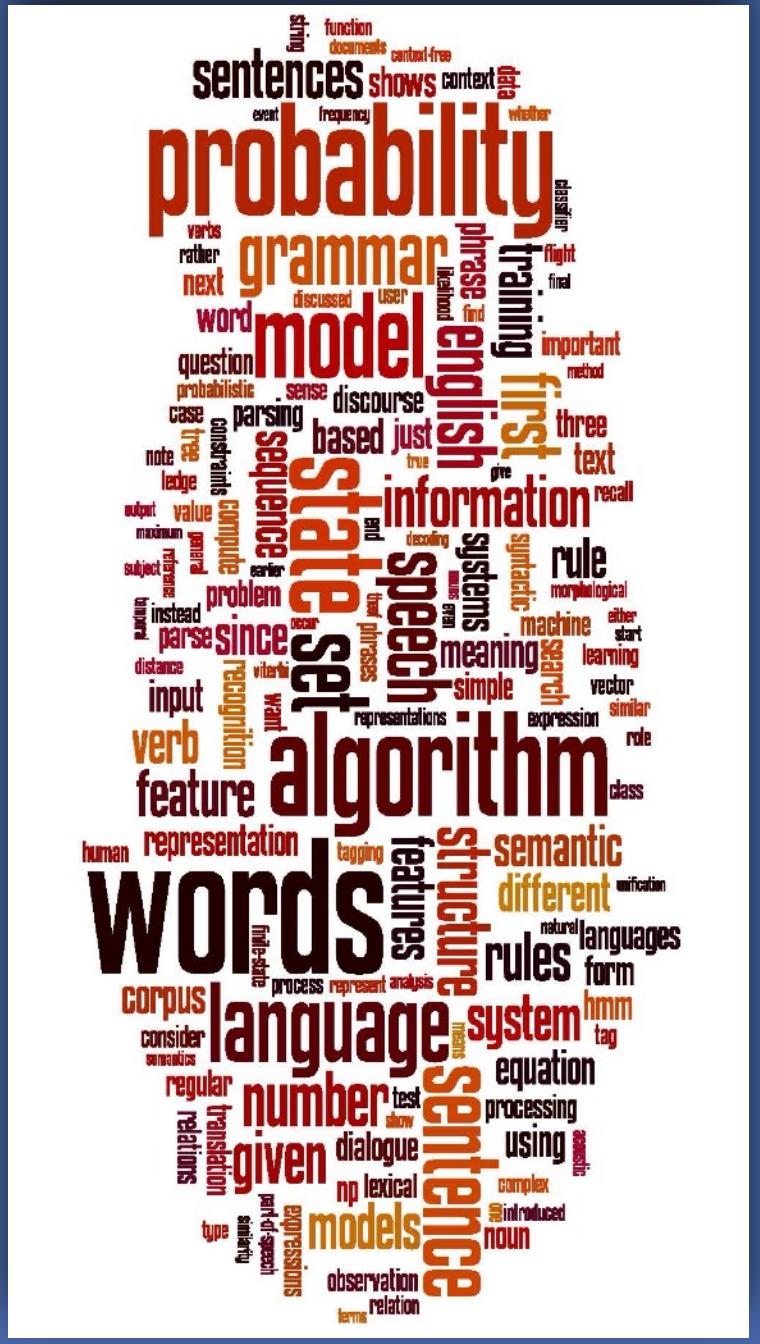
About myself

- 2000-2009: Worked in Danish IT Industry as programmer and consultant
  - Novo Nordisk, Catalog-International, Resultmaker, CSC
  - Security, cryptographic protocols, workflows etc.
  - Programming in C++, C#, Java, Webservices etc.
- 2009-2012: PhD in Theoretical Computer Science in Formal models and programming languages from ITU, Denmark
- 2015- : Asst. and Associate Professor @ CBS
  - 2018: Programme Director for Data Science Masters Programme at CBS
  - 2019: Director, Centre for Business Data Analytics (cbsBDA)
- 2015: - Adjunct Associate Professor at Kristiania University College
- Research: Data Science, NLP, LLMs, Blockchain and Cybersecurity

Outline

- The Task of Text Classification and Sentiment Analysis
- Unpacking the complexity of an Algorithm
  - Naïve Bayes Algorithm
  - Naïve Bayes: Learning
- Case Study for Domain-specific Text Classification Models

Machine Learning versus Text Analytics



CBS

Data Mining vs. Text Mining

Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data. Some examples of such questions are as follows:

- What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature?
- What are the useful tools and techniques used for mining text data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data?
- What are the key application domains in which such mining techniques are used, and how are they effectively applied?

The most important characteristic of text data is that it is *sparse* and *high dimensional*. For example, a given corpus may be drawn from a lexicon of about 100,000 words, but a given text document may contain only a few hundred words. Thus, a corpus of text documents can be represented as a *sparse term-document matrix* of size $n \times d$, when n is the number of documents, and d is the size of the lexicon vocabulary. The (i, j)th entry of this matrix is the (normalized) frequency of the jth word in the lexicon in document i. The large size and the sparsity of the matrix has immediate implications for a number of data analytical techniques such as dimensionality reduction. In such cases, the methods for reduction should be specifically designed while taking this characteristic of text data into account. The variation in word frequencies and document lengths also lead to a number of issues involving document representation and normalization, which are critical for text mining.

Furthermore, text data can be analyzed at different levels of representation. For example, text data can easily be treated as a bag-of-words, or it can be treated as a string of words. However, in most applications, it would be desirable to represent text information *semantically* so that more meaningful analysis and mining can be done. For example, representing text data at the level of named entities such as people, organizations, and locations, and their relations may enable discovery of more interesting patterns than representing text as a bag of words. Unfortunately, the state of the art methods in natural language processing are still not robust enough to work well in unrestricted text domains to generate accurate semantic representation of text. Thus most text mining approaches currently still rely on the more shallow word-based representations, especially the bag-of-wrods approach, which, while losing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach. In special domains (e.g., biomedical domain) and for special mining tasks (e.g., extraction of knowledge from the Web), natural language processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text.

Chapters 1-4 of Aggarwal, C. C., & Zhai, C. (2012). Mining text data: Springer Science & Business Media.

Bag of words vs. String of words

```
#Example 1: Bag of words and string of words
"the theology is the theory of the religion"

# Bag_of_words approach:
bag_of_words = {'the', 'theology', 'is', 'theory','of','
   religion'}

# String_of_words approach:
string_of_words = {('the',0), ('theology',1), ('is',2), ('the'
   ,3), ('theory',4) ,('of',5),('the',6),('religion',7)}
```

- Why is it so important?
  - Shakespeare plays contain 885 000 words, but the count of unique words is 31 000

Data Mining vs. Text Mining - Sparse and High Dimensionality Problem

Imagine that you have data about **60 Facebook posts** with features:

- #likes, #comments, #shares, #comment-replies
- The text associated with each post
  - Combine texts from all posts and prepare a dictionary ($\Longrightarrow$ 300 unique words)
  - words in the columns indicate how many times each word has occurred in the post

$$\mathbf{Z} = \begin{bmatrix} 6 & 2 & 1 & 8 \\ 12 & 21 & 0 & 2 \\ 26 & 19 & 0 & 2 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 9 & 7 & 2 \\ 7 & 13 & 81 & 2 \\ 11 & 9 & 38 & 2 \end{bmatrix}_{60\times 4}$$

Facebook Post Attributes

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 2 & 0 & 3 \\ 0 & 0 & 2 & \cdots & \cdots & 1 & 0 & 0 \\ 1 & 0 & 0 & \cdots & \cdots & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 2 & \cdots & \cdots & 1 & 0 & 0 \\ 3 & 1 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 3 & 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}_{60\times 300}$$

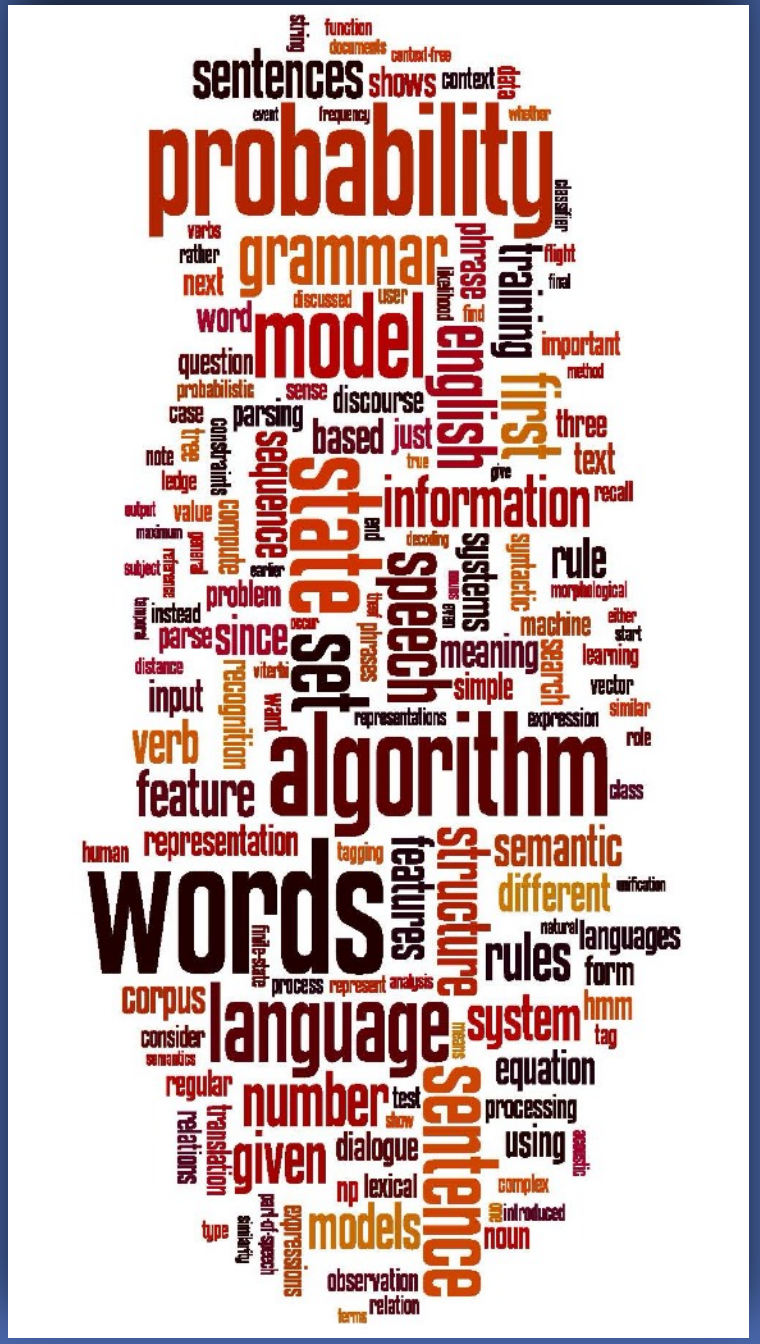
Text Features of Facebook Posts

	Terms				
	Camera	Digital	Memory	Print	...
Document 1	3	2	0	1	
Document 2	0	4	0	3	
...	...	...	...	...	...

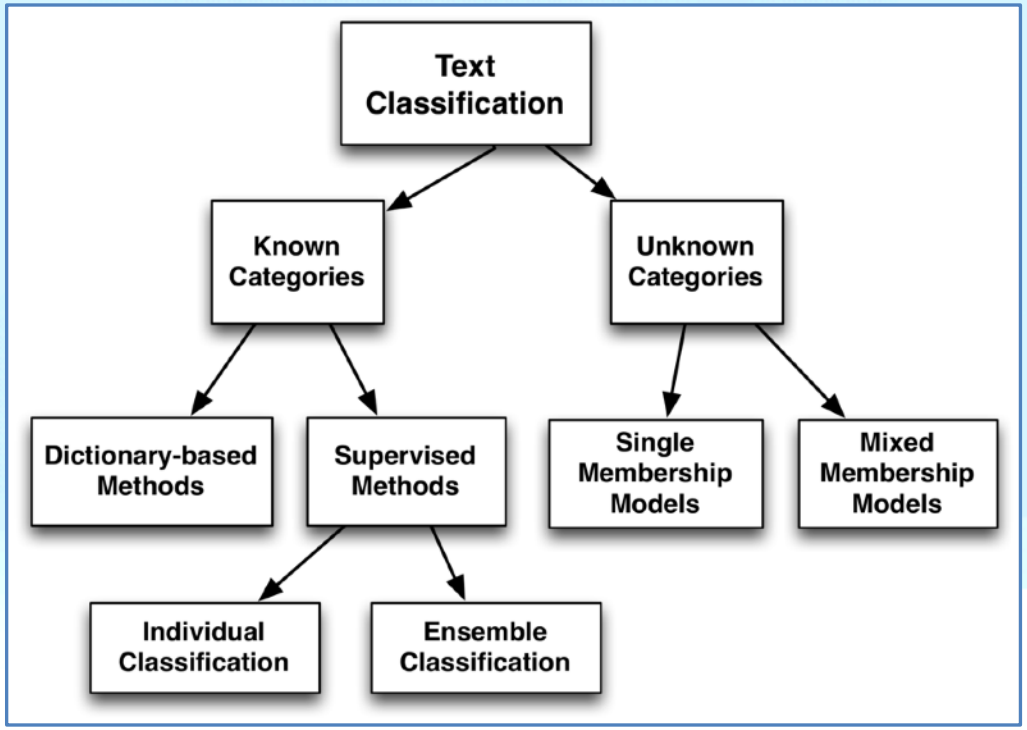
The Task of Text Classification

Many of the slides have been taken and adapted from: Speech and Language Processing by Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/

CBS



- Classification: Assigning text documents* to predefined categories
- Category: A set of labels for domain specific concept (E.g. Sentiment or Emotion Analysis)
- Classifying into Known Categories:
  - Dictionary based models
  - Supervised Learning Methods
- Unknown Categories: Topic Modeling



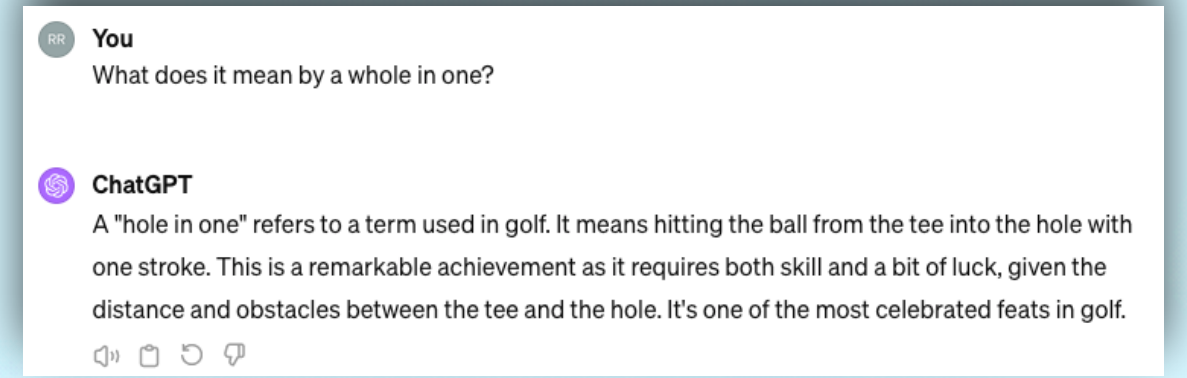
Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political analysis 21.3 (2013): 267-297.

* Text document: unit of text, which could be one or more words/sentences/paragraphs

Dictionary Based Methods

- Use rate at which keywords appear in documents
- Uses a list of words with scores to find out the document category label {+ve: +1 to +5} and {-ve: -1 to -5}
  - Boring: -1 , Disgust: -3, inspire: +2, masterpiece: +5
- Limited to categories for which dictionaries are available (Sentiment, Emotion etc.)
- domain specific i.e. accuracy depends on domain from which words are taken (Usage of word {crude} in “crude oil” vs “a crude joke”)
- Validation of dictionaries is bit hard

Ambiguity makes NLP hard



Real Newspaper headlines

- Teacher Strikes Idle Kids
  - **#1 The teacher is on strike, which idles the kids.**
  - **#2 A teacher strikes kids who are idle**
- Local High School Dropouts Cut In Half
- Grandmother of Eight Makes A Hole in One
- Ban on Nude Dancing on Governor's Desk
  - **#1 Ban on [Nude Dancing on Governor's Desk]**
  - **#2 [Ban on Nude Dancing] on Governor's Desk**

Students Cook & Serve Grandparents

On Thursday, September 9, Gorman School hosted the first annual Grandparent's Day.

All Grandparents were invited to a school wide pancake breakfast. Upper grade students served as excellent chefs, as well as taking responsibility for serving the food and the clean up after-

Dan Jurafsky and Christopher Manning. Natural Language Processing
https://www.coursera.org/course/nlp , http://www.teacherjoe.us/laughandlearnenglishheadlines.htm

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

Is this spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Positive or negative movie review?

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

?

- It was pathetic. The worst part about it was the boxing scenes.

Text Classification: definition

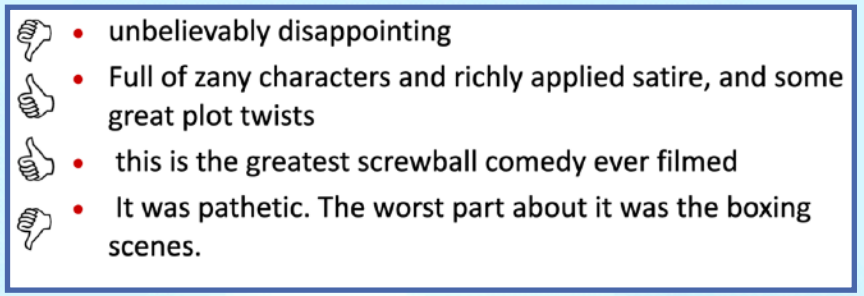
- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: a predicted class $c \in C$

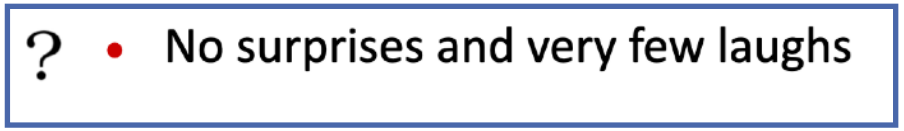
Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND“have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods:Supervised Machine Learning



- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $\{(d_1,c_1),\ldots,(d_m,c_m)\}$
- *Output:*
  - a learned classifier $\gamma:d \rightarrow c$



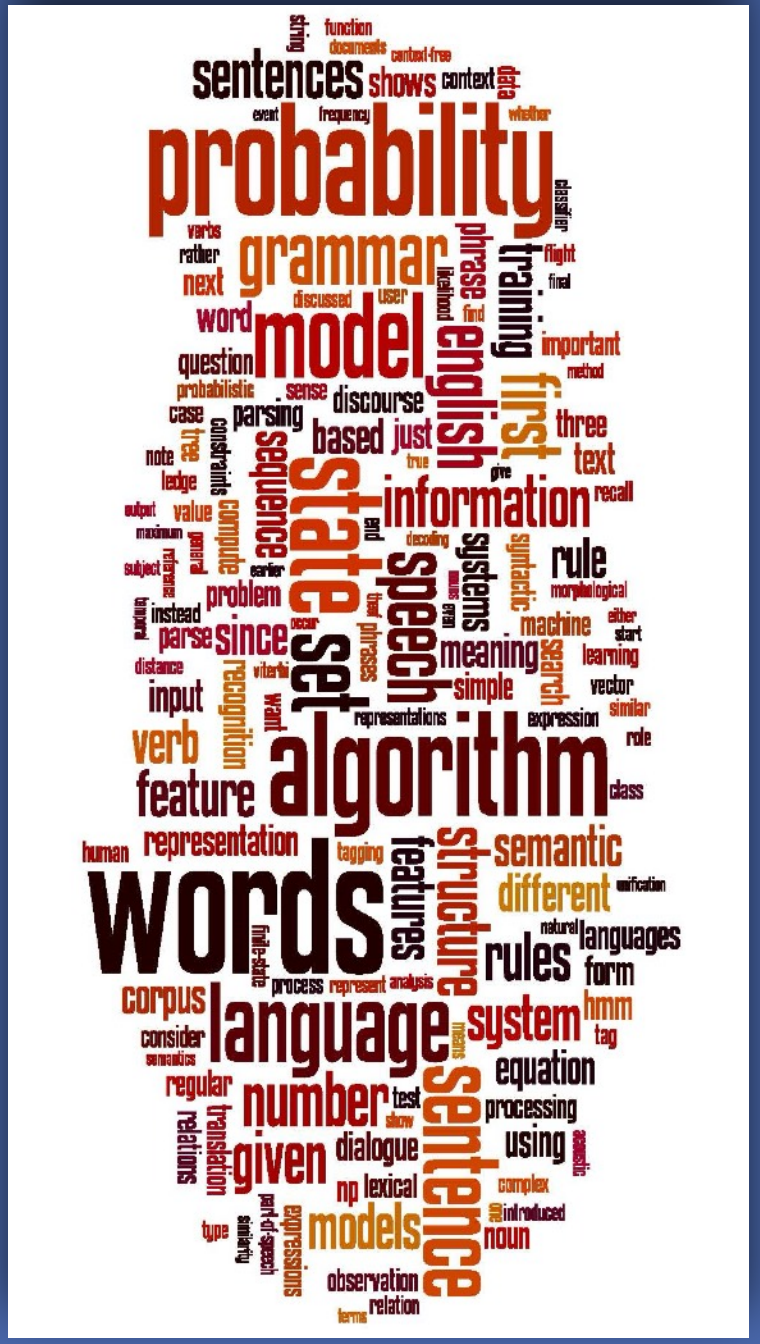
Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes (Generative classifiers)
  - Logistic regression (Discriminative classifiers)
  - Support-vector machines
  - k-Nearest Neighbors
  - …

Naïve Bayes Algorithm

Many of the slides have been taken and adapted from: Speech and Language Processing by Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/

CBS

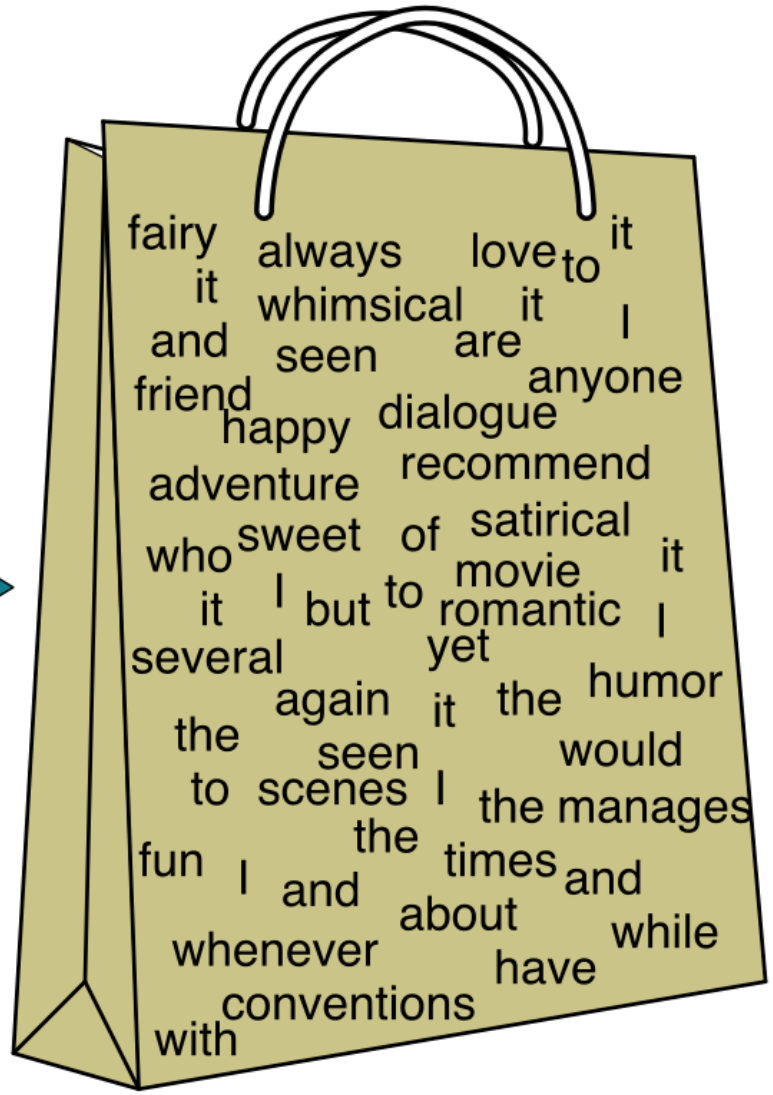


Naïve Bayes Intuition

- Simple ("naïve") classification method based on Bayes rule
- Relies on very simple representation of document
  - Bag of words

The Bag of Words Representation

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
…	…

The bag of words representation

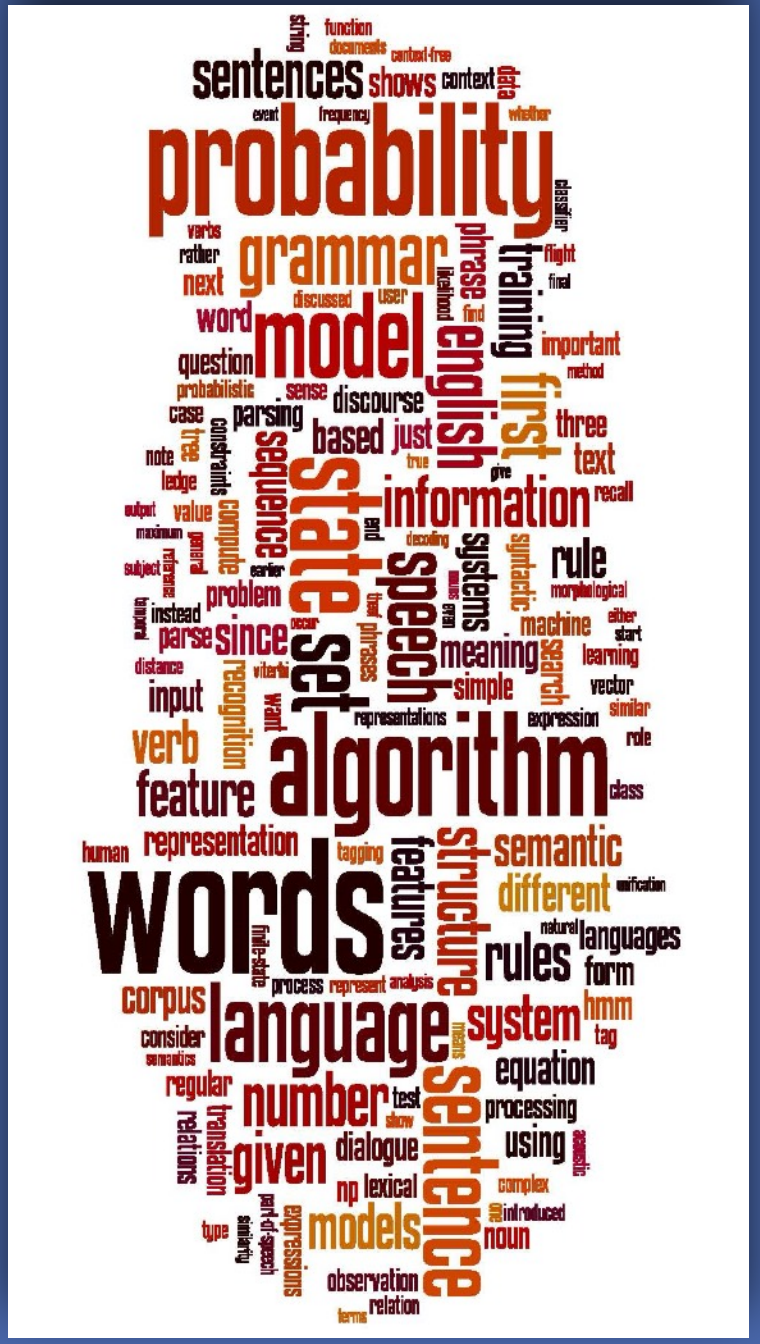
$\gamma($

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

$) = c$

Formalizing the Naïve Bayes Classifier

Many of the slides have been taken and adapted from:
Speech and Language Processing by Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/



Conditional Probability Example

Rolling of two six-sided dice and predicting sum of numbers on the dice sample space: 36 outcomes

+		B					
		1	2	3	4	5	6
A	1	2	3	4	5	6	7
	2	**3**	**4**	**5**	**6**	**7**	**8**
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

P(A==2) = 6/36 = 1/6

+		B					
		1	2	3	4	5	6
A	1	**2**	**3**	**4**	**5**	6	7
	2	**3**	**4**	**5**	6	7	8
	3	**4**	**5**	6	7	8	9
	4	**5**	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

P(A+B≦ 5) = 10/36

+		B					
		1	2	3	4	5	6
A	1	2	3	4	5	6	7
	2	**3**	**4**	**5**	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

P(A==2 | A+B≦ 5) = 3/10

https://en.wikipedia.org/wiki/Conditional_probability

Bayes' Theorem

- Thomas Bayes (1701–1761)
- Consider events X and Y and P(X) ≠ 0

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

  - P(X) and P(Y) probabilities of events A and B **without** regard to each other
  - P(X | Y) is conditional probability of observing event X when event Y occurred
  - P(Y | X) is conditional probability of observing event Y when event X occurred
  - Let's verify if P(A==2|A+B≦ 5) == 3/10 using Bayes' theorem.
    - X = (A+B≦ 5), Y = (A==2) and P(X) = P(A+B≦ 5) = 10/36, P(Y) = P(A==2) = 1/6
    - P(X|Y) = P(A+B≦ 5 | A==2) = 3/6. Event space = {(2,1), (2,2), (2,3)}
    - P(Y|X) = (3/6 * 1/6) / (10/36) = 3 /10

https://highdemandskills.com/bayes-theorem/

Bayes' Theorem

- Thomas Bayes (1701–1761)
- Consider events C and R and P(C) ≠ 0

$$P(R \mid C) = \frac{P(C \mid R) \quad P(R)}{P(C)}$$

  - C = cloudy weather, R=rain next day
  - P(R|C) = probability of raining the next day, given it is cloudy today
  - P(C|R) = probability of cloudy on the preceding day when it rains next day
  - P(C) = probability of cloudy day
  - P(R) = probability of rainy day

https://highdemandskills.com/bayes-theorem/

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c* [e.g. c ∈ {+,0,-}]

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Naïve Bayes Classifier (I)

[e.g. $c \in \{+,0,-\}$]

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

Dropping the denominator

Naïve Bayes Classifier (II)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c) P(c)$$

[e.g. c ∈ {+,0,-}]

$$= \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Document d represented as features x1..xn

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

Naïve Bayes Classifier (IV)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$O(|X|^n \bullet |C|)$ parameters

Could only be estimated if a very, very large number of training examples was available.

How often does this class occur?

We can just count the relative frequencies in a corpus

Multinomial Naïve Bayes Independence Assumptions

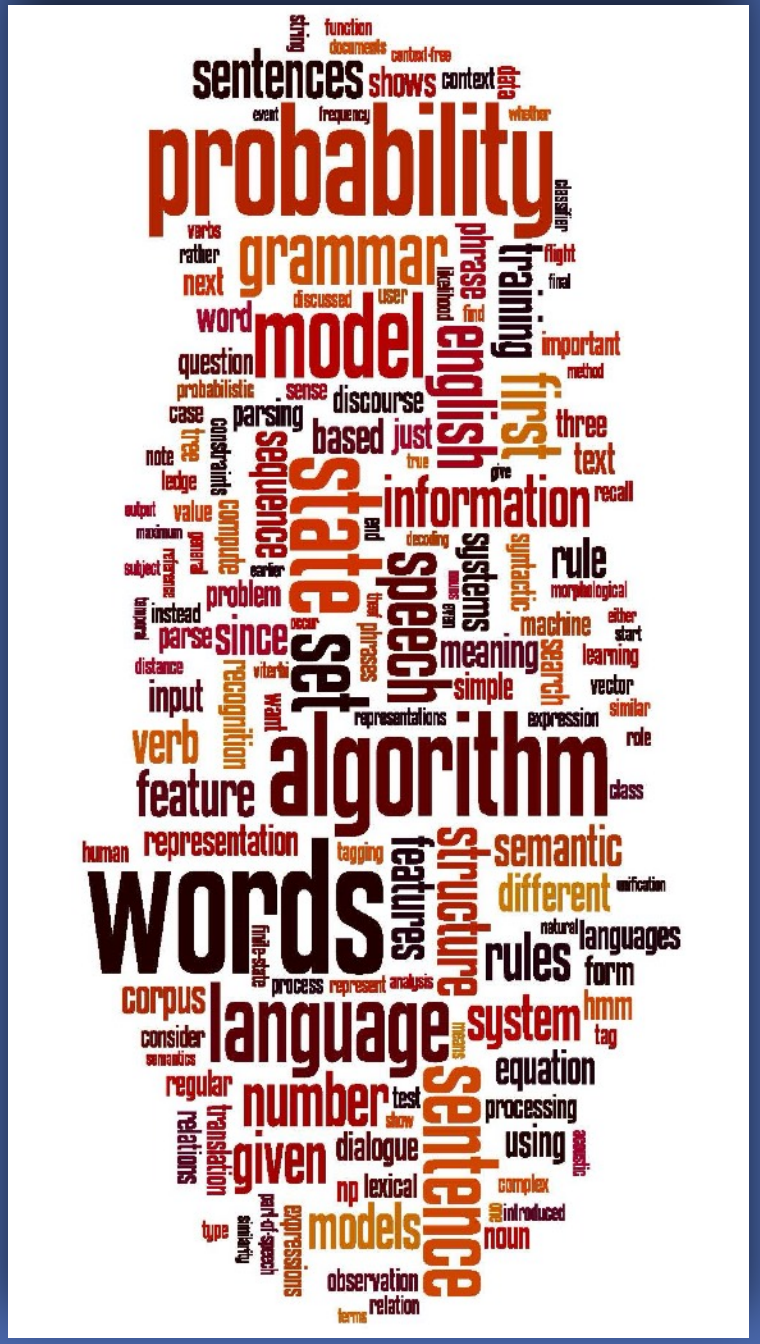
$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Bag of Words assumption**: Assume position doesn't matter
- **Conditional Independence**: Assume the feature probabilities $P(x_i \mid c_j)$ are independent given the class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Naïve Bayes: Learning

Many of the slides have been taken and adapted from:
Speech and Language Processing by Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/



CBS

Training Sets: Positive or negative movie review?

- unbelievably disappointing (- ve)

- Full of zany characters and richly applied satire, and some great plot twists (+ ve)

- this is the greatest screwball comedy ever filmed (+ ve)

- It was pathetic. The worst part about it was the boxing scenes. (- ve)

- No surprises and very few laughs

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

P(+) = ⅖ = 0.4

P(-) = ⅗ = 0.6

Parameter estimation

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

fraction of times word w_i appears among all words in documents of class c_j

- Create mega-document for class c_j by concatenating all docs in this class
  - Use frequency of w in mega-document

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive** (***thumbs-up)***?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Laplace (add-1) smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) \right) + 1)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Multinomial Naïve Bayes: Learning

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha \mid Vocabulary|}$$

Naïve Bayes Example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

- $W_{(-)} = 14$
- $W_{(+)} = 9$
- Vocabulary = 20 (3 words reperated: and, very, the)
- P(-) = 3/5 = 0.6
- P(+) = 2/5 = 0.4
- P("Predictable"| -) = (1+1)/(14+20) = 2/34
- P("Predictable"| +) = (0+1)/(9+20) = 1/29
- P("with"| -) = (0+1)/(14+20) = 1/34 [P("with"| +) = 1/29]
- P("no"| -) = (1+1)/(14+20) = 2/34
- P("no"| +) = (0+1)/(9+20) = 1/29
- P("originality"| -) = (0+1)/(14+20) = 1/34 [P("originality"| +) = 1/29]
- P(S| -) P(-) = (2x1x2x1 / 34^4) x 3/5 = 1.8 x 10^{-6}
- P(S| +) P(+) = (1x1x1x1 / 29^4) x 2/5 = 0.56 x 10^{-6}
- Therefore the sentence is negative (-).

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w, c) + 1}{count(c) + \mid V \mid}$$

Naïve Bayes Example - II

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Beijing Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

P(Beijing|*c*) = 2/14 = 1/7 P(Beijing|*j*) = 1/9

P(c|d5) $\propto$ 3/4 * (3/7) * 1/7 * 1/14 * 1/14
≈ 0.00023

P(j|d5) $\propto$ 1/4 * (2/9) * 1/9 * 2/9 * 2/9
≈ 0.00030

C(d5) = j

Domain-specific models - Public Health Case Study

Jonathan-Raphael Reichert, Klaus Langholz Kristensen, Raghava Rao Mukkamala, Ravi Vatrapu. A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes. 19th IEEE International Conference on e-Health Networking, Application & Services (HEALTHCOM 2017), Dalian, China, October, 2017
https://raghavamukkamala.github.io/files/pubs/2017_IEEE-Healthcom-Patient-journey-cam.pdf

CBS

Jonathan-Raphael Reichert, Klaus Langholz Kristensen, Raghava Rao Mukkamala, Ravi Vatrapu. A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes. 19th IEEE International Conference on e-Health Networking, Application & Services (HEALTHCOM 2017), Dalian, China, October, 2017
https://raghavamukkamala.github.io/files/pubs/2017_IEEE-Healthcom-Patient-journey-cam.pdf

Research Questions

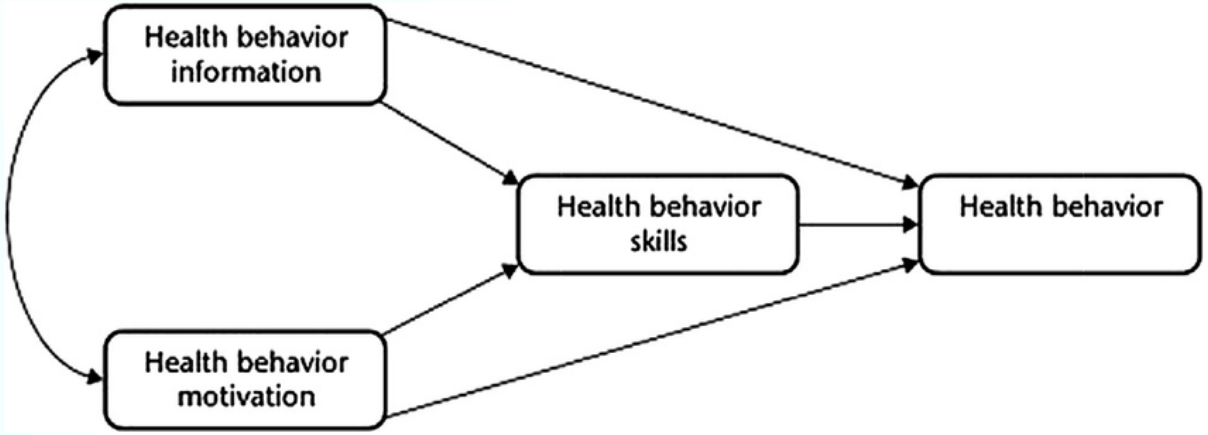
Main focus is to explore how supervised machine learning techniques can be applied to diabetes conversations to extract valuable insights from the user-generated content.

- Which insights can be derived from online forum conversations about type 2 diabetes?
- How can such insights be used to optimize healthcare communication and services to benefit diabetes patients, pharmaceutical companies, and healthcare organizations?
- In what way can insights from online conversations about type 2 diabetes lead to strategic recommendations?

Jonathan-Raphael Reichert, Klaus Langholz Kristensen, Raghava Rao Mukkamala, Ravi Vatrapu. A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes. 19th IEEE International Conference on e-Health Networking, Application & Services (HEALTHCOM 2017), Dalian, China, October, 2017
https://raghavamukkamala.github.io/files/pubs/2017_IEEE-Healthcom-Patient-journey-cam.pdf

Public Health Case Study

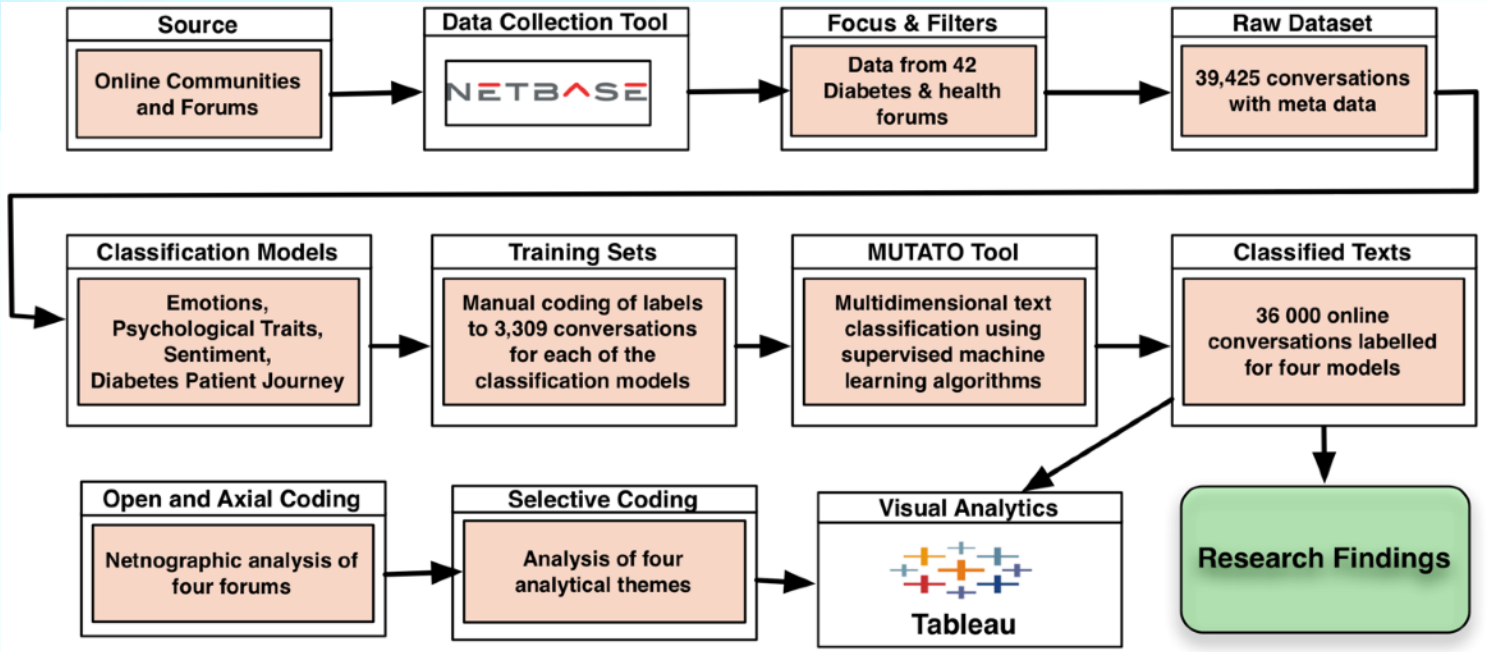
- Patient adherence model: Adequate information related to a patient's personal treatment regimen is necessary for good adherence and good health behavior (Fisher, et.al., 2003)
- Patients discussing their difficulties, conditions, and related problems in online communities become better at managing their disease and coping with related issues (Choudhury, 2014)



W. A. Fisher, J. D. Fisher, and J. Harman, The information-motivation- behavioral skills model: A general social psychological approach to understanding and promoting health behavior, Social psychological foundations of health and illness, 2003

Research Methodology

- Dataset consists of 39,425 texts collected from 42 online forums, including reddit, community.diabetes.org, diabetes.co.uk, healthunlocked.com, myfitnesspal.com etc.



Domain Specific Models - Emotion, Sentiment

Label	Definition
Joy	Feeling of well-being, often also stated as happiness.
Sadness	It is opposite of happiness or joy and is usually seen as a lowering the individuals' mood for a temporary period of time (where depression is a longer period of time).
Trust	Trust is concerned with believing in something. This might be trust in a person or a thing.
Disgust	A feeling of revulsion or strong disapproval aroused by something unpleasant or offensive. Feeling disgust often relates to something that we have tasted, which made us feel discomfort.
Fear	Fear is present in a human being, when he or she is trying to avoid some kind of pain or a situation where one's comfort or happiness is threatened.
Anger	Anger is an intense emotional state that includes feelings such as irritation, uncomfortableness, provocation or even, at the extreme, rage.
Anticipation	Anticipation is a kind of expectation towards future. The expectation can be of a positive kind (feeling exited) or can be of fear or in extreme cases anxiety.
Surprise	Surprise is the result of experiencing something unexpected. Surprise is only momentarily and does, in itself, not have positive or negative spectrum; it can be anything.

Table I
TEXT CLASSIFICATION MODEL: EMOTIONS [16]

Label	Definition
Positive	Positive means that something is good, beneficial and/or desirable in a given context.
Neutral	That something is neutral means that it is neither or. In this case, neither positive or negative.
Negative	Negative means that something is bad, hurtful or unwanted in a given context.

Table II
TEXT CLASSIFICATION MODEL: SENTIMENT

R. Plutchik, Emotions and life: Perspectives from psychology, biology, and evolution. American Psychological Association, 2003

Domain Specific Model - Personality

Label	Definition
Openness	Describes a general openness to new ideas, experiences and is related to curiosity, adventure and imagination.
Conscientious-ness	Describes an individual who aims for achievements [17] and expresses a propensity to be thoughtful, thorough, in control, a preference for planning and structural living.
Extraversion	Extraversion is often opposed to Introversion [17] and the extravert is often the centre of attention, out-going, socially comfortable, energetic and likes to talk.
Agreeableness	Describes an individual is focused on establishing con-sensus to achieve social harmony. Such individuals of-ten conform to social norms and are usually generous, trustworthy, optimistic, caring and emotionally support-ive [17].
Neuroticism	This trait is linked to emotional instability, anxiety and depression [17]. Individuals labeled with neuroticism will be vulnerable and emotionally reactive.

Table III
TEXT CLASSIFICATION MODEL: PERSONALITY TRAITS

M. Digman, Personality structure: Emergence of the five-factor model, Annual review of psychology, vol. 41, no. 1, 1990

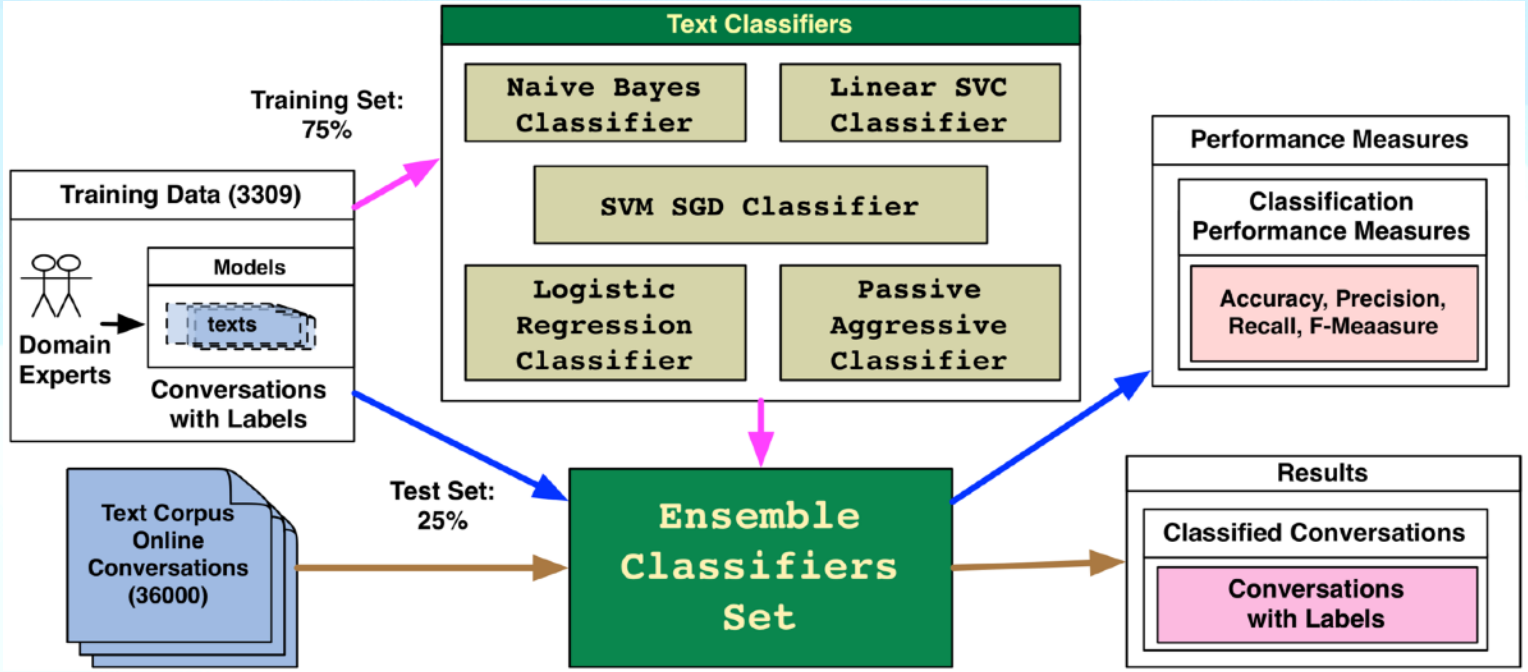
Domain Specific Model - Patient Journey

Label	Definition
Undiagnosed	People without diabetes, patients with pre-diabetes or gestational diabetes and factual texts about diabetes
Relatives of diabetes patients	People discussing topics on behalf of family/friends diagnosed with diabetes or in the risk zone
Diagnosis	Patients diagnosed with diabetes by a health care professional.
Clinical Treatment	Everything related to medical treatment of diabetes. Clinical treatment, managing, adhering to treatment.
Alternative Treatment	Conversations related to alternative treatment (e.g. Ayurveda or home remedies).
Living with diabetes - Lifestyle, social & psychological	Everything related to managing social and psychological life related to diabetes. Topics may include how diabetes have changed the social lifestyle or affects the patient psychologically
Living with diabetes - nutrition	Includes discussions about diet, recipes and other questions related to nutrition.
Living with diabetes - exercise	Includes discussions and questions related to an active lifestyle

Table IV
TEXT CLASSIFICATION MODEL: PATIENT JOURNEY

Dartmouth-Hitchcock, A typical patient's journey: Diabetes, http://www.dartmouth-hitchcock.org/diabetes/a_typical_patients_journey_diabetes.html

Text Classification Approach



Classifiers Performance

Model 1: Emotions				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.71	0.70	0.69	0.695
Linear SVC	0.75	0.75	0.74	0.748
Logistic Regression	0.68	0.58	0.51	0.576
Passive Aggressive	0.73	0.73	0.73	0.735
SVM SGD	0.66	0.61	0.58	0.613
Voted Accuracy	-	-	-	0.706
Model 2: Sentiment				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.78	0.77	0.77	0.772
Linear SVC	0.81	0.81	0.81	0.806
Logistic Regression	0.73	0.7	0.67	0.698
Passive Aggressive	0.81	0.81	0.81	0.807
SVM SGD	0.73	0.62	0.53	0.621
Voted Accuracy	-	-	-	0.789
Model 3: Personality Traits				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.67	0.66	0.66	0.661
Linear SVC	0.69	0.68	0.68	0.683
Logistic Regression	0.63	0.62	0.61	0.619
Passive Aggressive	0.69	0.69	0.69	0.691
SVM SGD	0.65	0.64	0.63	0.636
Voted Accuracy	-	-	-	0.675
Model 4: Patient Journey				
Classifiers	**Precision**	**Recall**	**F1-score**	**Accuarcy**
Multinomial NB	0.84	0.84	0.84	0.843
Linear SVC	0.87	0.87	0.87	0.872
Logistic Regression	0.81	0.79	0.78	0.794
Passive Aggressive	0.88	0.88	0.88	0.881
SVM SGD	0.8	0.79	0.77	0.786
Voted Accuracy	-	-	-	0.865

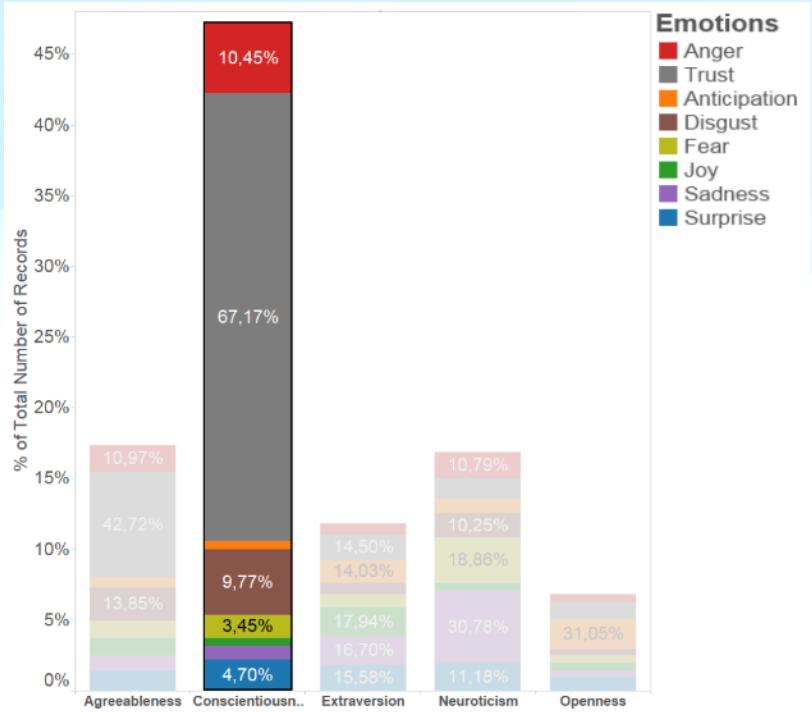
Table II

PERFORMANCE MEASURES OF THE CLASSIFIERS

Results - I

High Degree of Trust

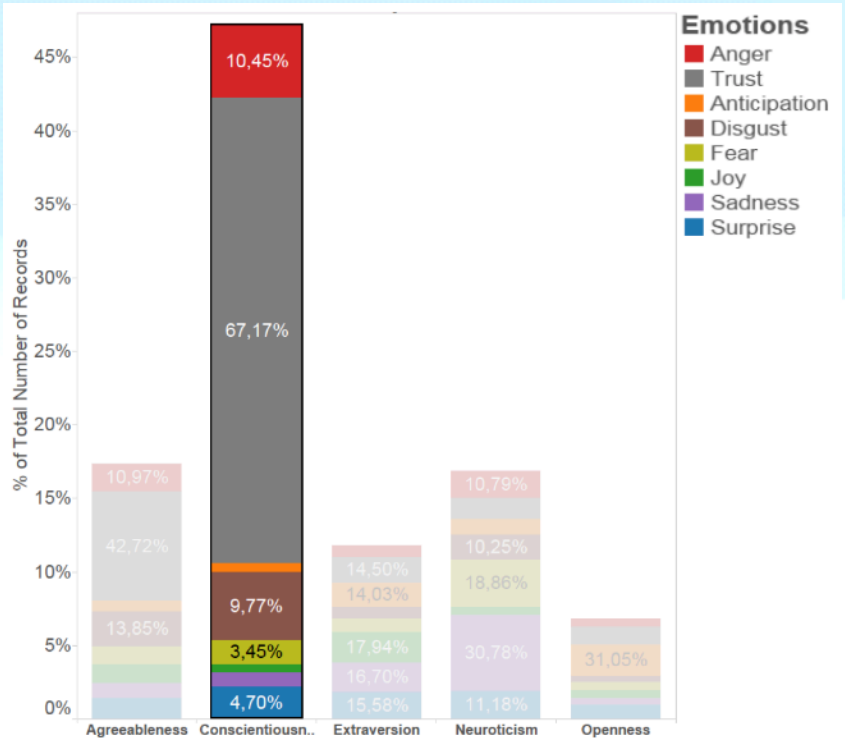
- online communities play a vital role in knowledge-sharing and general discussions around disease-related information
- conscientiousness is the most prevalent personality trait among the users
- combined with emotions trust represent the largest share of conscientiousness posts
- indicates that the users have high amount of trust in these online forums



Results - 2

Support in Digital Space

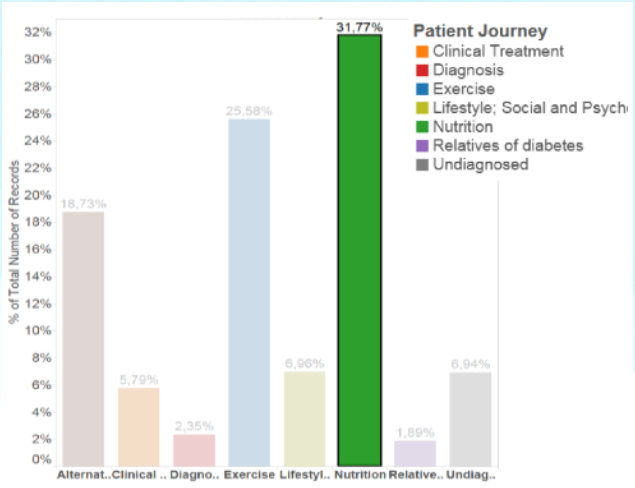
- Online communities not only exist to fulfill information needs, but also being supportive for patients seeking comfort and empathy.
- Patients who have been diagnosed recently, as they often express a need for support and motivation
- people in online communities draw on each other's experiences
- agreeableness is the major personality trait after the conscientiousness.

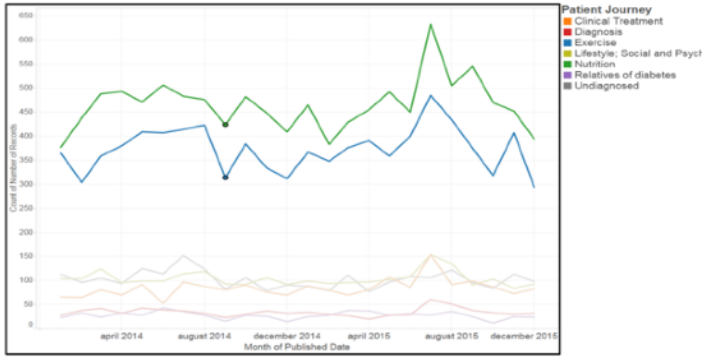


Results - 3

Perceptions of Health - Diet and Nutrition

- One of the main concerns about living with type 2 diabetes is the constant struggle to control the body's blood glucose levels
- many patients advocate that diet and nutrition can either cure or at least slow down disease progression
- majority of posts classified in relation to the Patient Journey model are labelled with Nutrition
- there existed a high association between Exercise and Nutrition





Thank you!

Raghava Mukkamala

rrm.digi@cbs.dk

https://www.cbs.dk/staff/rrmdigi

https://raghavamukkamala.github.io/

https://cbsbda.github.io/