Natural Language Processing

Text classification

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

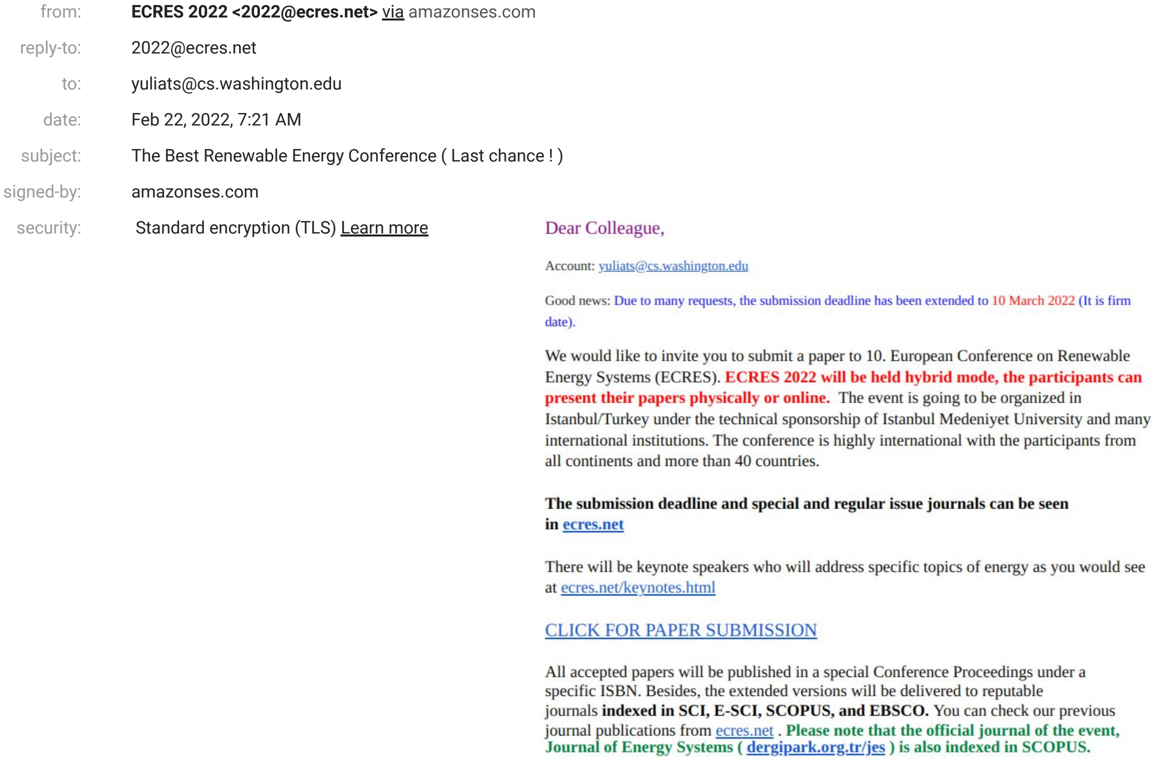
Some questions from the previous lecture

- NLP for sign languages

Announcements

- HW1 is out today, please start early!

Is this spam?



from: **ECRES 2022 <2022@ecres.net>** via amazonses.com
reply-to: 2022@ecres.net
to: yuliats@cs.washington.edu
date: Feb 22, 2022, 7:21 AM
subject: The Best Renewable Energy Conference (Last chance !)
signed-by: amazonses.com
security: Standard encryption (TLS) Learn more

Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been extended to 10 March 2022 (It is firm date).

We would like to invite you to submit a paper to 10. European Conference on Renewable Energy Systems (ECRES). **ECRES 2022 will be held hybrid mode, the participants can present their papers physically or online.** The event is going to be organized in Istanbul/Turkey under the technical sponsorship of Istanbul Medeniyet University and many international institutions. The conference is highly international with the participants from all continents and more than 40 countries.

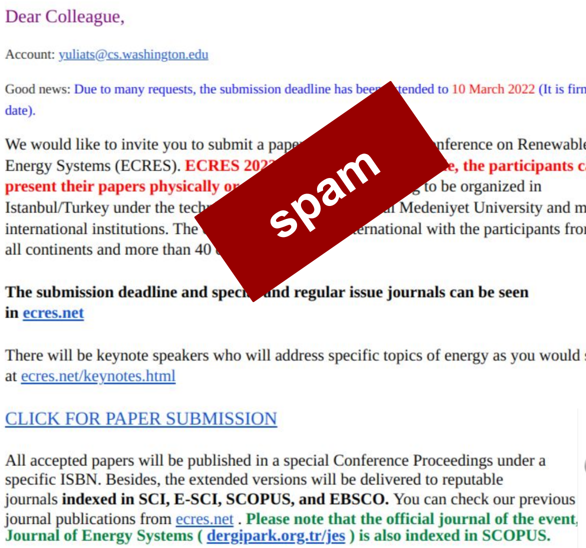
The submission deadline and special and regular issue journals can be seen in ecres.net

There will be keynote speakers who will address specific topics of energy as you would see at ecres.net/keynotes.html

CLICK FOR PAPER SUBMISSION

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals **indexed in SCI, E-SCI, SCOPUS, and EBSCO.** You can check our previous journal publications from ecres.net . **Please note that the official journal of the event, Journal of Energy Systems (dergipark.org.tr/jes) is also indexed in SCOPUS.**

Spam classification



Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been tended to 10 March 2022 (It is firm date).

We would like to invite you to submit a pape nference on Renewable Energy Systems (ECRES). **ECRES 202 e, the participants can present their papers physically o** to be organized in Istanbul/Turkey under the tech Medeniyet University and many international institutions. The ernational with the participants from all continents and more than 40

The submission deadline and spec nd regular issue journals can be seen in ecres.net

There will be keynote speakers who will address specific topics of energy as you would see at ecres.net/keynotes.html

CLICK FOR PAPER SUBMISSION

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals **indexed in SCI, E-SCI, SCOPUS, and EBSCO.** You can check our previous journal publications from ecres.net . **Please note that the official journal of the event, Journal of Energy Systems (dergipark.org.tr/jes) is also indexed in SCOPUS.**

Invitation to present at the February 2022 Wikimedia Research Showcase

Emily Lescak <elescak@wikimedia.org>
to yuliats@cs.washington.edu

Tue, Nov 23, 2021, 11:00 AM

Hi Yulia,

My name is Emily Lescak and I am a member of the Research team at the Wikimedia Foundation. On be o present your research on social biases on Wikipedia at our Research Showcase in February 2022. This topic fits into our theme for this showcase, which is gaps and biases on Wikipedia.

The Wikimedia Research Showcase is a monthly, public lecture series where Foundation, aca ent work related to Wikipedia, Wikimedia, peer production, and open-source software. We focus on topics and projects that we think our audience— a global community of academic researchers, Wiki on staff—would find interesting and relevant to their work.

Research Showcase presentations are generally 20 minutes long, with an additio nvite two presenters to every showcase. Most presenters choose to use slides to present their work.

The February showcase takes place on the 16th at 9:15AM Pacific / 17:15 UT ouTube and also archived for later viewing on the Wikimedia Foundation's YouTube channel

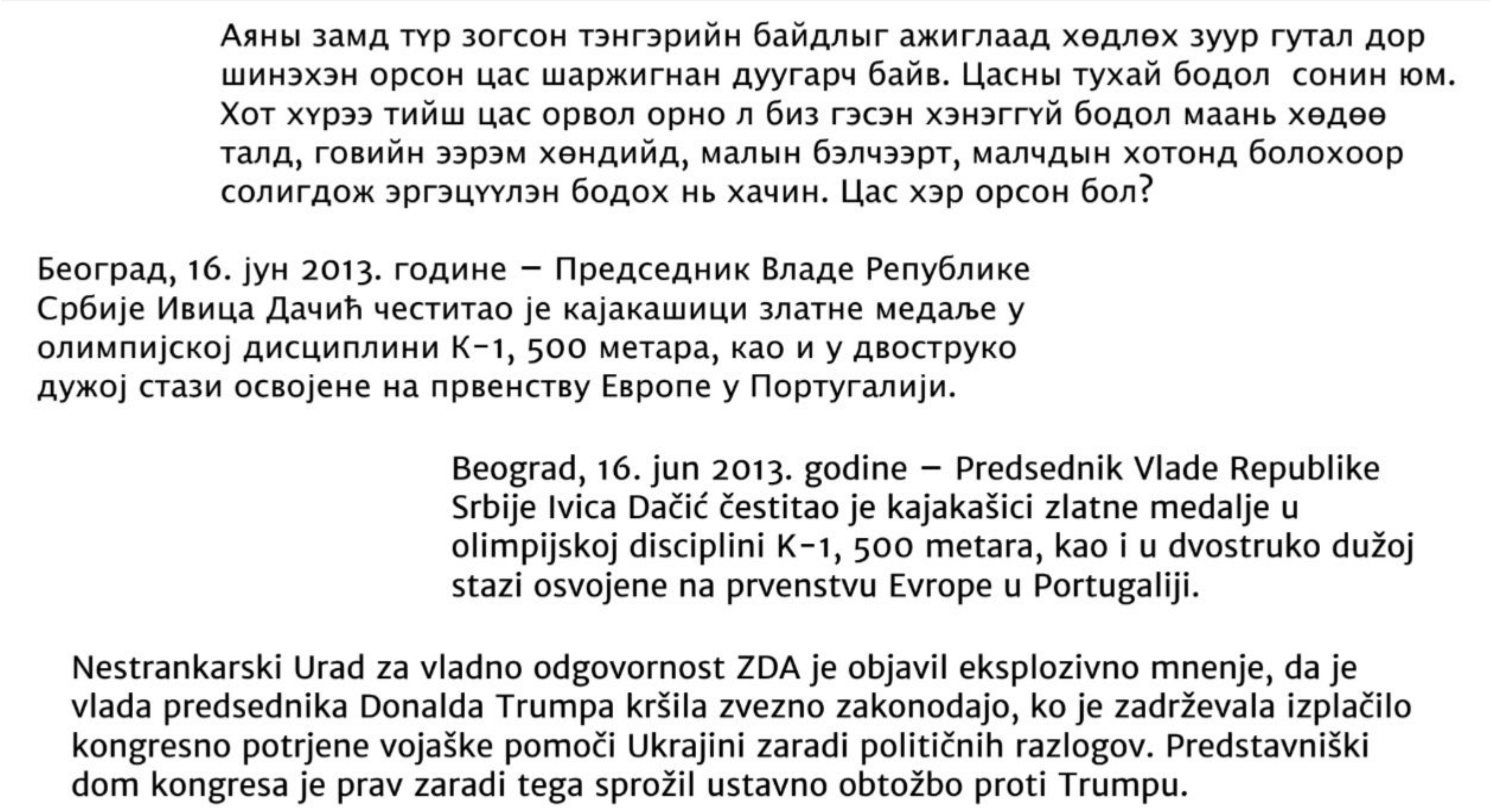
If this date does not work for you, but you are still interested in giving a showcase we can discuss other options.

I hope to get a chance to see your work presented at the Research Showcase!

Sincerely,

Emily

Language ID



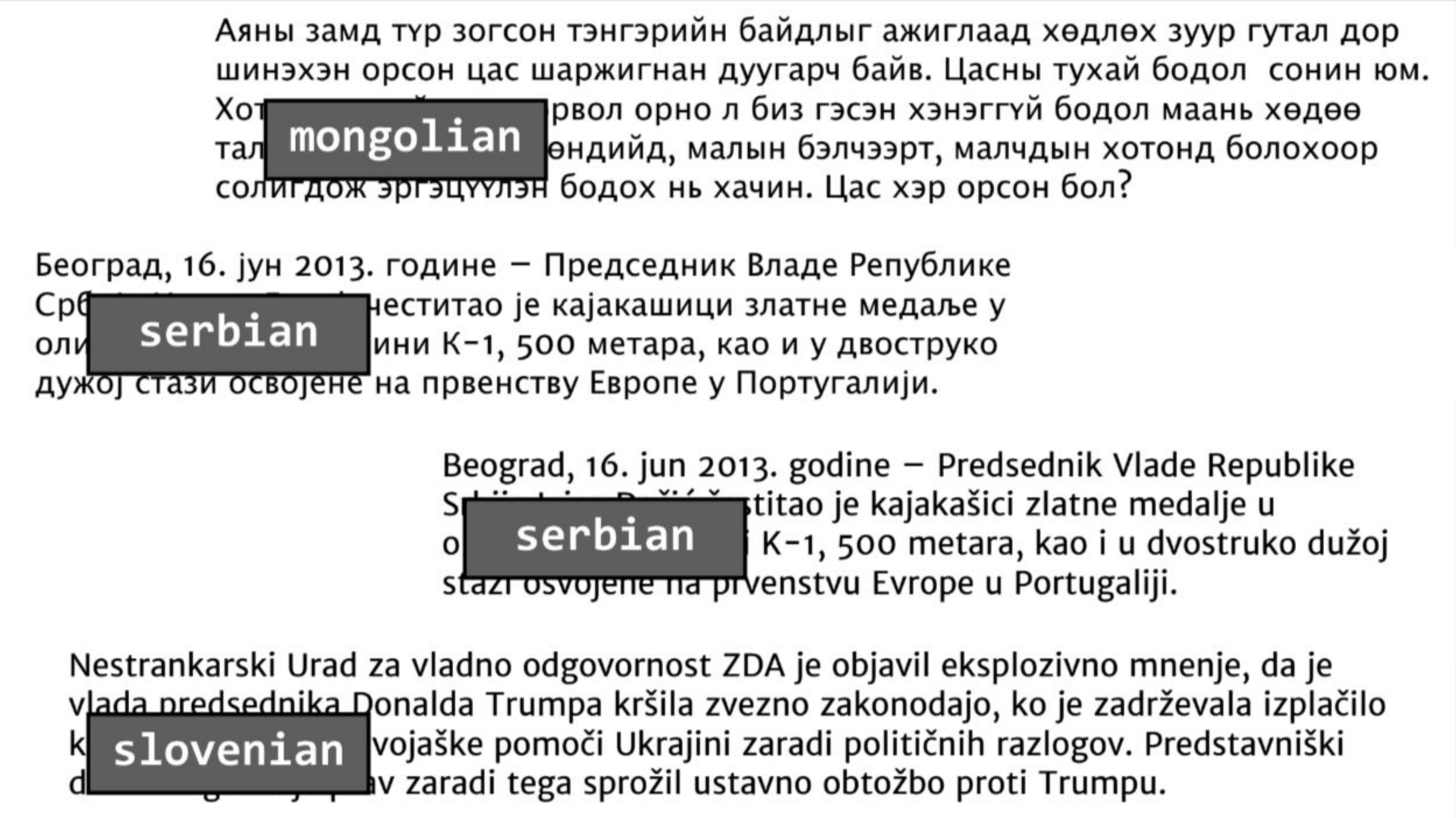
Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хөдөө талд, говийн ээрэм хөндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије Ивица Дачић честитао је кајакашици златне медаље у олимпијској дисциплини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. jun 2013. godine – Predsednik Vlade Republike Srbije Ivica Dačić čestitao je kajakašici zlatne medalje u olimpijskoj disciplini K-1, 500 metara, kao i u dvostruko dužoj stazi osvojene na prvenstvu Evrope u Portugaliji.

Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

Language ID



Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор
шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм.
Хот рвол орно л биз гэсэн хэнэггүй бодол маань хөдөө
тал өндийд, малын бэлчээрт, малчдын хотонд болохоор
солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике
Ср честитао је кајакашици златне медаље у
оли ини К-1, 500 метара, као и у двоструко
дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. jun 2013. godine – Predsednik Vlade Republike
S titao je kajakašici zlatne medalje u
o K-1, 500 metara, kao i u dvostruko dužoj
stazi osvojene na prvenstvu Evrope u Portugaliji.

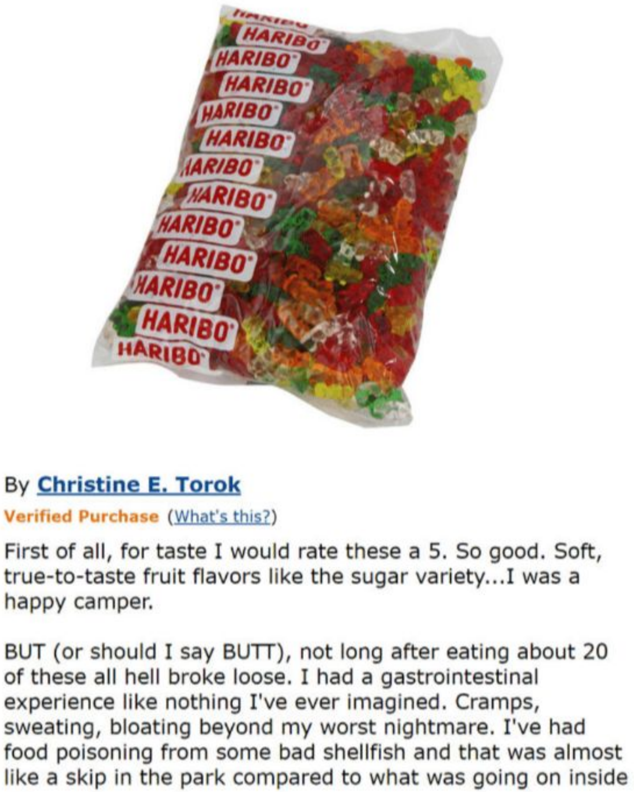
Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je
vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo
k vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški
d v zaradi tega sprožil ustavno obtožbo proti Trumpu.

Sentiment analysis

By **John Neal**

This review is from: Accoutrements Horse Head Mask (Toy)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.



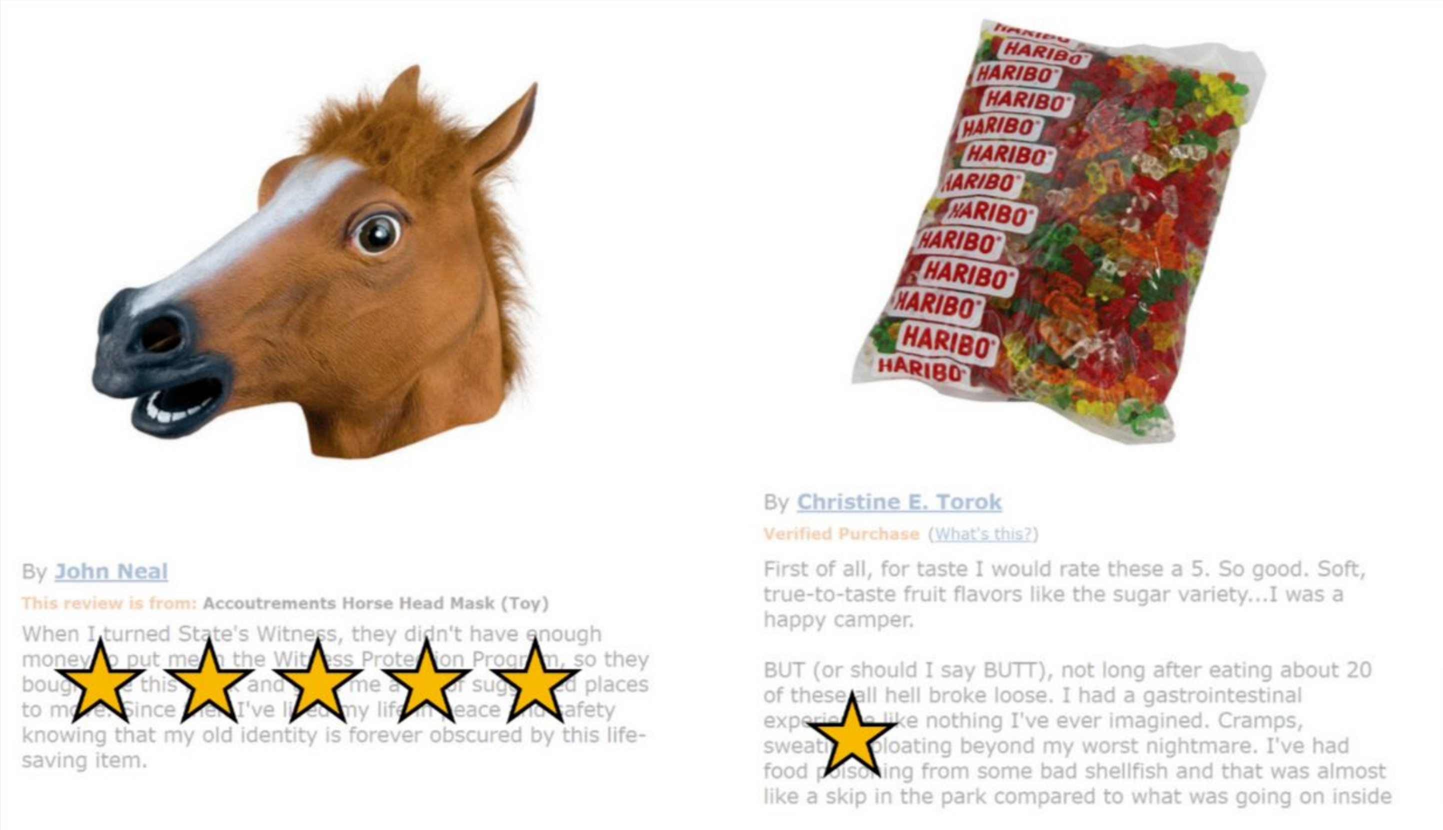
By **Christine E. Torok**

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

Sentiment analysis



By **John Neal**

This review is from: **Accoutrements Horse Head Mask (Toy)**

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

★★★★★

By **Christine E. Torok**

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

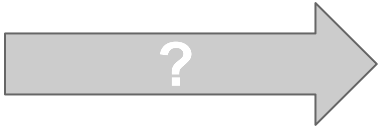
★

Topic classification

MEDLINE Article

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences



MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Authorship attribution: is the author male of female?

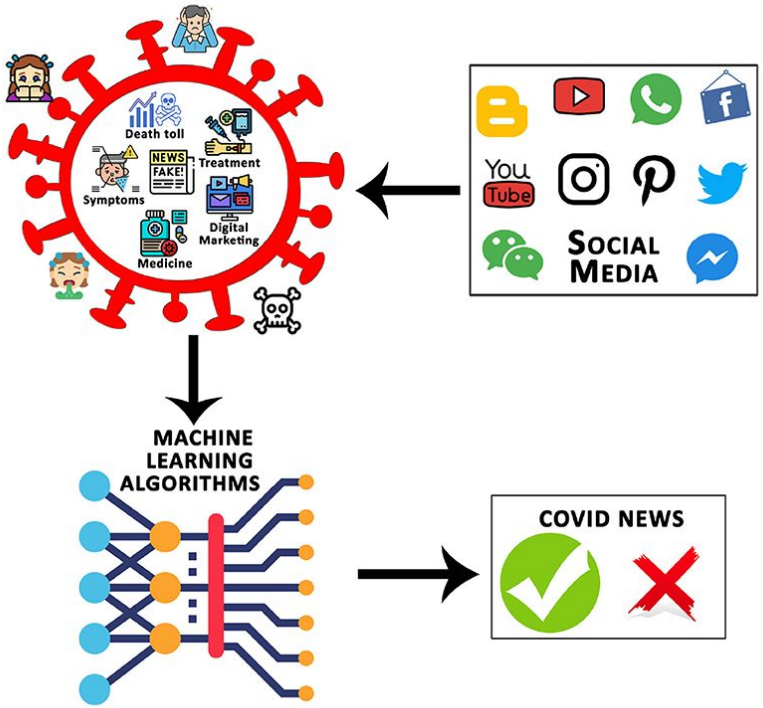
By 1925 Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam.

Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of the greatest assets...

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, volume 23, number 3, pp. 321–346

Fact verification: trustworthy or fake?

Have Covid19?
Drink bleach!



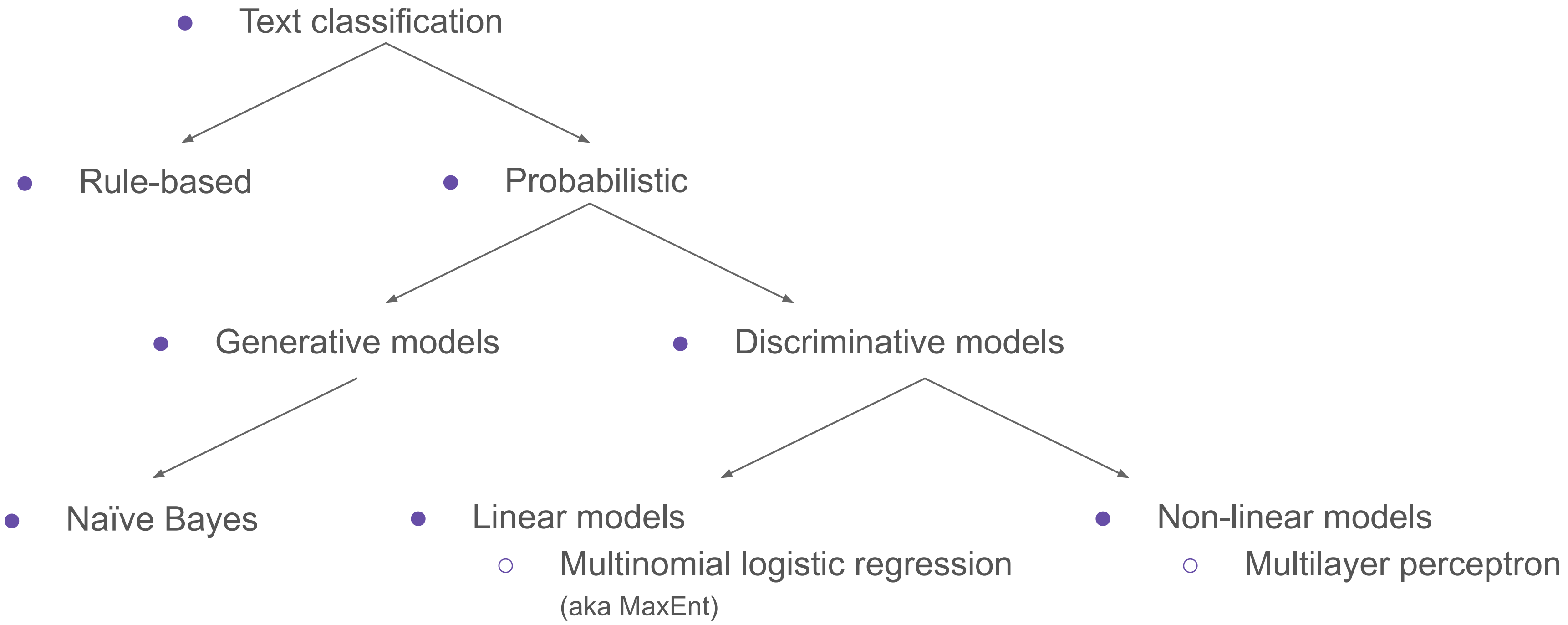
Detecting COVID-19-Related Fake News Using Feature Extraction

Suleman Khan, Saqib Hakak, N. Deepa, B. Prabadevi, Kapal Dev and Silvia Trelova

Text classification

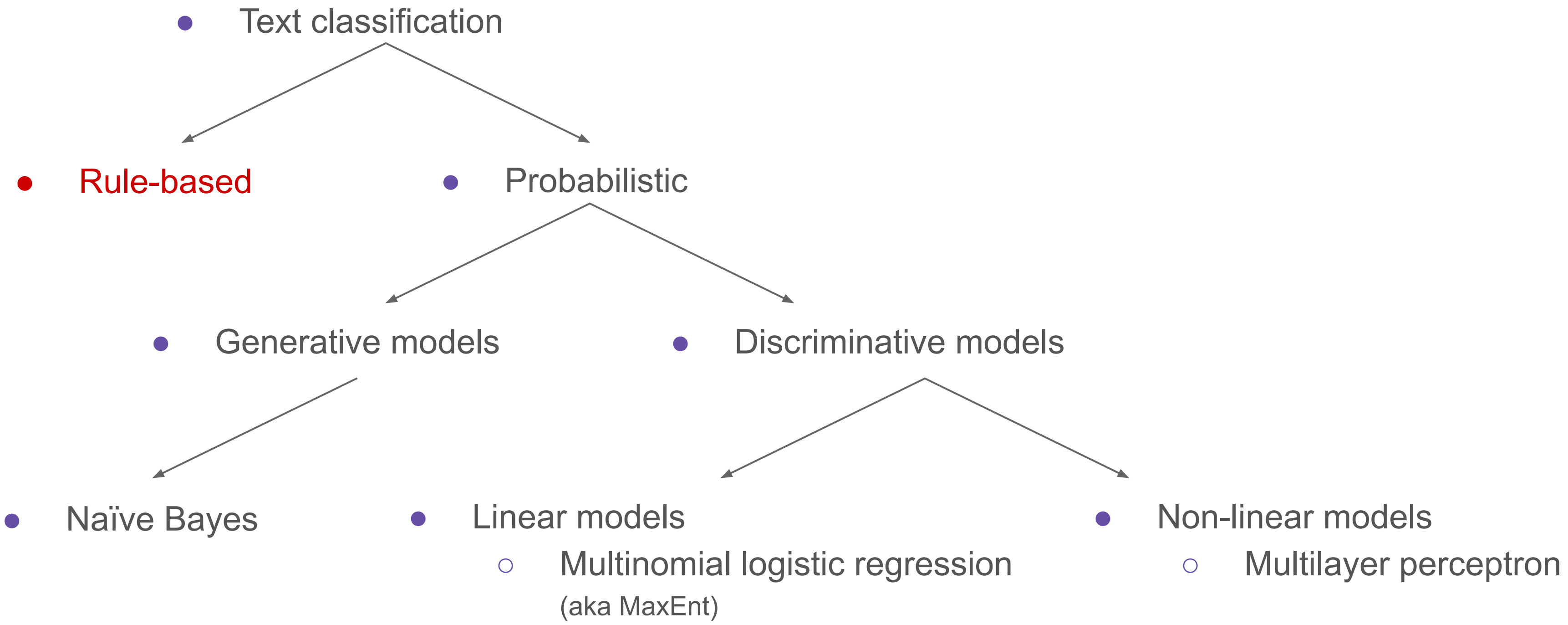
- We might want to categorize the content of the text:
    - Spam detection (binary classification: spam/not spam)
    - Sentiment analysis (binary or multiway)
        - movie, restaurant, product reviews (pos/neg, or 1-5 stars)
        - political argument (pro/con, or pro/con/neutral)
        - Topic classification (multiway: sport/finance/travel/etc)
    - Language Identification (multiway: languages, language families)
    - …
- Or we might want to categorize the author of the text (authorship attribution)
    - Human- or machine generated?
    - Native language identification (e.g., to tailor language tutoring)
    - Diagnosis of disease (psychiatric or cognitive impairments)
    - Identification of gender, dialect, educational background, political orientation (e.g., in forensics [legal matters], advertising/marketing, campaigning, disinformation)
    - …

We’ll consider alternative models for classification



- Text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

We’ll consider alternative models for classification



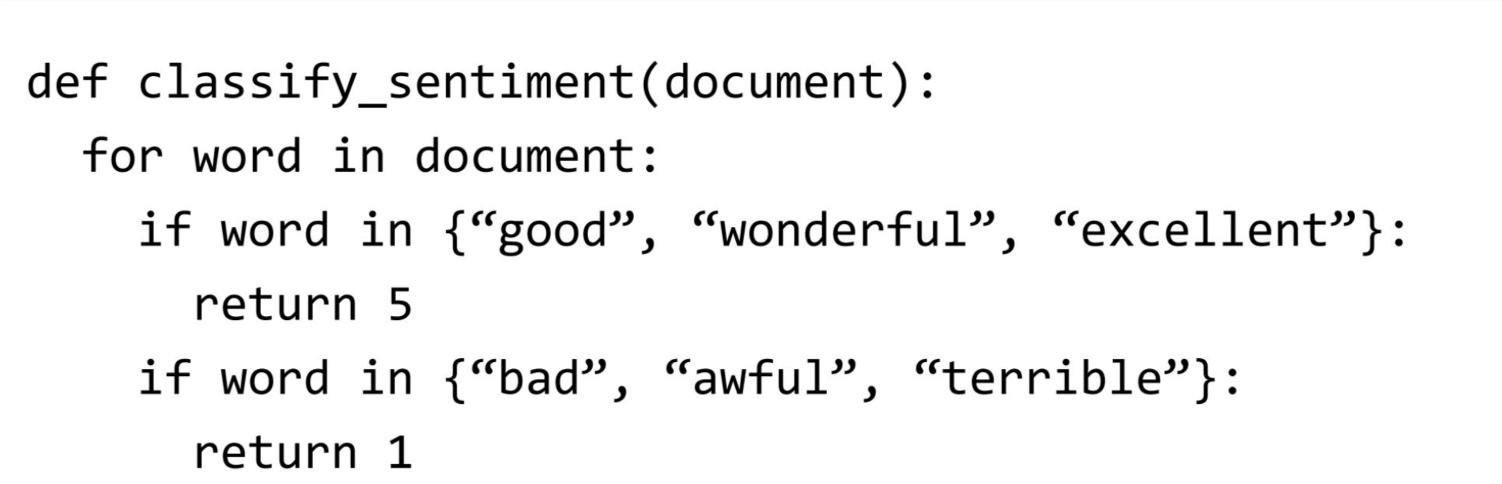
- Text classification

- Rule-based
- Probabilistic

- Generative models
- Discriminative models

- Naïve Bayes
- Linear models
    - Multinomial logistic regression (aka MaxEnt)
- Non-linear models
    - Multilayer perceptron

Rule-based classifier



```
def classify_sentiment(document):
  for word in document:
    if word in {“good”, “wonderful”, “excellent”}:
      return 5
    if word in {“bad”, “awful”, “terrible”}:
      return 1
```

Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

⟶ hard to identify a priori which words are informative (and what information they carry!)

Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

⟶ hard to identify a priori which words are informative (and what information they carry!)

Sentiment: It’s not life-affirming, it’s vulgar, it’s mean, but I liked it.

Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

⟶ hard to identify a priori which words are informative (and what information they carry!)

Sentiment: It’s not life-affirming, it’s vulgar, it’s mean, but I liked it.

⟶word order matters, but hard to encode in rules!

Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

⟶ hard to identify a priori which words are informative (and what information they carry!)

Sentiment: It's not life-affirming, it's vulgar, it's mean, but I liked it.

⟶word order matters, but hard to encode in rules!

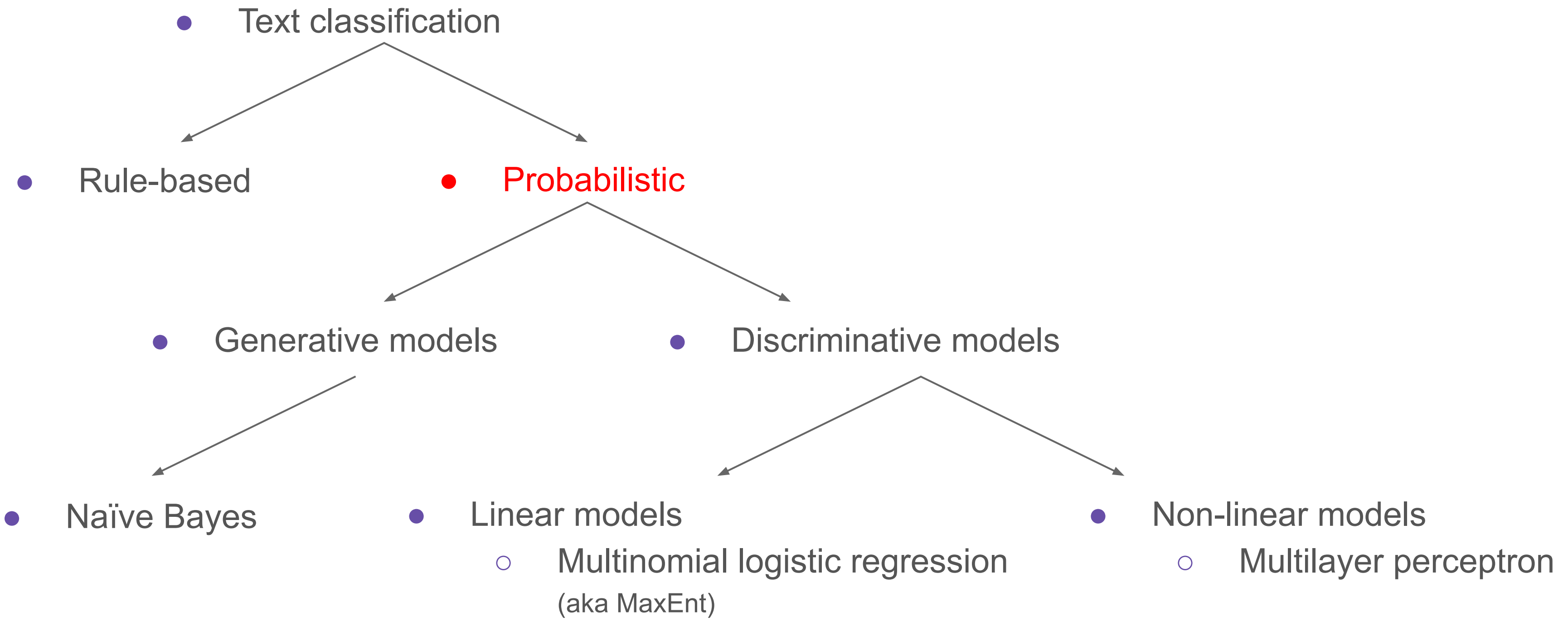
Language ID: All falter, stricken in kind. "LINGERIE SALE"

⟶simple features can be misleading!

Rule-based classification

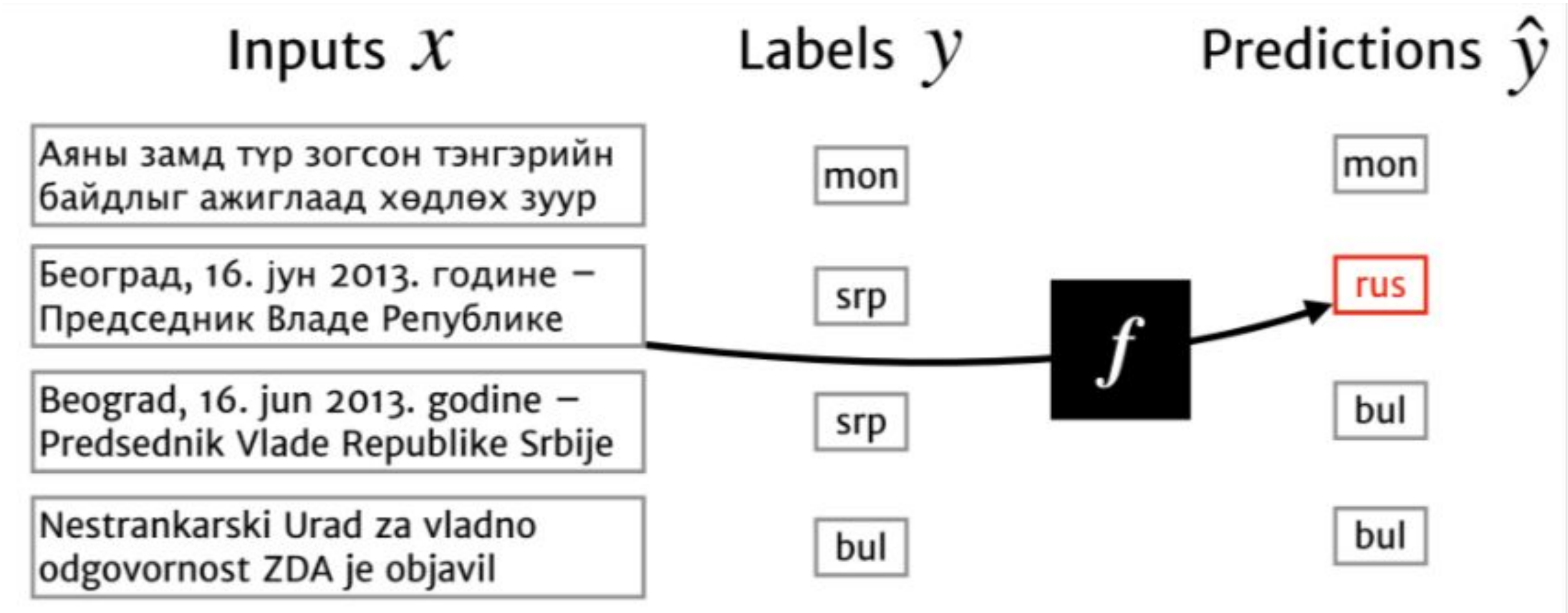
But don't forget: if you don't have access to data, speaker intuition and a bit of coding get you pretty far!

We’ll consider alternative models for classification



- Text classification
- Rule-based
- Probabilistic
- Generative models
- Discriminative models
- Naïve Bayes
- Linear models
    - Multinomial logistic regression (aka MaxEnt)
- Non-linear models
    - Multilayer perceptron

Learning-based classification



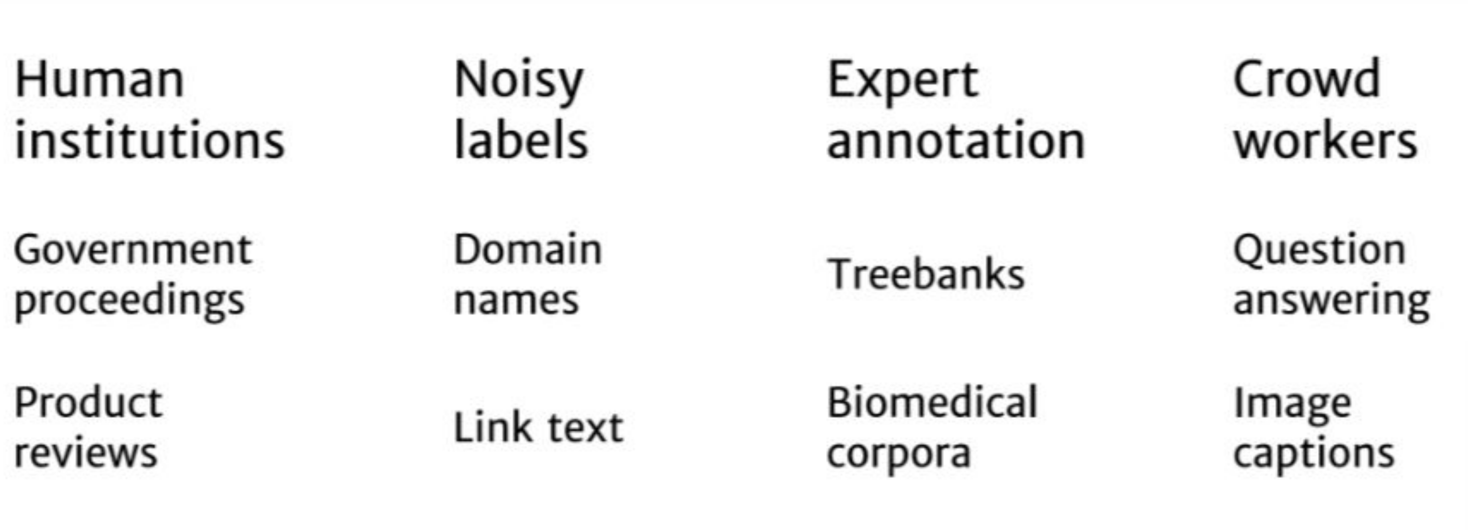
Inputs x	Labels y	Predictions $\hat{y}$
Аяны замд тур зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур	mon	mon
Београд, 16. јун 2013. године – Председник Владе Републике	srp	rus
Beograd, 16. jun 2013. godine – Predsednik Vlade Republike Srbije	srp	bul
Nestrankarski Urad za vladno odgovornost ZDA je objavil	bul	bul

Goal: pick the function f that does “best” on training data

Classification: learning from data

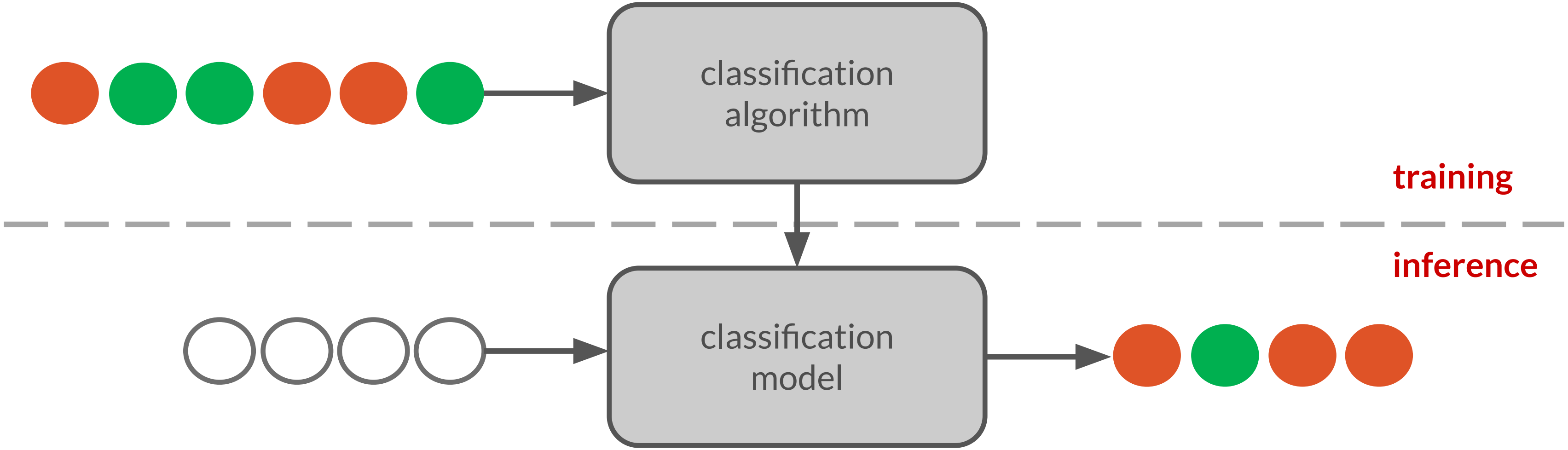
- Supervised
  - labeled examples
    - Binary (true, false)
    - Multi-class classification (politics, sports, gossip)
    - Multi-label classification (#party #FRIDAY #fail)
- Unsupervised
  - no labeled examples
- Semi-supervised
  - labeled examples + non-labeled examples
- Weakly supervised
  - heuristically-labeled examples

Where do datasets come from?

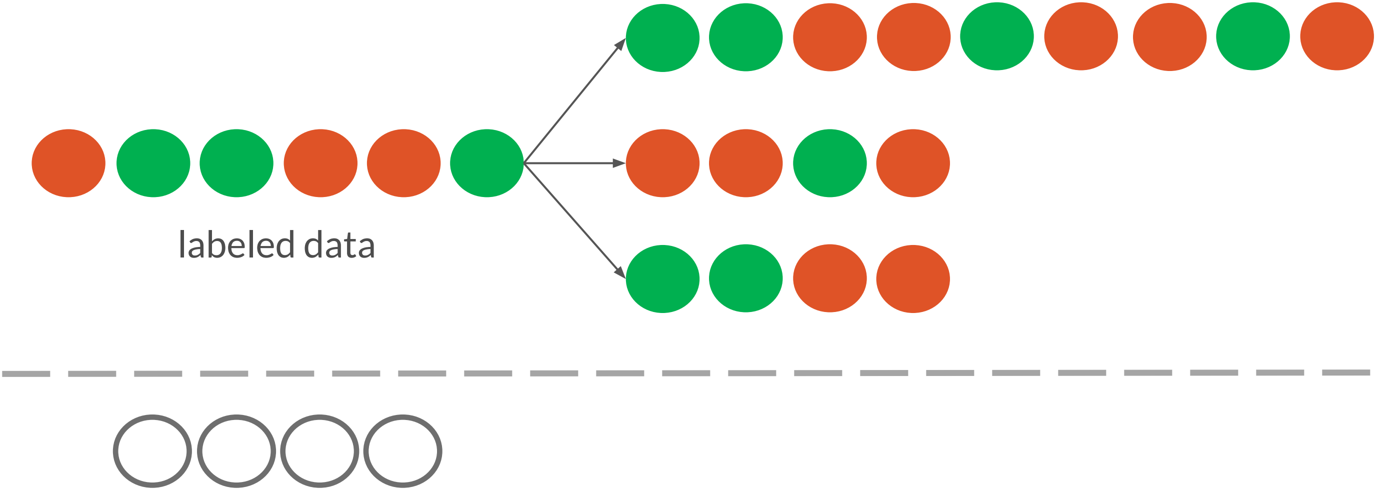


Human institutions	Noisy labels	Expert annotation	Crowd workers
Government proceedings	Domain names	Treebanks	Question answering
Product reviews	Link text	Biomedical corpora	Image captions

Supervised classification



Training, validation, and test sets



training set

training

validation set

labeled data

test set

inference

unlabeled data

Classification: features (measurements)

- Perform measurements and obtain features

4.2, 212, 3.4, 1332

↓ ↓ ↓ ↓

diameter, weight, softness, color

5.2, 315, 5.7, 4567

↓ ↓ ↓ ↓

diameter, weight, softness, color

Supervised classification: formal setting

- Learn a classification model from labeled data on
  - properties (“features”) and their importance (“weights”)
- X: set of attributes or features $\{x_1, x_2, \ldots, x_n\}$
  - e.g. fruit measurements, or word counts extracted from an input documents
- y: a “class” label from the label set $Y = \{y_1,y_2, \ldots, y_k\}$
  - e.g., fruit type, or spam/not spam, positive/negative/neutral

Supervised classification: formal setting

- Learn a classification model from labeled data on
  - properties (“features”) and their importance (“weights”)
- X: set of attributes or features $\{x_1, x_2, \ldots, x_n\}$
  - e.g. fruit measurements, or word counts extracted from an input documents
- y: a “class” label from the label set $Y = \{y_1,y_2, \ldots, y_k\}$
  - e.g., fruit type, or spam/not spam, positive/negative/neutral

- Given data samples $\{x_1, x_2, \ldots, x_n\}$ and corresponding labels $Y = \{y_1,y_2, \quad \ldots, y_k\}$
- We train a function $f: x \in X \rightarrow y \in Y$ (the model)

Supervised classification: formal setting

- Learn a classification model from labeled data on
  - properties (“features”) and their importance (“weights”)
- X: set of attributes or features $\{x_1, x_2, \ldots, x_n\}$
  - e.g. fruit measurements, or word counts extracted from an input documents
- y: a “class” label from the label set $Y = \{y_1, y_2, \ldots, y_k\}$
  - e.g., fruit type, or spam/not spam, positive/negative/neutral

- At inference time, apply the model on new instances to predict the label

Text classification – feature extraction

What can we measure over text? Consider this movie review:

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 just to 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 before.

Text classification – feature extraction

What can we measure over text? Consider this movie review:

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 just to 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 before.

Text classification – feature extraction

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 just to 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 before.

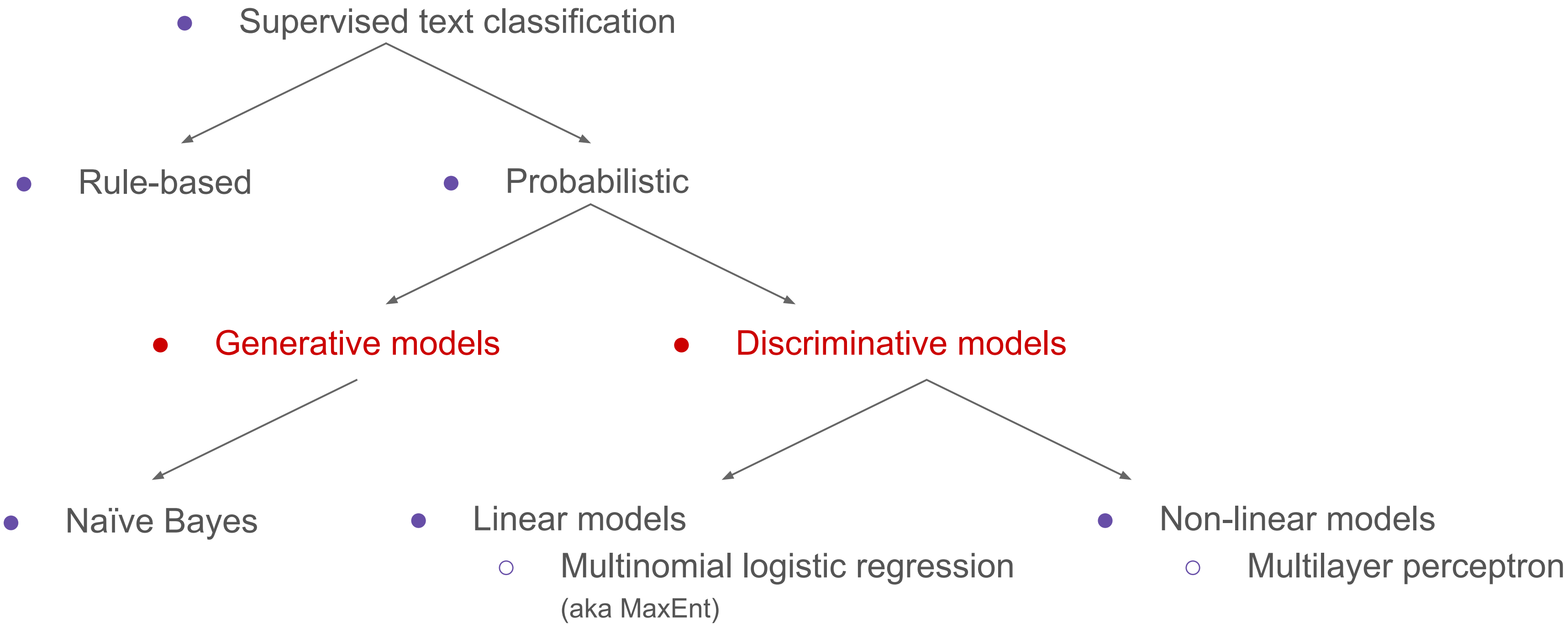
(almost) the entire lexicon

word	count	relative frequency
love	10	0.0007
great	…	
recommend		
laugh		
happy		
…		
several		
boring		
…		

Types of textual features

- Words
    - content words, stop-words
    - punctuation? tokenization? lemmatization? lowercase?
- Word sequences
    - bigrams, trigrams, n-grams
- Grammatical structure, sentence parse tree
- Words’ part-of-speech
- Word vectors
- …

We’ll consider alternative models for classification



- Supervised text classification
- Rule-based
- Probabilistic
- Generative models
- Discriminative models
- Naïve Bayes
- Linear models
    - Multinomial logistic regression (aka MaxEnt)
- Non-linear models
    - Multilayer perceptron

Generative and discriminative models

- Generative model: a model that calculates the probability of the input data itself

$$P(X, Y)$$

joint

- Discriminative model: a model that calculates the probability of a latent trait given the data

$$P(Y \mid X)$$

conditional

Generative and discriminative models

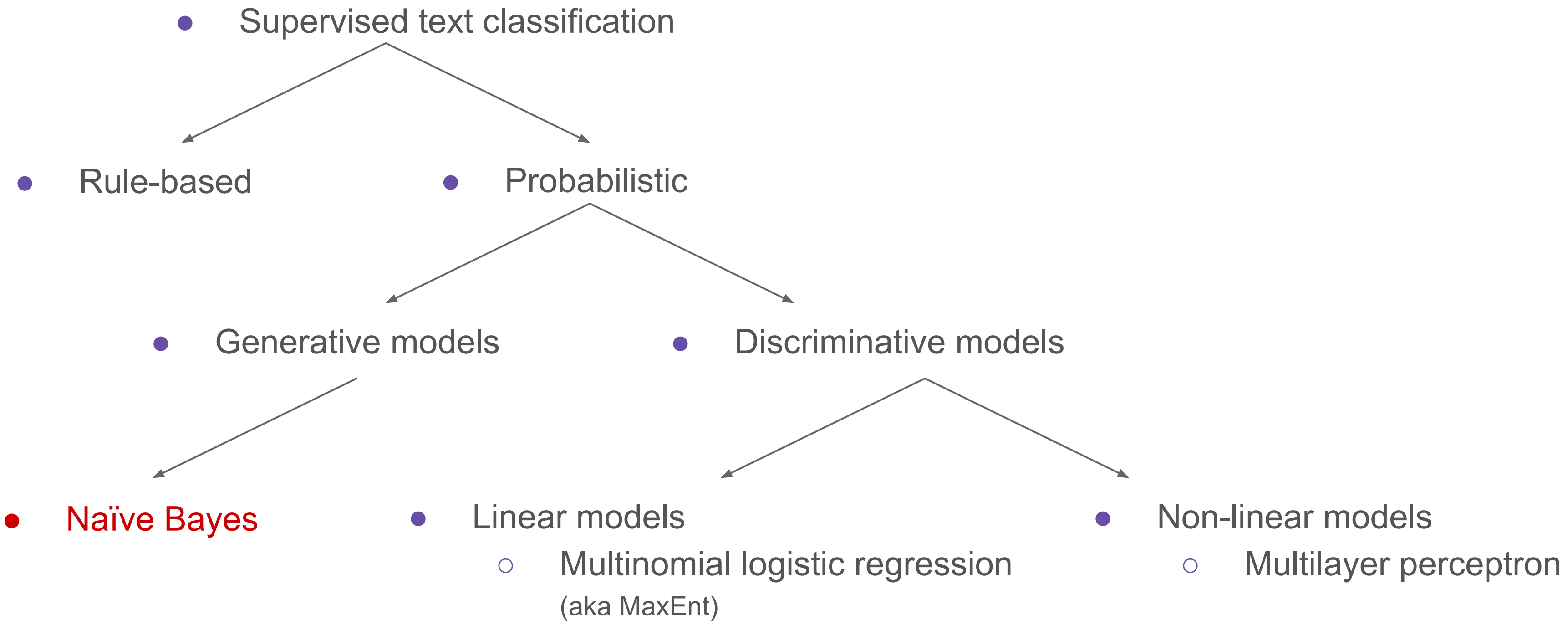
- Generative text classification: Learn a model of the joint P(X, y), and find

$$\hat{y} = \underset{\tilde{y}}{\text{argmax}}\ P(X, \tilde{y})$$

- Discriminative text classification: Learn a model of the conditional P(y | X), and find

$$\hat{y} = \underset{\tilde{y}}{\text{argmax}}\ P(\tilde{y}|X)$$

We'll consider alternative models for classification



- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Generative text classification: naïve Bayes

- Simple (naïve) classification method
  - based on the Bayes rule
- Relies on very simple representation of a documents
  - bag-of-words, no relative order
- A good baseline for more sophisticated models

Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems 14 (NIPS), 2001.

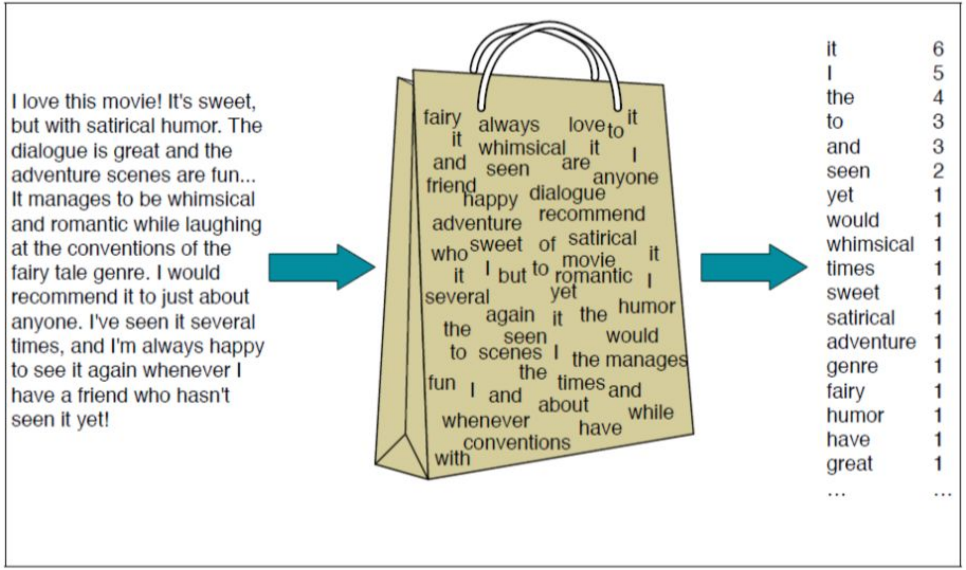
Naïve Bayes

Sentiment analysis: movie reviews

- Given a document d (e.g., a movie review)
- Decide which class c it belongs to: positive, negative, neutral
- Compute P(c | d) for each c
    - P(positive | d), P(negative | d), P(neutral | d)
    - select the one with max P

Bag-of-Words (BOW)

- Given a document d (e.g., a movie review) – how to represent d ?



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
...	...

Figure 7.1 Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.

Figure from J&M 3rd ed. draft, sec 7.1

Possible representations for text

- Bag-of-Words (BOW)
  - Easy, no effort required
  - Variable size, ignores sentential structure
- Hand-crafted features
  - Full control, can use NLP pipeline, class-specific features
  - Over-specific, incomplete, makes use of NLP pipeline
- Learned feature representations
  - Can learn to contain all relevant information
  - Needs to be learned

Naïve Bayes

- Given a document d and a class c, Bayes' rule:

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

Naïve Bayes

- Given a document d and a class c, Bayes' rule:

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

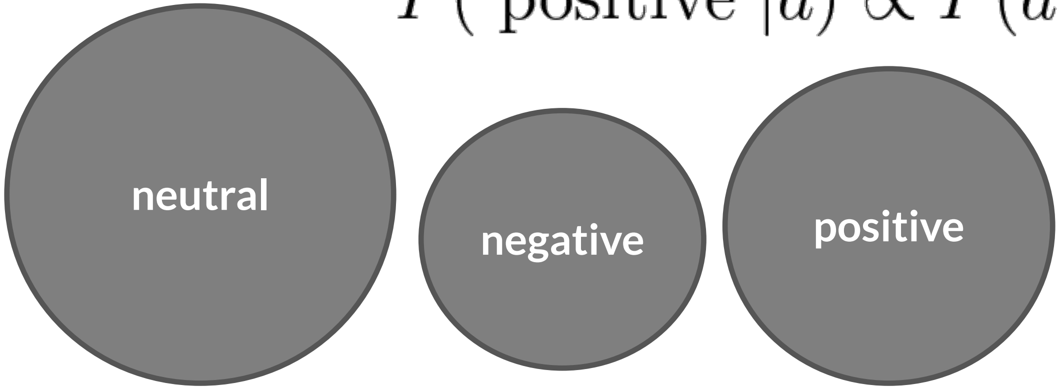
$$P(\text{‘positive’}|d) \propto \underset{\substack{\big\downarrow \\ \text{likelihood}}}{P(d|\text{‘positive’})}\,\underset{\substack{\big\downarrow \\ \text{prior}}}{P(\text{‘positive’})}$$

Naïve Bayes

- Given a document d and a class c, Bayes' rule:

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

$$P(\text{‘positive’}|d) \propto P(d|\text{‘positive’})\underbrace{P(\text{‘positive’})}_{\text{prior}}$$



Naïve Bayes

- Given a document d and a class c, Bayes' rule:

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

$$P(\text{‘positive’}|d) \propto \underset{\downarrow}{P(d|\text{‘positive’})}P(\text{‘positive’})$$

likelihood

Naïve Bayes independence assumptions

$$P(w_1, w_2, \ldots, w_n | c)$$

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

$$P(w_1, w_2, \ldots, w_n | c) = P(w_1|c) \times P(w_2|c) \times P(w_3|c) \times \ldots \times P(w_n|c)$$

Document 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!

→ **bag of words (BOW)** →

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
…	…

Document 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!

→ **bag of words (BOW)** →

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
…	…

$$P(d|c) = P(w_1, w_2, \ldots, w_n|c) = \prod_i P(w_i|c)$$

Generative text classification: Naïve Bayes

$$\mathrm{C}_{NB} = \underset{c}{\operatorname{argmax}} P(c|d) = \underset{c}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)} \propto$$

Bayes rule

$$\underset{c}{\operatorname{argmax}} P(d|c)P(c) =$$

same denominator

$$\underset{c}{\operatorname{argmax}} P(w_1, w_2, \ldots, w_n|c)P(c) =$$

representation

$$\underset{c_j}{\operatorname{argmax}} \; P(c_j) \prod_i P(w_i|c)$$

conditional independence

Underflow prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow
- Since log(xy) = log(x) + log(y)
    - better to sum logs of probabilities instead of multiplying probabilities
- Class with highest un-normalized log probability score is still most probable

$$\mathrm{C}_{NB} = \underset{c_j}{\operatorname{argmax}}\, P(c_j) \textstyle\prod_i P(w_i|c)$$

$$\mathrm{C}_{NB} = \underset{c_j}{\operatorname{argmax}}\, log(P(c_j)) + \textstyle\sum_i log(P(w_i|c))$$

- Model is now just max of sum of weights

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?
- We learn $P(c)$ and $P(w_i|c)$ from training (labeled) data

$$\mathrm{C}_{NB} = \underset{c_j}{\operatorname{argmax}}\, log(\underline{P(c_j)}) + \sum_i log(\underline{P(w_i|c)})$$

Parameter estimation

- Parameter estimation during training
- Concatenate all documents with category c into one mega-document
- Use the frequency of w_i in the mega-document to estimate the word probability

$$\mathrm{C}_{NB} = \underset{c_j}{\operatorname{argmax}}\, log(P(c_j)) + \sum_i log(P(w_i|c))$$

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

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

Parameter estimation

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

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

- Create mega-document for topic j by concatenating all docs in this topic
  - Use frequency of w in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

$$\hat{P}(\text{“}fantastic\text{”}|c = \text{positive}) = \frac{count(\text{“}fantastic\text{”}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

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

$$\underset{c_j}{\text{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

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

Multinomial naïve Bayes : learning

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
    - For each c_j do
        - $docs_j \leftarrow$ all docs with class = c_j
        - $P(c_j) \leftarrow \dfrac{|docs_j|}{total\ \#\ documents}$

Multinomial naïve Bayes : learning

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
  - For each c_j do
    - $docs_j \leftarrow$ all docs with class = c_j
    - $P(c_j) \leftarrow \dfrac{|docs_j|}{total\ \#\ documents}$
- Calculate $P(w_i | c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_i in *Vocabulary*
    - $n_i \leftarrow$ # of occurrences of w_i in $Text_j$
    - $P(w_j | c_j) \leftarrow \dfrac{n_i + \alpha}{n + \alpha|Vocabulary|}$

Example

	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 Chinese Chinese Tokyo Japan	?

Example

	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 Chinese Chinese Tokyo Japan	?

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

Priors:

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

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

Example

	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 Chinese Chinese Tokyo Japan	?

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

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

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

Example

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

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

	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 Chinese Chinese 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

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

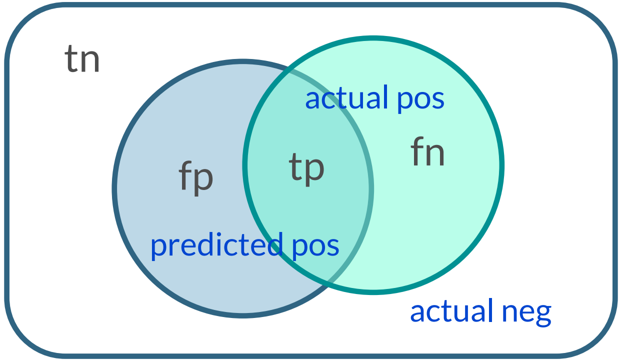
Summary: naïve Bayes is not so naïve

- Naïve Bayes is a probabilistic model
- Naïve because is assumes features are independent of each other for a class
- Very fast, low storage requirements
- Robust to Irrelevant Features
    - Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features
    - Decision Trees suffer from fragmentation in such cases – especially if little data
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
    - But we will see other classifiers that give better accuracy

Classification evaluation

- Contingency table: model's predictions are compared to the correct results
  - a.k.a. confusion matrix

	actual pos	actual neg
predicted pos	true positive (tp)	false positive (fp)
predicted neg	false negative (fn)	true negative (tn)



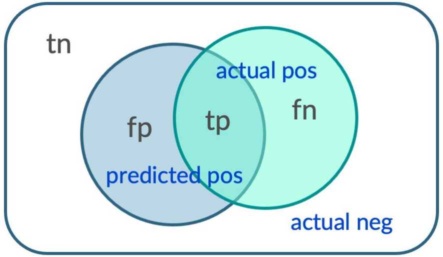
Classification evaluation

- Borrowing from Information Retrieval, empirical NLP systems are usually evaluated using the notions of precision and recall

Classification evaluation

- Precision (P) is the proportion of the selected items that the system got right in the case of text categorization
  - it is the % of documents classified as “positive” by the system which are indeed “positive” documents
- Reported per class or average

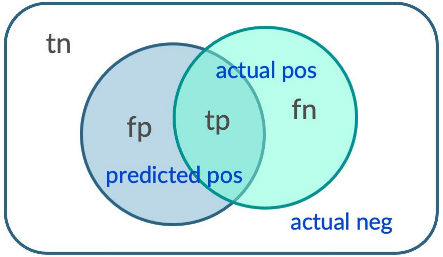
$$\text{precision} = \frac{true\ positives}{true\ positives + false\ positives} = \frac{tp}{tp + fp}$$



Classification evaluation

- Recall (R) is the proportion of actual items that the system selected in the case of text categorization
    - it is the % of the “positive” documents which were actually classified as “positive” by the system
- Reported per class or average

$$\text{recall} = \frac{true\ positives}{true\ positives + false\ negatives} = \frac{tp}{tp + fn}$$



Classification evaluation

- We often want to trade-off precision and recall
    - typically: the higher the precision the lower the recall
    - can be plotted in a precision-recall curve
- It is convenient to combine P and R into a single measure
    - one possible way to do that is F measure

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \qquad \text{for } \beta=1,\ F_1 = \frac{2PR}{P+R}$$

Classification evaluation

- Additional measures of performance: accuracy and error
  - accuracy is the proportion of items the system got right
  - error is its complement

$$\text{accuracy} = \frac{tp+tn}{tp+fp+tn+fn}$$

Micro- vs. macro-averaging

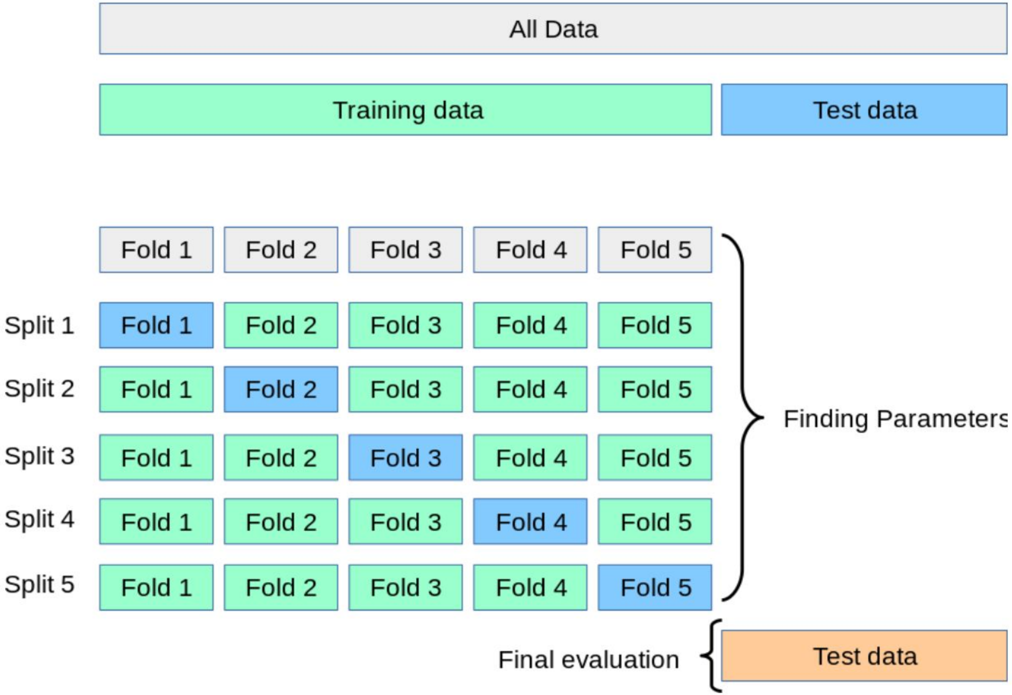
If we have more than one class, how do we combine multiple performance measures into one quantity?

- Macroaveraging
  - Compute performance for each class, then average.
- Microaveraging
  - Collect decisions for all classes, compute contingency table, evaluate.

Classification common practices

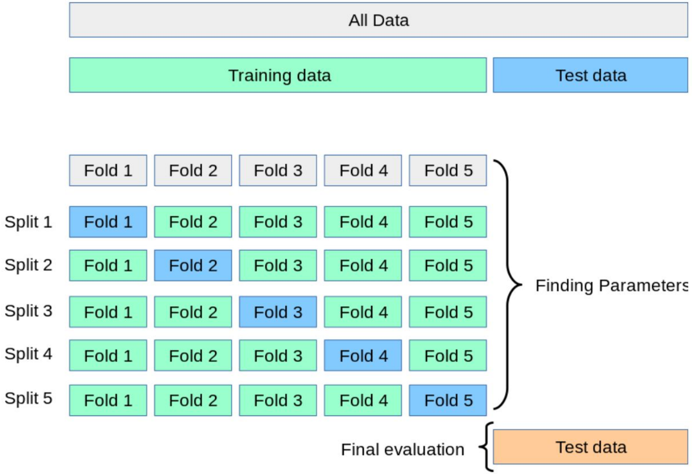
- Divide the training data into k folds (e.g., k=10)
- Repeat k times: train on k-1 folds and test on the holdout fold, cyclically
- Average over the k folds' results

K-fold cross-validation

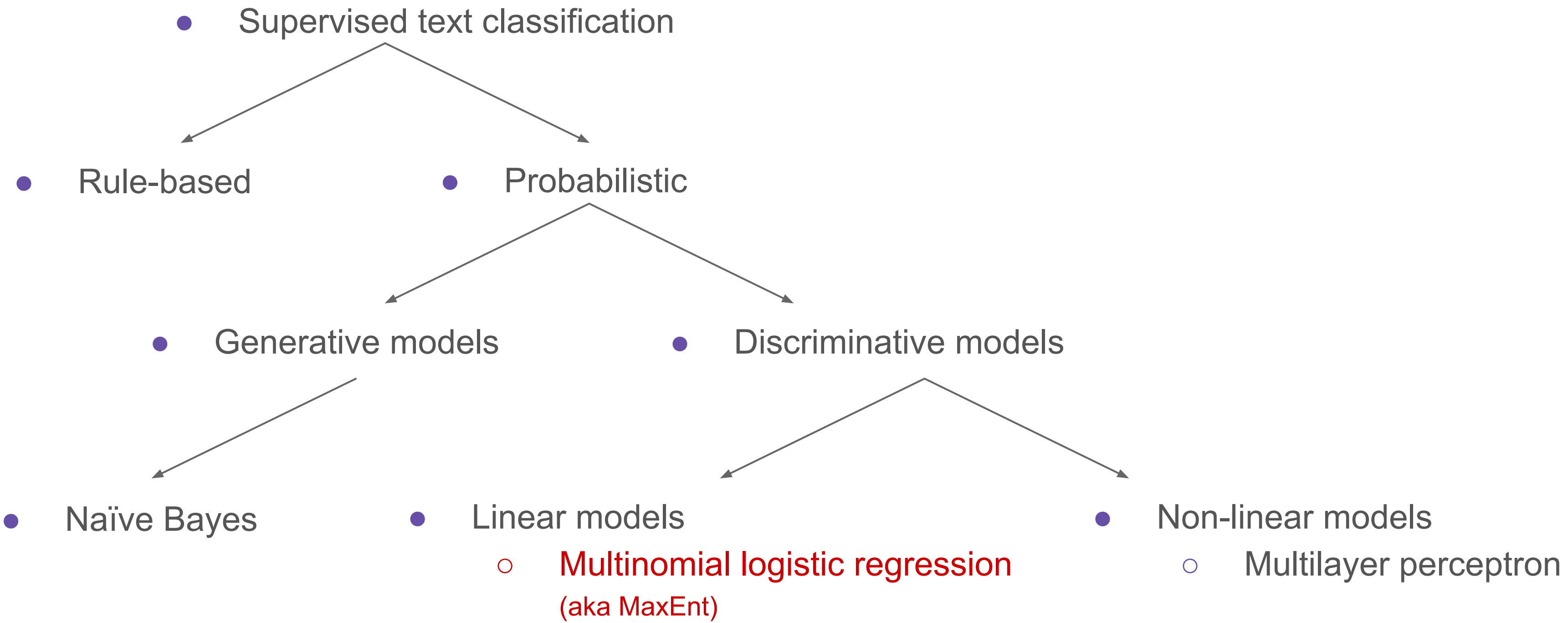


K-fold cross-validation

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting (‘tuning to the test set’)
  - more conservative estimate of performance
- Cross-validation over multiple splits
    - Handles sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance



Next class



- Supervised text classification
- Rule-based
- Probabilistic
- Generative models
- Discriminative models
- Naïve Bayes
- Linear models
    - Multinomial logistic regression (aka MaxEnt)
- Non-linear models
    - Multilayer perceptron

Readings

- Eis 2
- J&M III 4
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP, 2002
- Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, In Proceedings of NeurIPS, 2001.