Natural Language Processing

Text classification, Feature engineering

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Announcements

- Quiz 1:
  - Lecture materials through Monday
    - Linguistic structure
    - Language corpora properties
    - Intro to classification
    - Text feature engineering
- HW1 releases
  - Make sure to attend next lecture
  - Kabir will be giving an overview of HW1
  - Kavel will be showing how to access quiz on canvas

Text Classification

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 bee[n] extended to 10 March 2022 (It is firm date).

We would like to invite you to submit a pape[r] nference on Renewable Energy Systems (ECRES). **ECRES 202[2] e, the participants can present their papers physically o[n]** to be organized in Istanbul/Turkey under the tech[nical] Medeniyet University and many international institutions. The [conference] ternational with the participants from all continents and more than 40 [countries].

spam

The submission deadline and speci[al] 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.**

Invitation to present at the February 2022 Wikimedia Research Showcase

Emily Lescak <elescak@wikimedia.org> Tue, Nov 23, 2021, 11:00 AM
to yuliats@cs.washington.edu

Hi Yulia,

My name is Emily Lescak and I am a member of the Research team at the Wikimedia Foundation. On be[half] 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, ac[ademic] nt 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[media] n staff—would find interesting and relevant to their work.

Research Showcase presentations are generally 20 minutes long, with an additio[nal] vite 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[C] uTube 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

not spam

Language ID

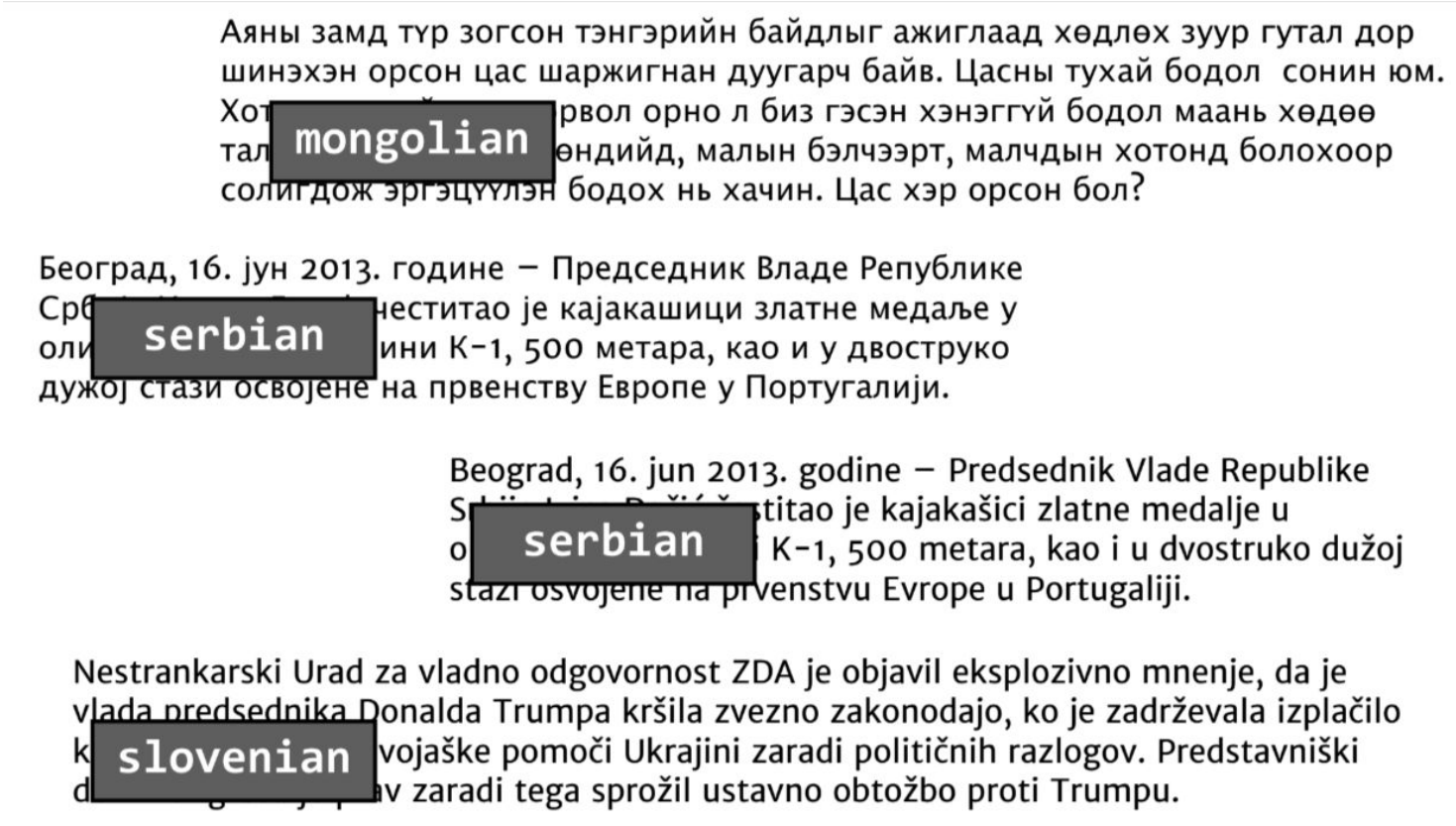
Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хөдөө талд, говийн ээрэм хөндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох 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?
mongolian
Београд, 16. јун 2013. године – Председник Владе Републике
Ср нештитао је кајакашици златне медаље у
оли ини К-1, 500 метара, као и у двоструко
дужој стази освојене на првенству Европе у Португалији.
serbian
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.
serbian
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.
slovenian

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 some other 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

MeSH Subject Category Hierarchy

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

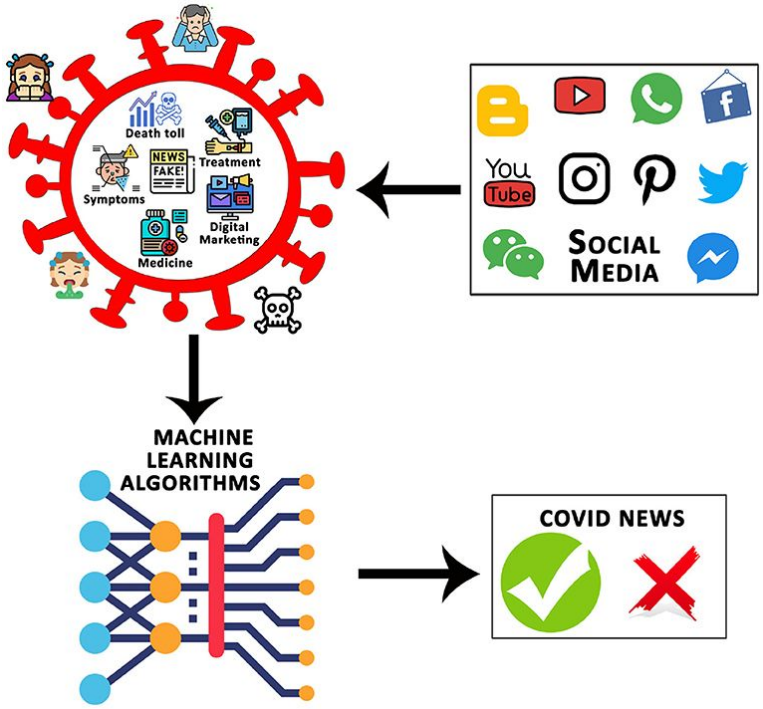
Authorship attribution: is the author male or female?

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?



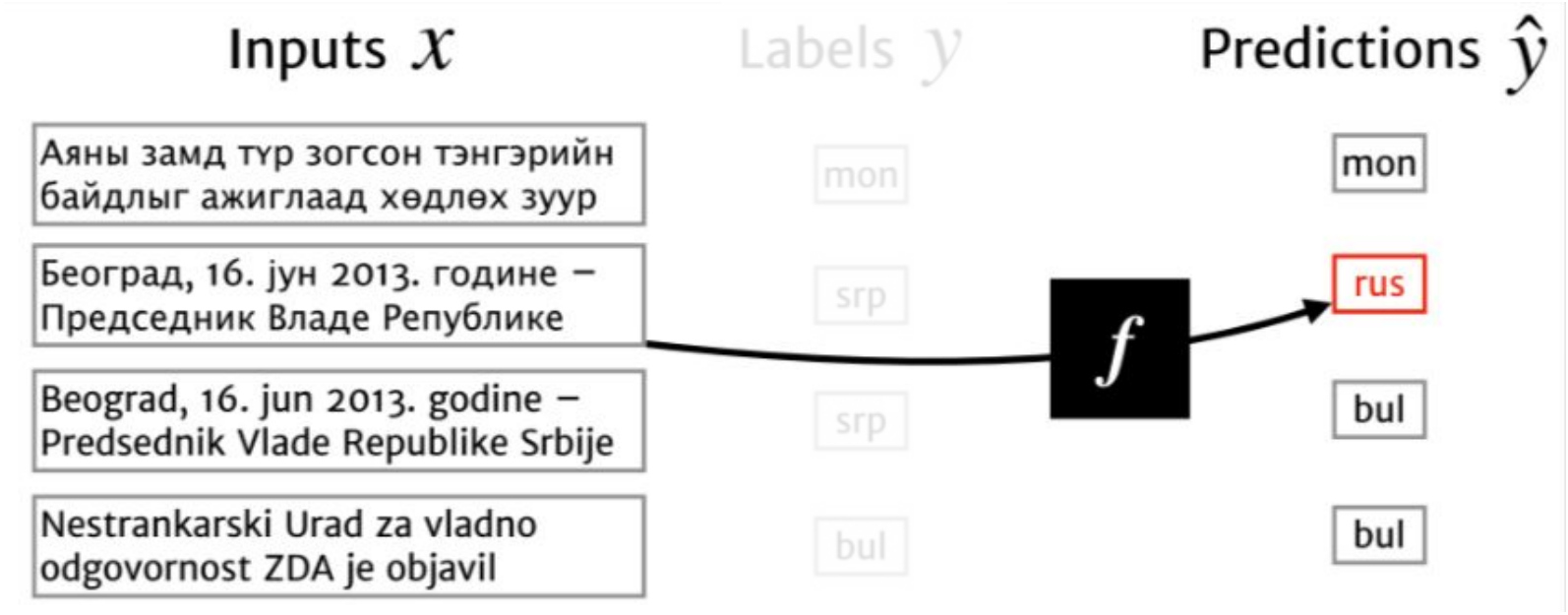
Detecting COVID-19-Related Fake News Using Feature Extraction

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)
  - …

Text classification



Goal: create a function f that makes a prediction $\hat{y}$ given an input x

Over the next couple of classes, we'll investigate:

1. How do we "digest" text into a form usable by a function?

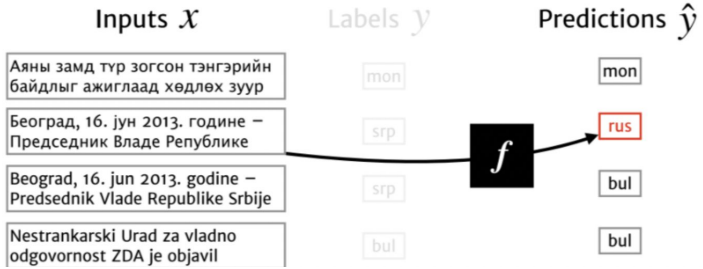
   (Keywords for this section: features, feature extraction, feature selection, representations)

2. What kinds of strategies might we use to create our function *f*?

   (Keyword for this section: models)

3. How do we evaluate our function *f*?

   (Keyword for this section: … evaluation)



How do we “digest” text into a form usable by a function?

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

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.

Bag-of-Words (BOW)

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

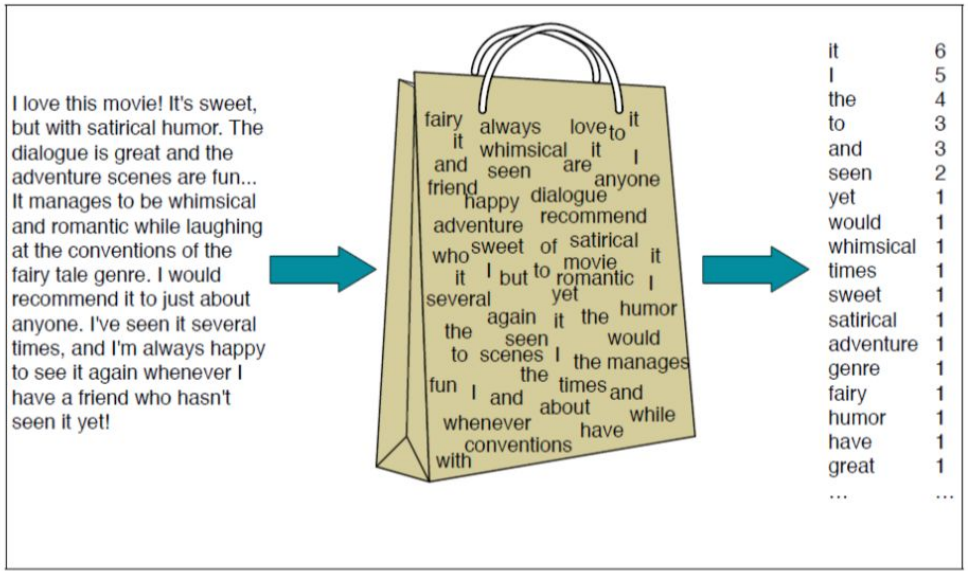


Figure from J&M 3rd ed. draft, sec 7.1

BOW feature extraction, independence assumption

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 beyond BOW

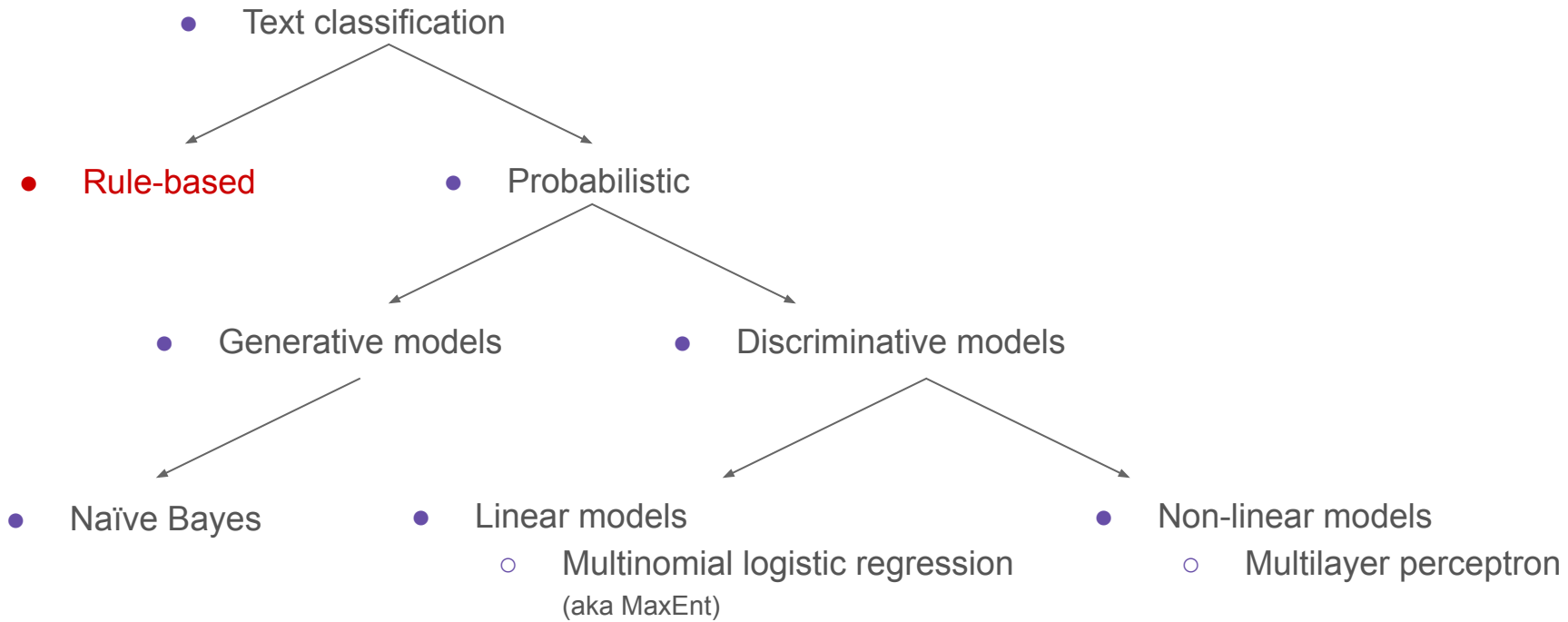
- Words
  - content words, stop-words
  - punctuation? tokenization? lemmatization? lowercase?
- Word sequences
  - bigrams, trigrams, n-grams
- Grammatical structure, sentence parse tree
- Words' part-of-speech
- Word vectors
- …

Summary: 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

What kinds of strategies might we use to create our function *f*?

We'll consider alternative models for classification



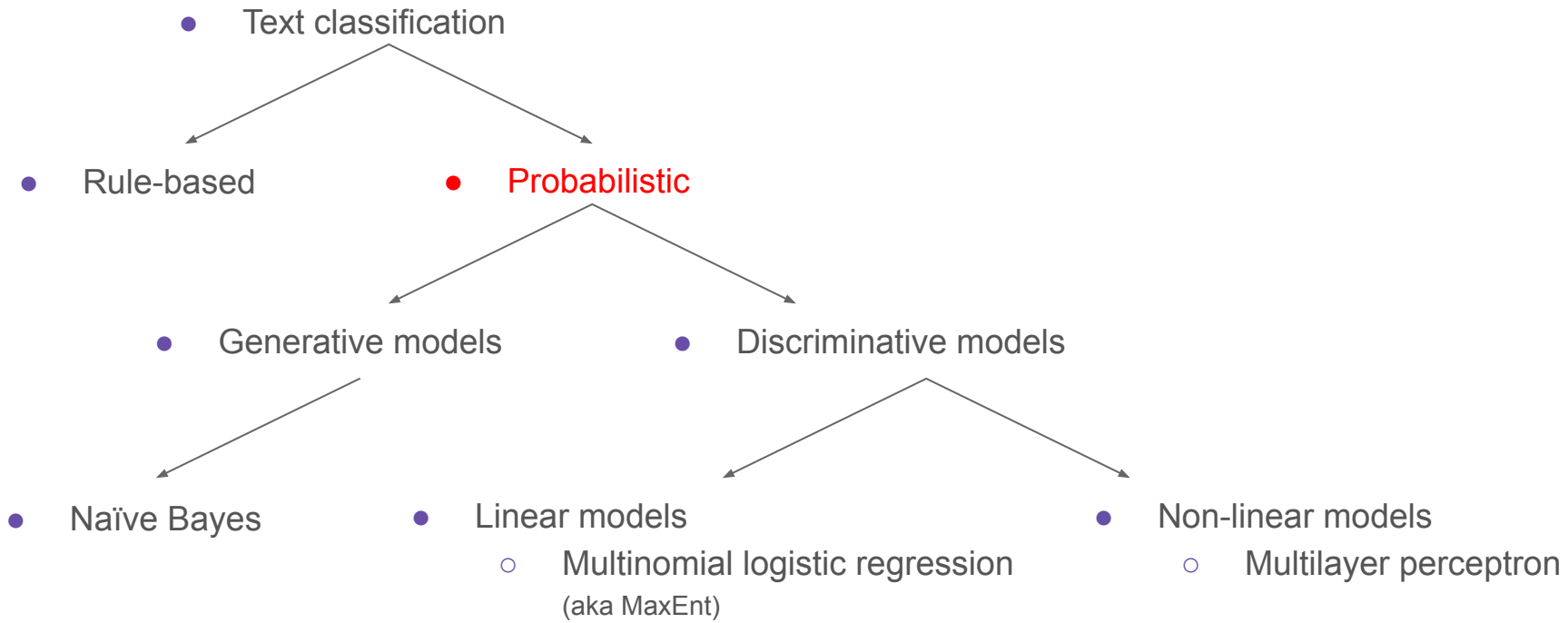
Rule-based classifier

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

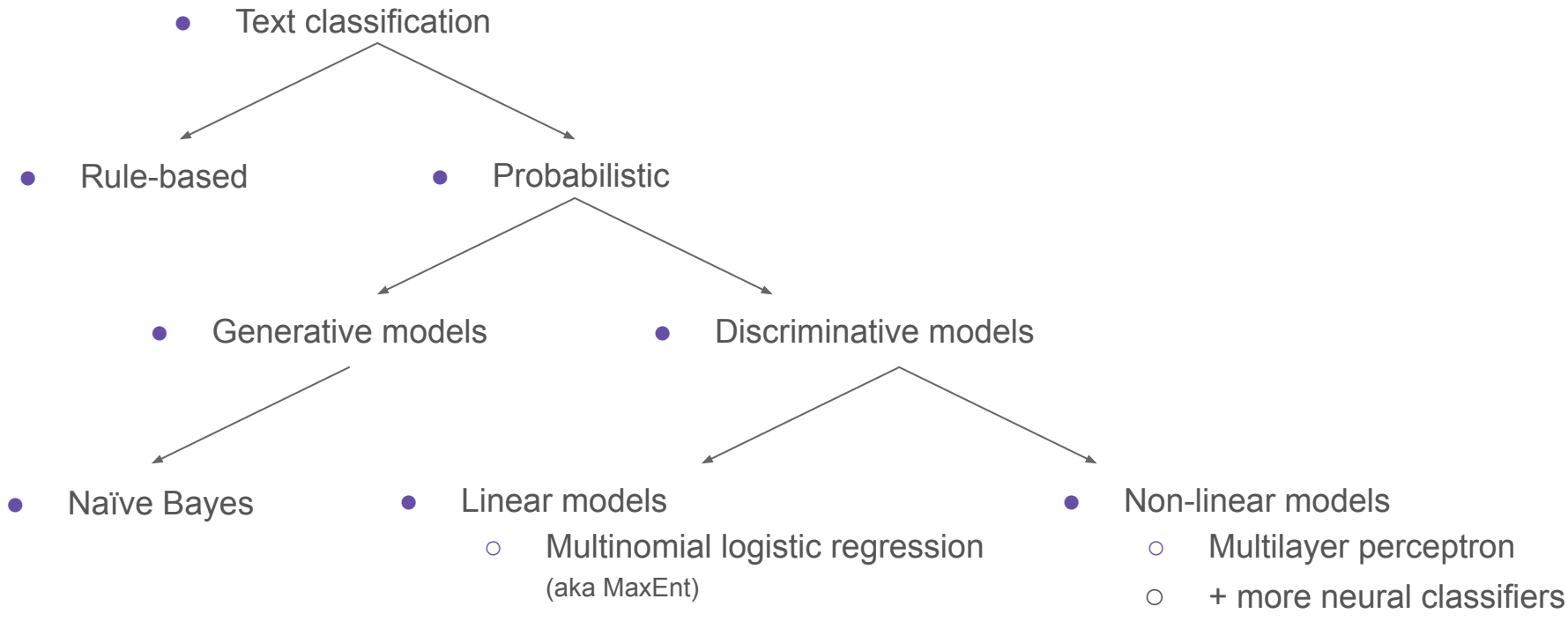
Rule-based classification

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

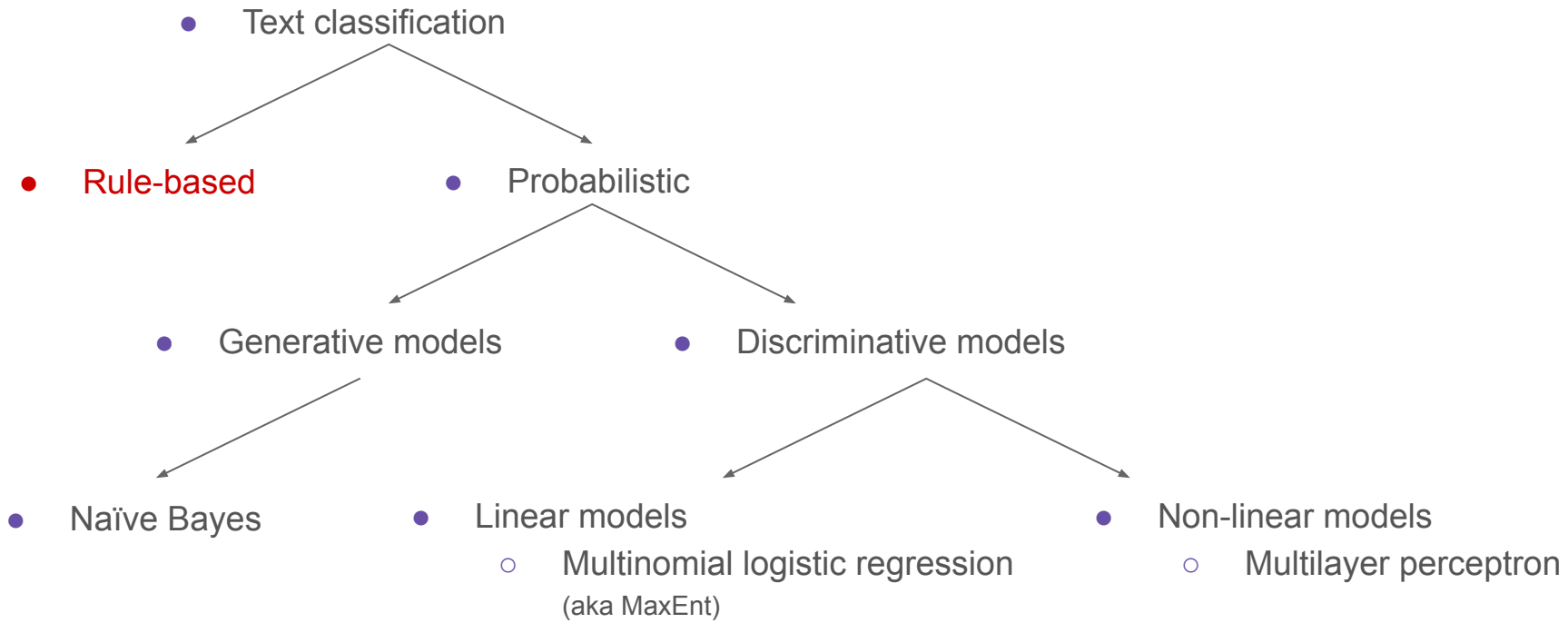
We'll consider alternative models for classification



We'll consider alternative models for classification



We'll consider alternative models for classification



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.

⟶language pragmatics is complex to model at word level, word order (syntax) 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.

⟶language pragmatics is complex to model at word level, word order (syntax) matters, but hard to encode in rules!

Language ID: All falter, stricken in kind.

⟶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