University of BRISTOL

2.1 Text Classification with Naïve Bayes

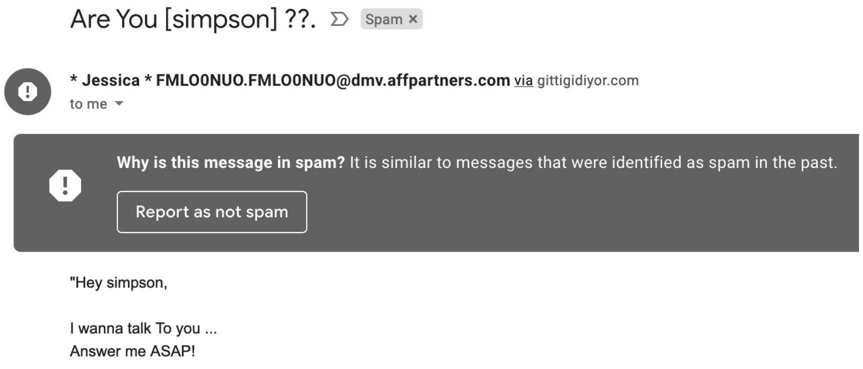
Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

Spam Filtering



Are You [simpson] ??.

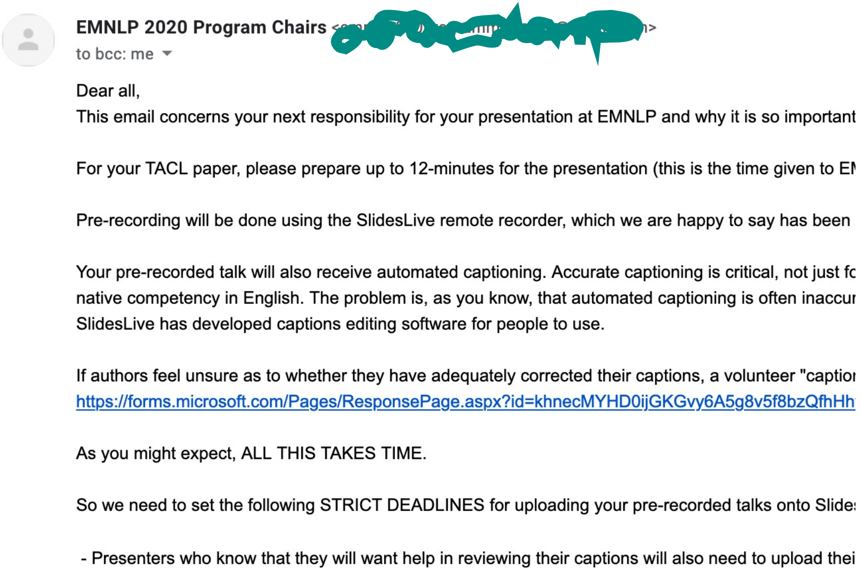
* Jessica * FMLO0NUO.FMLO0NUO@dmv.affpartners.com via gittigidiyor.com

to me

Why is this message in spam? It is similar to messages that were identified as spam in the past.

"Hey simpson,

I wanna talk To you ...
Answer me ASAP!



EMNLP 2020 Program Chairs <[redacted]>

to bcc: me

Dear all,
This email concerns your next responsibility for your presentation at EMNLP and why it is so important

For your TACL paper, please prepare up to 12-minutes for the presentation (this is the time given to EI

Pre-recording will be done using the SlidesLive remote recorder, which we are happy to say has been

Your pre-recorded talk will also receive automated captioning. Accurate captioning is critical, not just fc
native competency in English. The problem is, as you know, that automated captioning is often inaccur
SlidesLive has developed captions editing software for people to use.

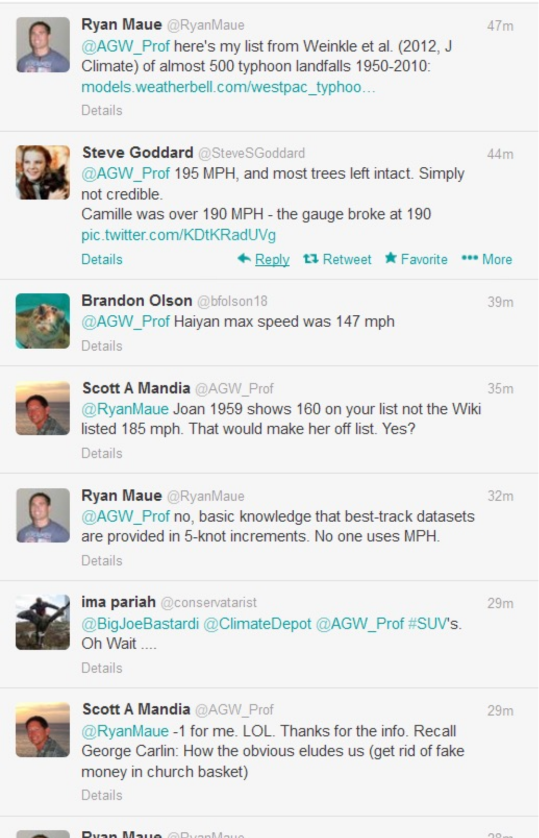
If authors feel unsure as to whether they have adequately corrected their captions, a volunteer "captioi
https://forms.microsoft.com/Pages/ResponsePage.aspx?id=khnecMYHD0ijGKGvy6A5g8v5f8bzQfhHh

As you might expect, ALL THIS TAKES TIME.

So we need to set the following STRICT DEADLINES for uploading your pre-recorded talks onto Slide

 - Presenters who know that they will want help in reviewing their captions will also need to upload thei

Social Media in a Crisis



Ryan Maue @RyanMaue 47m
@AGW_Prof here's my list from Weinkle et al. (2012, J Climate) of almost 500 typhoon landfalls 1950-2010: models.weatherbell.com/westpac_typhoo…
Details

Steve Goddard @SteveSGoddard 44m
@AGW_Prof 195 MPH, and most trees left intact. Simply not credible.
Camille was over 190 MPH - the gauge broke at 190
pic.twitter.com/KDtKRadUVg
Details

Brandon Olson @bfolson18 39m
@AGW_Prof Haiyan max speed was 147 mph
Details

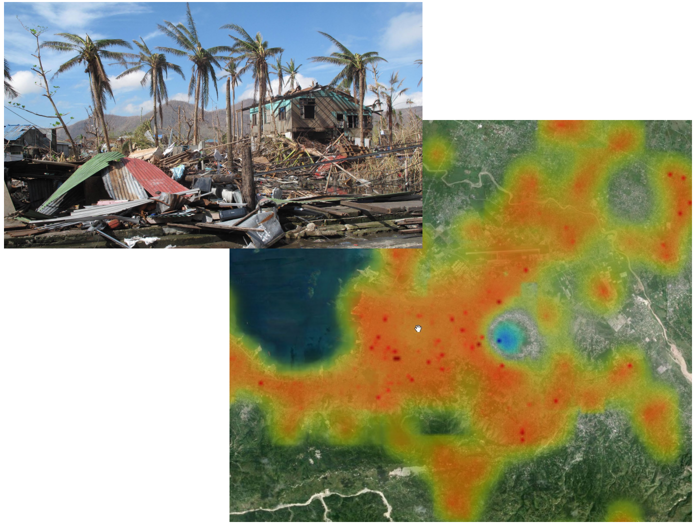
Scott A Mandia @AGW_Prof 35m
@RyanMaue Joan 1959 shows 160 on your list not the Wiki listed 185 mph. That would make her off list. Yes?
Details

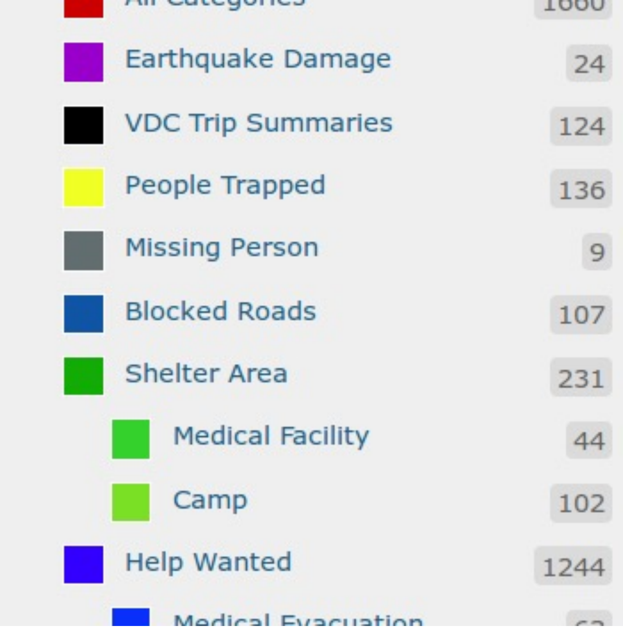
Ryan Maue @RyanMaue 32m
@AGW_Prof no, basic knowledge that best-track datasets are provided in 5-knot increments. No one uses MPH.
Details

ima pariah @conservatarist 29m
@BigJoeBastardi @ClimateDepot @AGW_Prof #SUV's.
Oh Wait ....
Details

Scott A Mandia @AGW_Prof 29m
@RyanMaue -1 for me. LOL. Thanks for the info. Recall George Carlin: How the obvious eludes us (get rid of fake money in church basket)
Details

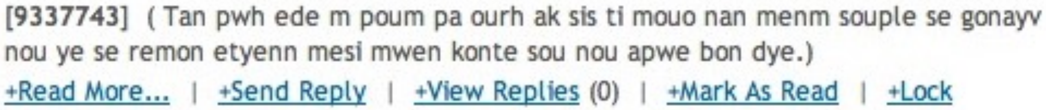
Ryan Maue @RyanMaue 28m





🟥	All Categories	1660
🟪	Earthquake Damage	24
⬛	VDC Trip Summaries	124
🟨	People Trapped	136
■	Missing Person	9
🟦	Blocked Roads	107
🟩	Shelter Area	231
🟩	Medical Facility	44
🟩	Camp	102
🟦	Help Wanted	1244
🟦	Medical Evacuation	

Categories of earthquake reports
Nepal, 2015, Quakemap.org



[9337743] (Tan pwh ede m poum pa ourh ak sis ti mouo nan menm souple se gonayv nou ye se remon etyenn mesi mwen konte sou nou apwe bon dye.)

+Read More... | +Send Reply | +View Replies (0) | +Mark As Read | +Lock

Ushahidi – From Haiti 2010 earthquake
Morrow et al. 2011

Sentiment Analysis

- Classify movie reviews as positive or negative:

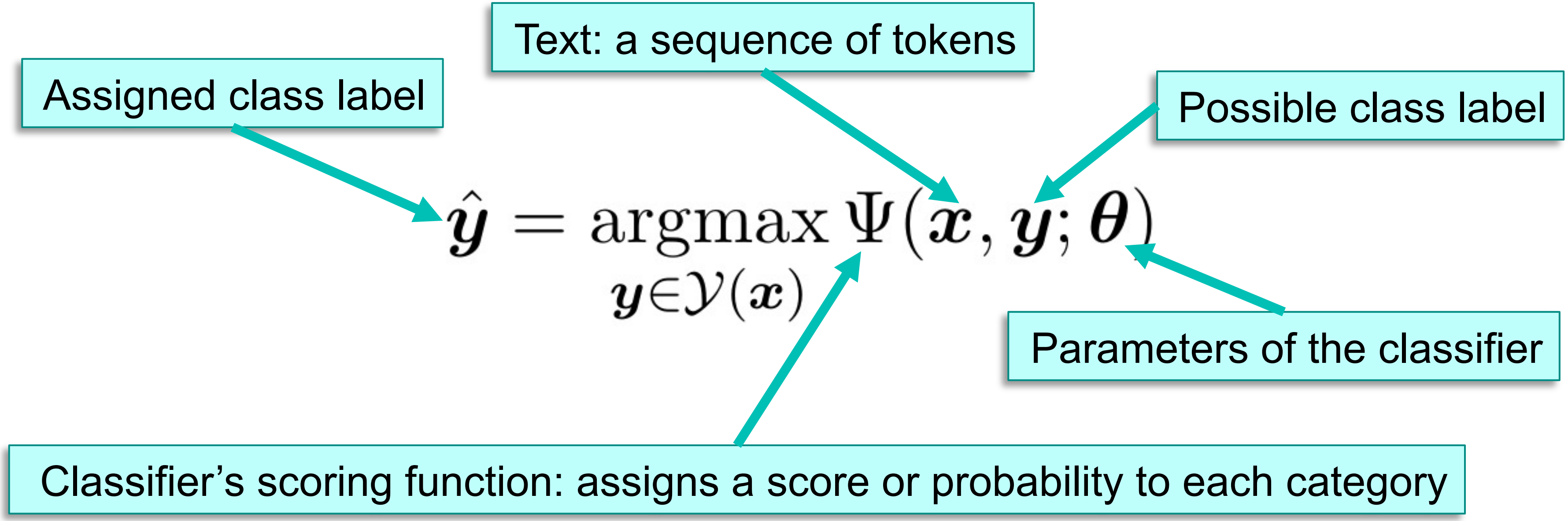
+	...zany characters and richly applied satire, and some great plot twists
−	It was pathetic. The worst part about it was the boxing scenes...
+	...awesome caramel sauce and sweet toasty almonds. I love this place!
−	...awful pizza and ridiculously overpriced...

Excerpts from movie reviews. Chapter 4, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2021).

Text Classification

- Identify sentences or documents of interest from a large **corpus.**
- Categorise documents so that we can process them differently.
- Perform statistical analysis on the documents in each category.

Text Classification



$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\operatorname{argmax}} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Text: a sequence of tokens

Assigned class label

Possible class label

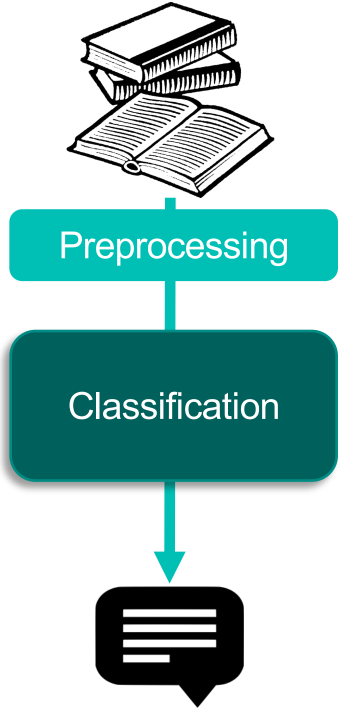
Parameters of the classifier

Classifier’s scoring function: assigns a score or probability to each category

[Section 1.2.2 and Chapter 2 of Introduction to Natural Language Processing, Eisenstein, J. (2019).]

Naïve Bayes Classifier

- A generative probabilistic model: learn what the documents in each class look like;
- Use **Bayes' rule** to determine class probabilities.
- Make some very strong simplifying assumptions:
  - Documents are represented as a **bag of words**
  - Features are **conditionally independent**



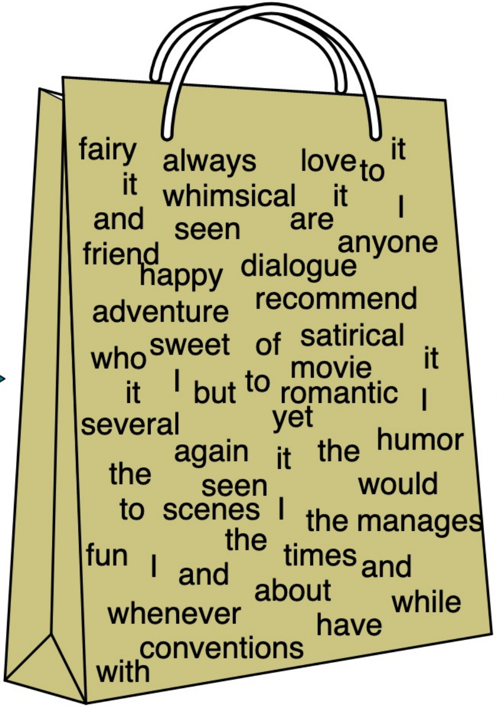
Preprocessing

Classification

Shallow text processing

Bag of Words Representation

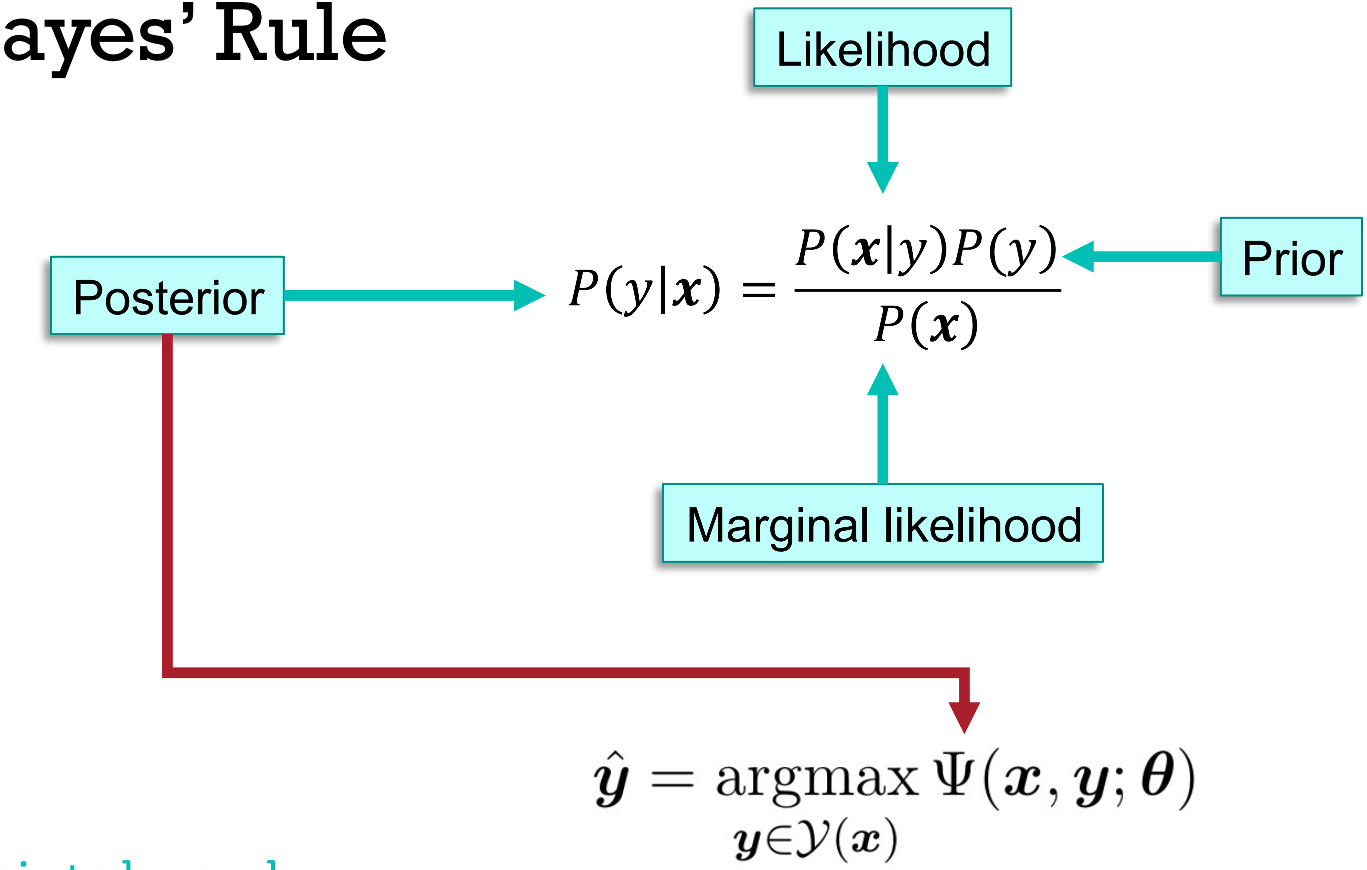
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy 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
…	…

Chapter 4, Speech and Language Processing (3rd edition draft), Jurafsky & Martin (2021).

Bayes' Rule



Likelihood

Posterior

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

Prior

Marginal likelihood

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

Naïve Bayes Classifier

$$P(y|\boldsymbol{x}) = \frac{\prod_{i=1}^{N} P(x_i|y)P(y)}{\sum_{y'} \prod_{i=1}^{N} P(x_i|y')\, P(y')}$$

- Bag of words ignores the word order when computing likelihood.
- The **naïve Bayes** assumption: the likelihood of each word is **conditionally independent** of the other words given the class label.
- Makes it simple to learn each term and fast to make predictions.

Maximum Likelihood Estimation

$$P(y = c) = \frac{\text{num\_docs\_in\_class\_c}}{\text{total\_num\_docs}}$$

$$P(x_i = w | y = c) = \frac{count(w\ in\ c) + 1}{\sum_{w' \in V}(count(w'\ in\ c) + 1)}$$

Summary

- Text classification has a wide range of applications including spam filtering, detecting crisis reports and sentiment analysis.
- Naïve Bayes' is a generative probabilistic model for classification.
- It makes strong assumptions: bag of words document representation; the naïve Bayes conditional independence assumption.
- The assumptions makes for a simple learning algorithm and the model performs well in many tasks despite these assumptions.