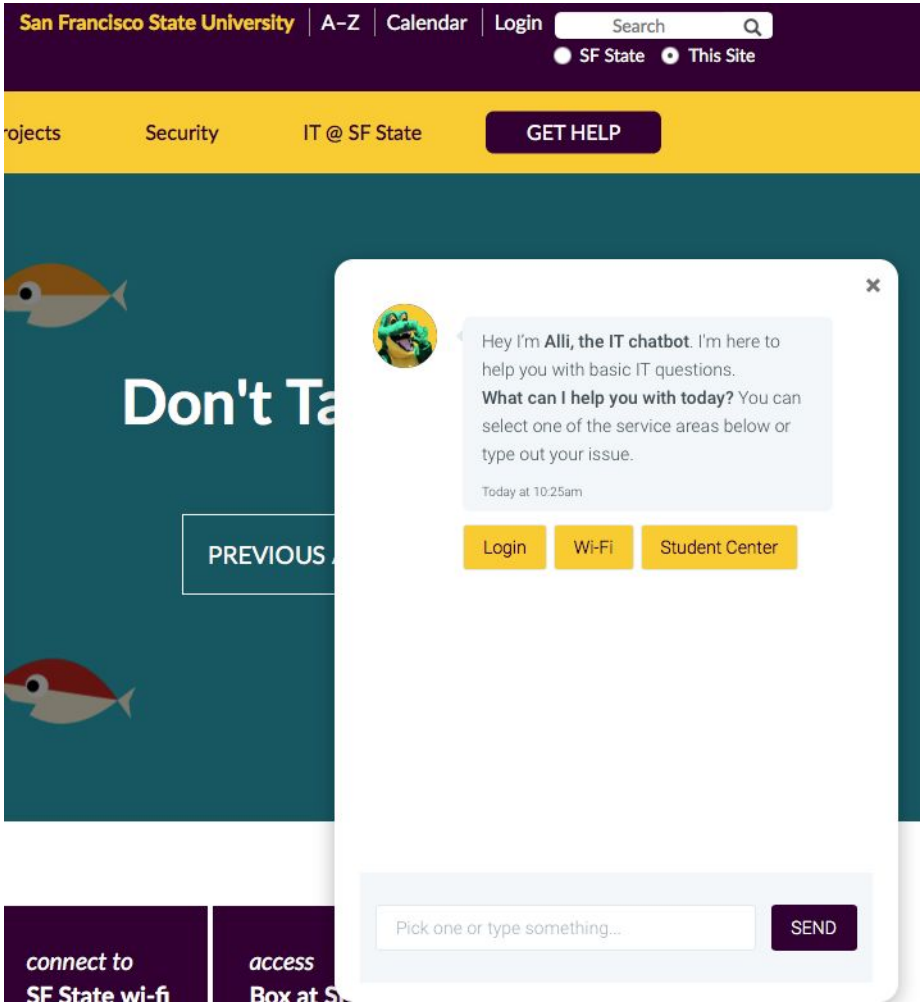
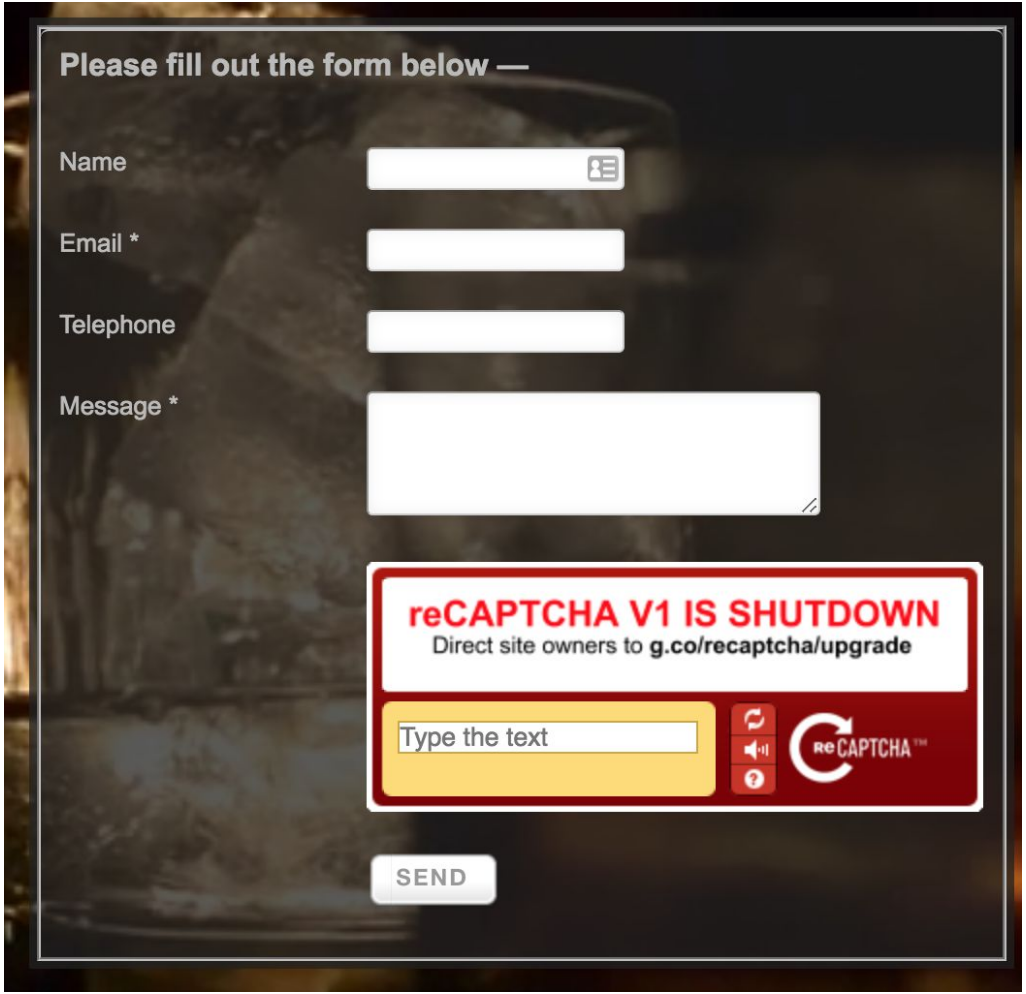
NLP for Online Conversation

Katie Bauer
Senior Data Scientist, Reddit
@imightbemary


San Francisco State University
A–Z
Calendar
Login
Search
SF State
This Site
rojects
Security
IT @ SF State
GET HELP
Don't Ta
PREVIOUS
Hey I'm Alli, the IT chatbot. I'm here to help you with basic IT questions.
What can I help you with today? You can select one of the service areas below or type out your issue.
Today at 10:25am
Login
Wi-Fi
Student Center
Pick one or type something...
SEND
connect to
SF State wi-fi
access
Box at S

Source: SFSU


Please fill out the form below —
Name
Email *
Telephone
Message *
reCAPTCHA V1 IS SHUTDOWN
Direct site owners to g.co/recaptcha/upgrade
Type the text
reCAPTCHA™
SEND

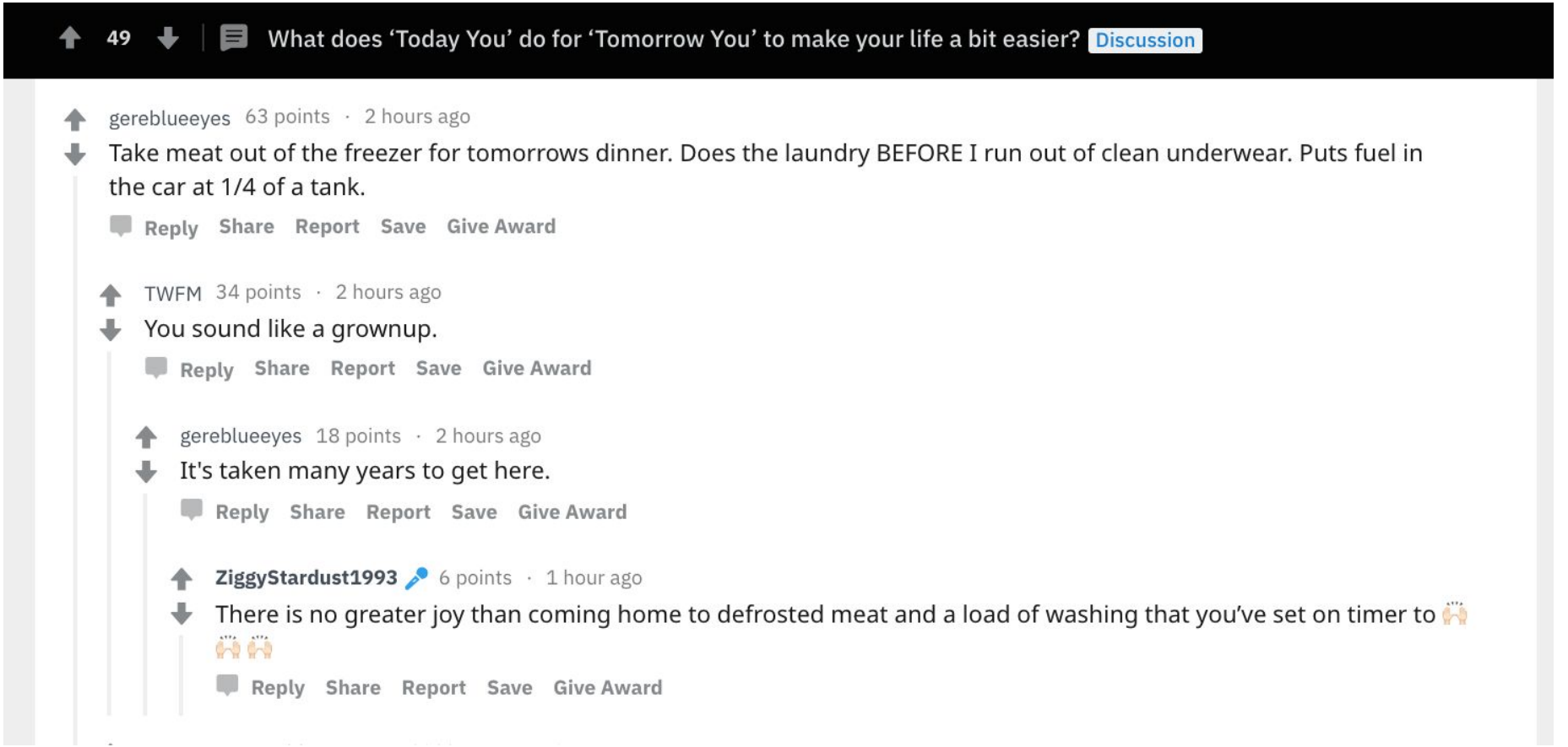
It’s going pretty well, although it’s still pretty cold for here. We’ve had frost here a couple of nights this week and it even snowed at higher elevations

We’re feeling a little cooped up, since it’s been pouring nonstop, but it’s supposed to be nicer tomorrow anyway

I hope the weather improves soon. It’s been very nice here since we came back from
Ohio. It’s breezy today and only in the 60’s but still nice.

I would take 60’s. It’s not the worst thing in the world for it to be cold here.

It’s been pretty much sunny and mid 70’s to low 80’s this week. Which is perfect if its not humid. Only about 2 1/2 months till that starts again.



Source: AskReddit

NLP

(Natural Language Processing)

A Toy Corpus

Document 1: My brother likes to play guitar.

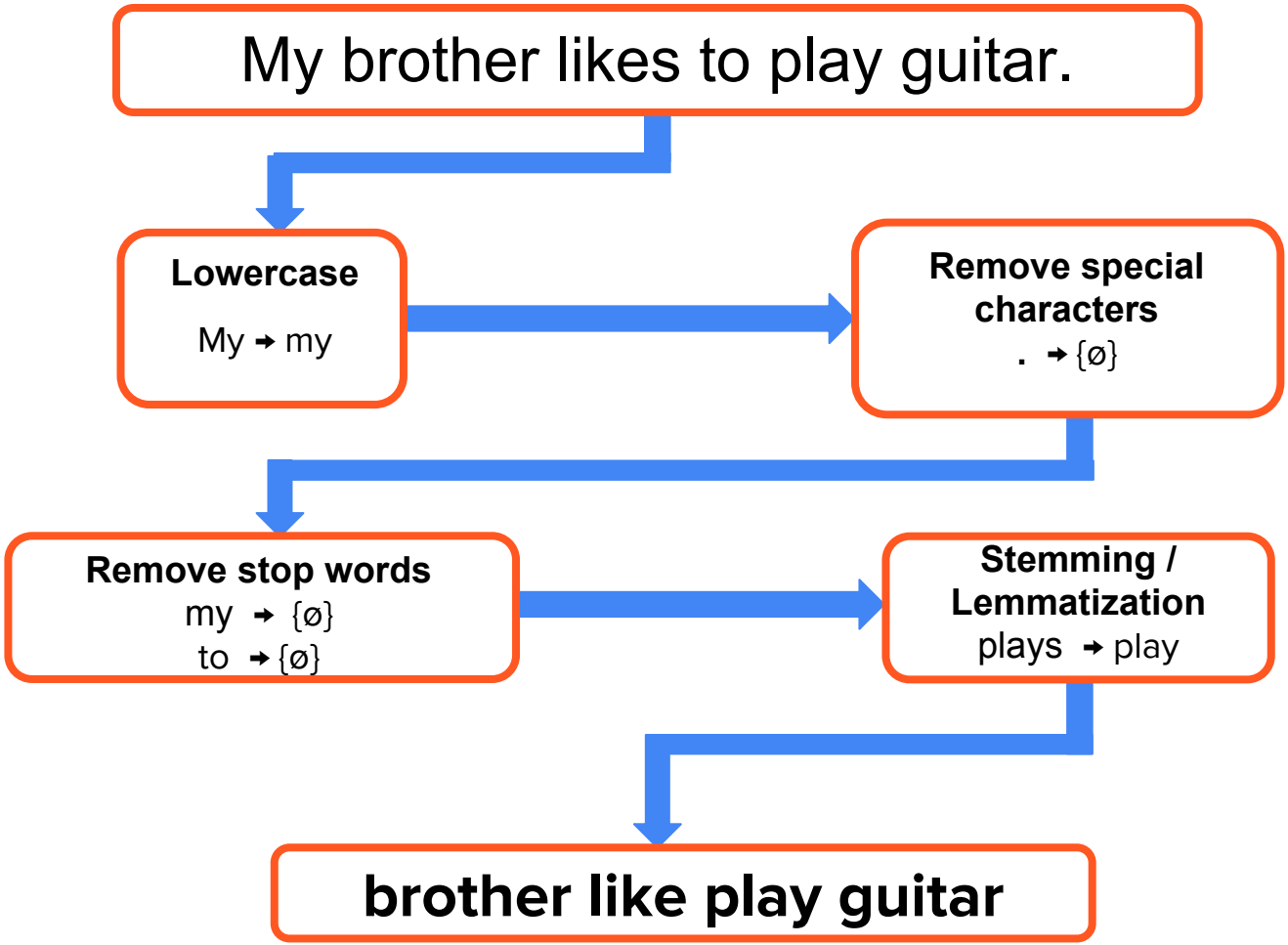
Document 2: She also likes guitar music.

Document 3: The music was loud!

Common NLP Preprocessing Steps

Preprocessing = normalization

1. Lowercasing
2. Removing special characters
3. Remove words with low information content (“stop words”)
4. Stemming / lemmatization (removing prefixes and suffixes)


My brother likes to play guitar.
Lowercase
My → my
Remove special characters
. → {ø}
Remove stop words
my → {ø}
to → {ø}
Stemming / Lemmatization
plays → play
brother like play guitar

Bag of Words

	also	brother	guitar	like	loud	music	play	she	was
Document 1	0	1	1	1	0	0	1	0	0
Document 2	1	0	1	1	0	1	0	1	0
Document 3	0	0	0	0	1	1	0	0	1

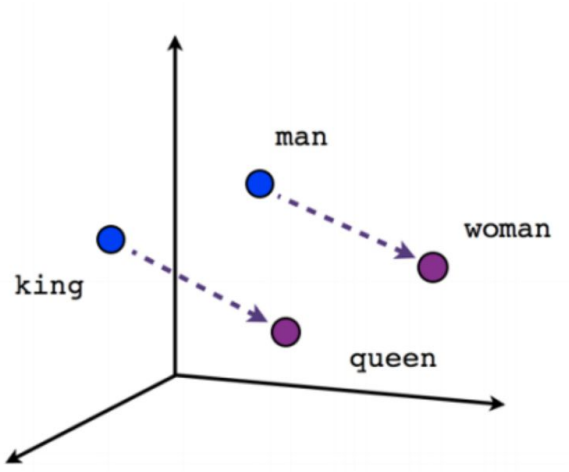
Bag of Words (TF-IDF)

	also	brother	guitar	likes	loud	music	play	she	was
Document 1	0	0.44	0.33	0.33	0	0	0.44	0	0
Document 2	0.52	0	0.39	0.39	0	.39	0	0.52	0
Document 3	0	0	0	0	0.53	0.53	0	0	0.53

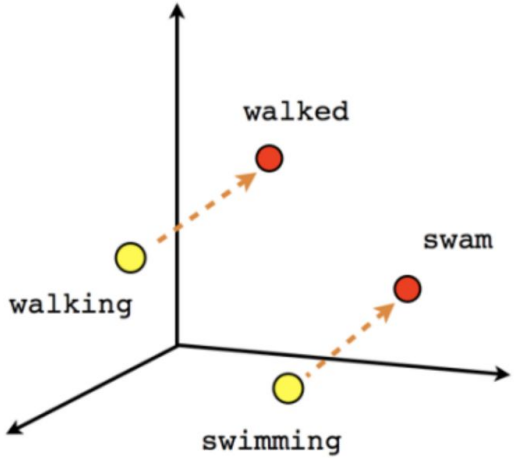
Word Embeddings

array([2.57366691e-02, 2.34292820e-01, -2.25516841e-01, 9.60306749e-02, 1.73759654e-01, -1.41218662e-01, 9.81499925e-02, -2.59603351e-01, 3.64628360e-02, 2.17744994e+00,
1.03816390e-03, 2.27650166e-01, 2.06617918e-02, 1.34981677e-01, -2.31201991e-01, 8.12402815e-02, -1.02147840e-01, 6.37144148e-01, -1.40572324e-01, 2.42056668e-01,
1.39532506e-01, -1.45041004e-01, -8.17376673e-02, 1.05790339e-01, 3.92300040e-02, 2.11964488e-01,   3.14404294e-02, 1.18907504e-01, 1.23028837e-01, -1.09470012e-02,
1.99595001e-02, 2.58022159e-01, -8.81183371e-02, 4.22833376e-02, 1.32231489e-01, -3.95340025e-02, 1.96540002e-02, -9.19334125e-03, 2.22775340e-01, -3.20076674e-01,
-3.53998393e-02, 1.06605671e-01, 9.30931568e-02, -8.17540064e-02, -3.45943347e-02, 1.39929727e-01, 4.31735031e-02, 5.78506626e-02,-8.81391466e-02, 6.88674971e-02,
-2.19402835e-01, -7.92200193e-02, 7.30716661e-02, 5.96557818e-02, 5.59951626e-02, 3.74743342e-02, -7.04365000e-02, -3.81000005e-02, 3.94041613e-02, -2.10180685e-01,
-8.99049938e-02, 1.00973777e-01, -9.76605043e-02, 6.28866032e-02, 3.47779989e-02, -2.11527035e-01, 1.52280673e-01, 6.20059967e-02, -7.02575147e-02, -7.92802647e-02,
-3.48336734e-02, 1.33269176e-01, 6.39401674e-02, -9.75561664e-02, -6.42970055e-02, 4.36808355e-02, 8.62745047e-02, 3.32753249e-02, -7.49828294e-02, -4.28231657e-02,
6.01169877e-02, -1.12035172e-02, -1.36125162e-01, 8.18746686e-02, 1.08909659e-01, -2.21970007e-01, 1.72260687e-01, -2.06738651e-01, 3.16586494e-01, -2.35769972e-02,
-3.23651671e-01, -1.10867716e-01, -1.34719506e-01, 2.45065495e-01, 1.46379828e-01, -2.05476657e-01, 4.03196186e-01, 9.90266576e-02, -2.39653170e-01, 4.20101695e-02,
-8.25038329e-02, -9.89145041e-02, 9.84141752e-02, 1.15932666e-01, -1.53431162e-01, -1.52546510e-01, 2.47379676e-01, 1.19438671e-01, 3.37863378e-02, -7.88366720e-02,
2.81721562e-01, -3.26300323e-01, -9.57049951e-02, 1.84794977e-01, 1.41419964e-02, -1.69688351e-02, 1.62748378e-02, -1.93686679e-01, 3.72637324e-02, 1.42149672e-01,
9.82453451e-02, -4.33968902e-02, 4.67049964e-02, -3.24265026e-02, 1.69481680e-01, 2.05937400e-01, -2.98313320e-01, 1.33571491e-01, -4.91788387e-02, 1.42029980e-02,
3.43808234e-02, 8.02036673e-02, -2.64681667e-01, 1.81744769e-01, 1.43906668e-01, 1.15626007e-01, -2.69594938e-02, 1.62795000e-02, 5.59616685e-02, -4.63583320e-02,
-2.37236667e+00, -1.05177395e-01, -1.10730499e-01, -1.79794982e-01,-2.09184170e-01, 1.57856658e-01, 7.06513747e-05, -2.48830155e-01, -5.59633225e-02, -3.01599354e-01,
1.55764177e-01, 1.25454649e-01, 4.32921685e-02, -5.51541634e-02, 1.09350085e-02, -2.66731501e-01, -1.14671635e-02, -6.81034103e-02, -8.11251178e-02, -2.98056692e-01,
6.87590465e-02, 1.06871665e-01, 1.43281832e-01, -1.74336001e-01, 6.69990182e-02, -2.70821333e-01, 8.78356621e-02, -1.74323499e-01, 1.29566833e-01, 1.06875993e-01,
-2.06408814e-01, 3.22793163e-02, 3.69264990e-01, -9.60004807e-04, 8.93274918e-02, -7.02276751e-02, -1.74998328e-01, 5.31848259e-02, 8.81608352e-02, -4.89149988e-03,
-1.14898168e-01, -1.73477292e-01, 5.48801683e-02, 1.42810019e-02, 1.49053320e-01, 2.65511721e-02, -2.51292169e-01, -1.38427420e-02, -3.92835028e-02, -2.01444998e-01,
-1.50919661e-01, 6.23006672e-02, -1.16433166e-01, -6.69260100e-02, -8.52531567e-02, 4.14474942e-02, -4.26016785e-02, -1.15648322e-01, -7.47454688e-02, 1.54296651e-01,
-1.92898333e-01, -7.39091709e-02, -1.35480508e-01, -3.00855041e-02, 1.76558182e-01, -2.56533455e-03, -1.29030449e-02, -2.53484517e-01, 1.10509999e-01, -1.68382838e-01,
-3.00106823e-01, -2.09489968e-02, 4.62292023e-02, -1.86726674e-01, -1.25473008e-01, 2.42444992e-01, 1.96976840e-01, -1.72747001e-01, 2.61533353e-02, 1.58475175e-01,
7.04178289e-02, -1.56576663e-01, -2.94274330e-01, -7.69834220e-03, 4.62271720e-02, 7.47152343e-02, 8.41124952e-02, 3.60805005e-01, 2.20953509e-01, 1.41697675e-01,
-2.53109664e-01, -8.59503374e-02, 2.92333420e-02, 9.61689949e-02, -1.23739004e-01, -1.70312658e-01, 4.37483191e-03, 1.89571992e-01, 3.58111672e-02, 1.64738640e-01,
2.63903350e-01, 1.25095502e-01, -3.69083397e-02, 2.00202823e-01,1.08082175e-01, -7.63366595e-02, -4.80566584e-02, -1.12933338e-01, -2.19192818e-01, 2.46603325e-01,
9.07684937e-02, -7.56993368e-02, -9.53274965e-02, 2.90466815e-01, 7.15833306e-02, -4.01766710e-02,-1.56948999e-01, 2.82724965e-02, -2.09465817e-01,
-2.37945020e-02,-7.12763369e-02, 9.25166607e-02, -9.51055065e-02, -1.14205003e-01,-7.53961727e-02, -1.15151502e-01, -2.67045647e-01, 4.39883322e-02,9.89266634e-02,
1.10445321e-01, 8.85200128e-02, -5.21190017e-02,1.27731748e-02, -4.02626656e-02, -1.05934501e-01, 1.66156683e-02, 8.97403359e-02, 1.64340004e-01, 2.74958368e-02,
-2.11101651e-01, -1.25387624e-01, 2.25980163e-01, -1.86898515e-01, -2.04148337e-01,1.00976668e-01, -7.27016628e-02, 2.80245334e-01, -2.76825994e-01,-6.80672005e-04,
-1.43736318e-01, -1.42790824e-01, -4.05921489e-02, 6.83716685e-02, 2.70542711e-01, 3.65885019e-01, 1.74611852e-01,1.45135105e-01, 6.53208271e-02, 7.04205036e-02,
1.65292546e-01]

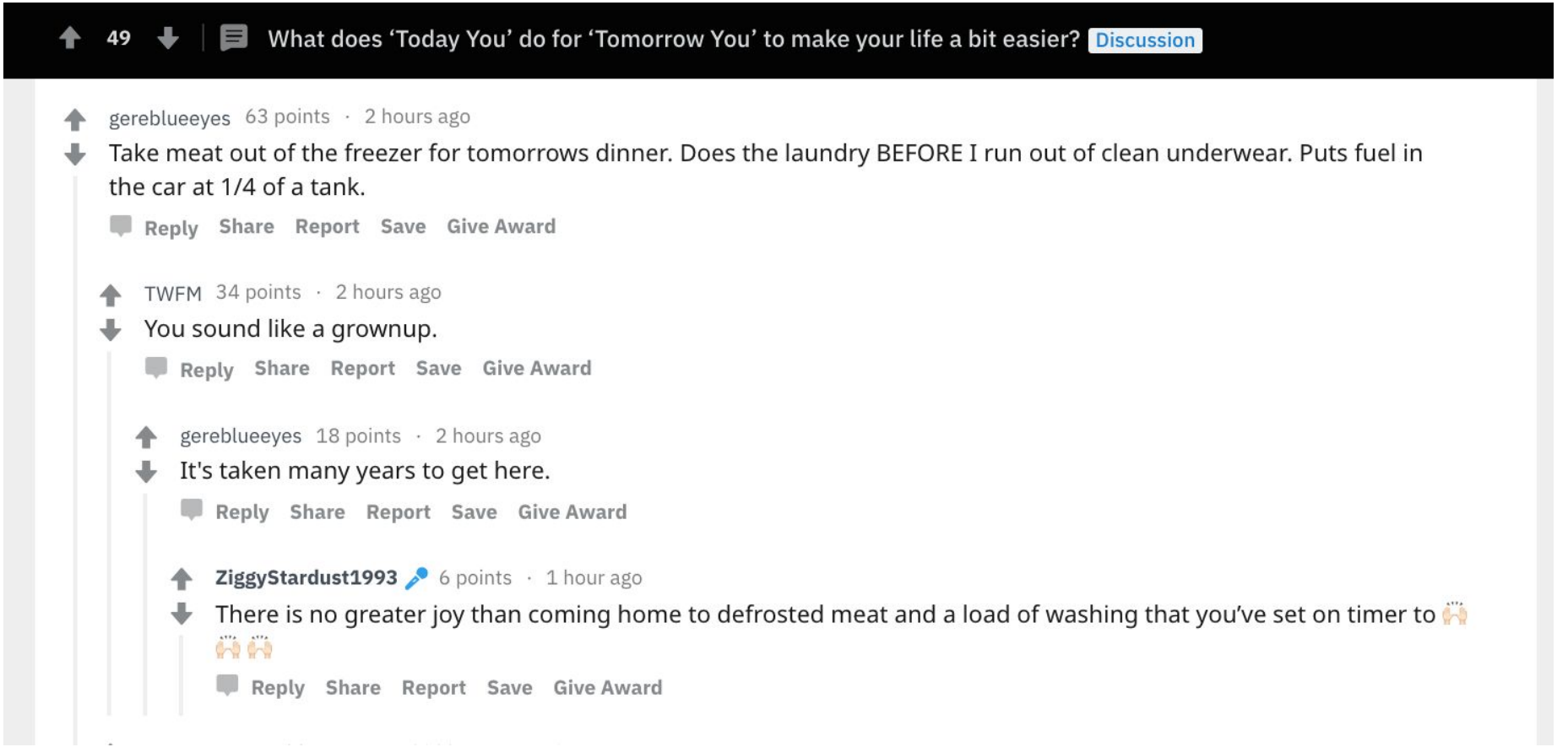
Word Embeddings



Male-Female



Verb tense



Source: AskReddit

I don’t know!

I don’t know…

I don’t know?

I don’t know

Punctuation Matters!

That’s obviously what you should do next

That’s OBVIOUSLY what you should do next

That’s obviously what YOU should do next

Capitalization MATTERS!

Thanks 🙌

Thanks 🙄

Thanks 😍

Thanks 🖕

Emojis DEFINITELY matter! 🙏

Paralanguage

“a component of meta-communication that may modify meaning, give nuanced meaning, or convey emotion”

Source: Wikipedia

Forms of Digital Paralanguage

Capitalization

Punctuation

Emoji / Emoticons

Formatting / Markdown

Sentiment

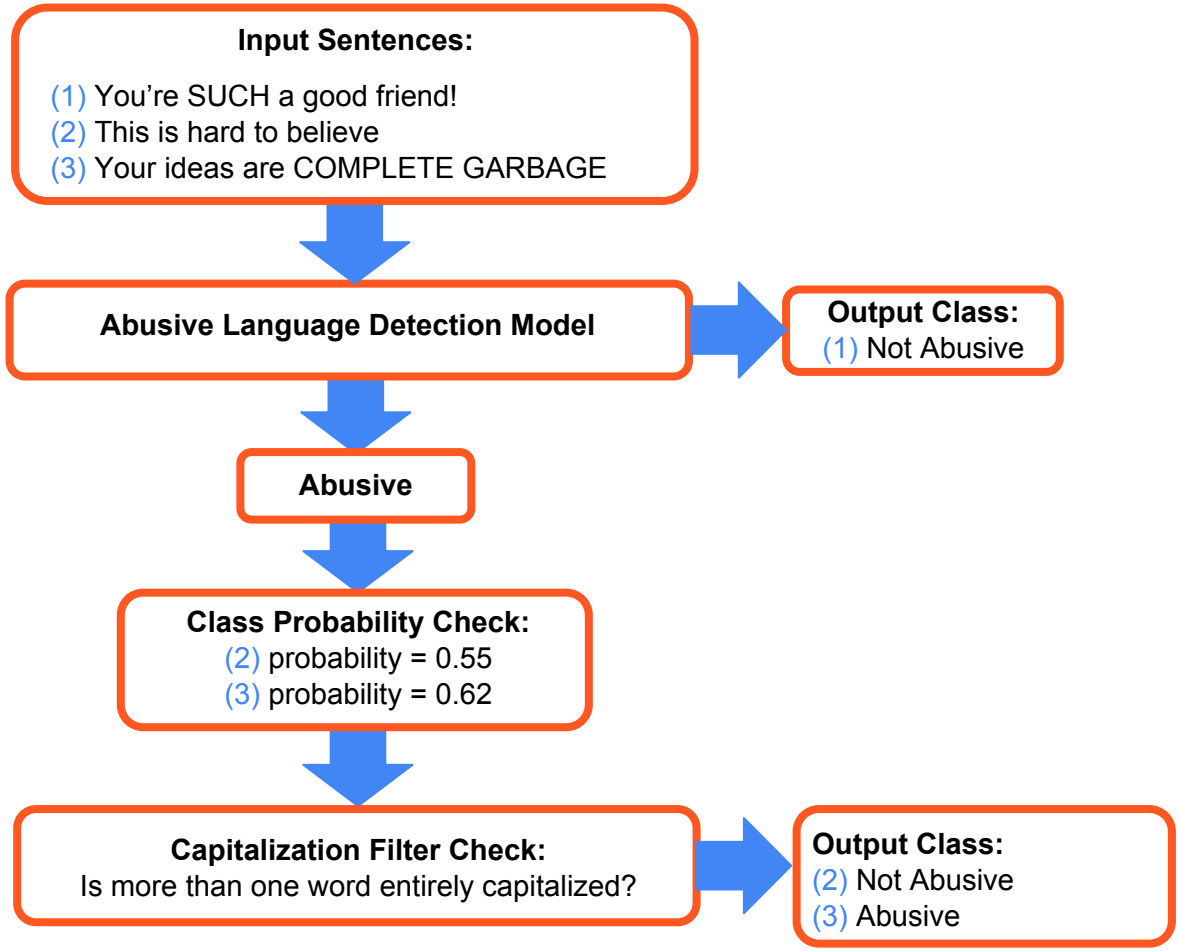
Pronoun Usage

Thread Structure

¯_(ツ)_/¯

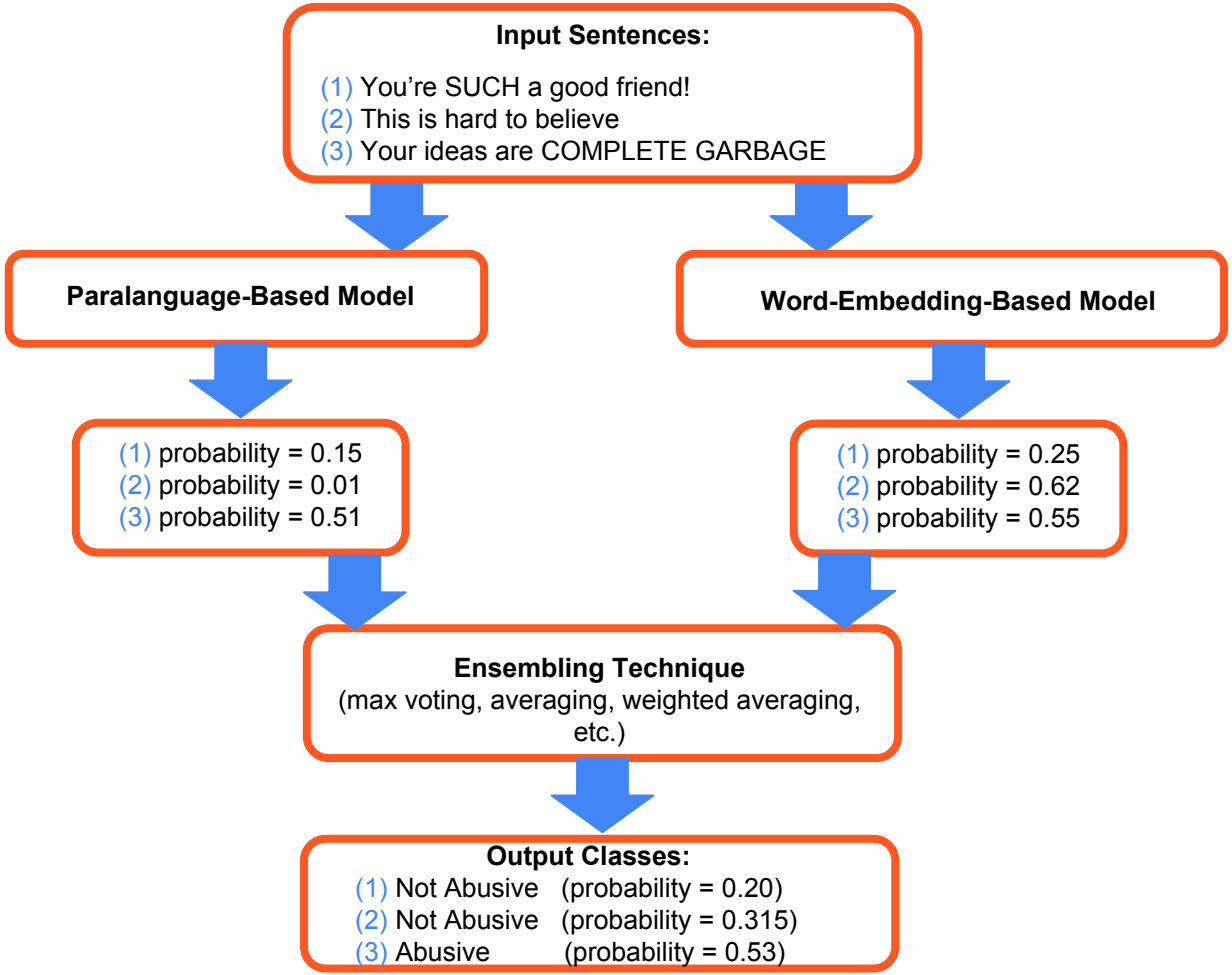
Filtering

(LOW EFFORT)


Input Sentences:
(1) You're SUCH a good friend!
(2) This is hard to believe
(3) Your ideas are COMPLETE GARBAGE
Abusive Language Detection Model
Output Class:
(1) Not Abusive
Abusive
Class Probability Check:
(2) probability = 0.55
(3) probability = 0.62
Capitalization Filter Check:
Is more than one word entirely capitalized?
Output Class:
(2) Not Abusive
(3) Abusive

Ensembling

(MEDIUM EFFORT)


Input Sentences:
(1) You're SUCH a good friend!
(2) This is hard to believe
(3) Your ideas are COMPLETE GARBAGE
Paralanguage-Based Model
Word-Embedding-Based Model
(1) probability = 0.15
(2) probability = 0.01
(3) probability = 0.51
(1) probability = 0.25
(2) probability = 0.62
(3) probability = 0.55
Ensembling Technique
(max voting, averaging, weighted averaging, etc.)
Output Classes:
(1) Not Abusive (probability = 0.20)
(2) Not Abusive (probability = 0.315)
(3) Abusive (probability = 0.53)

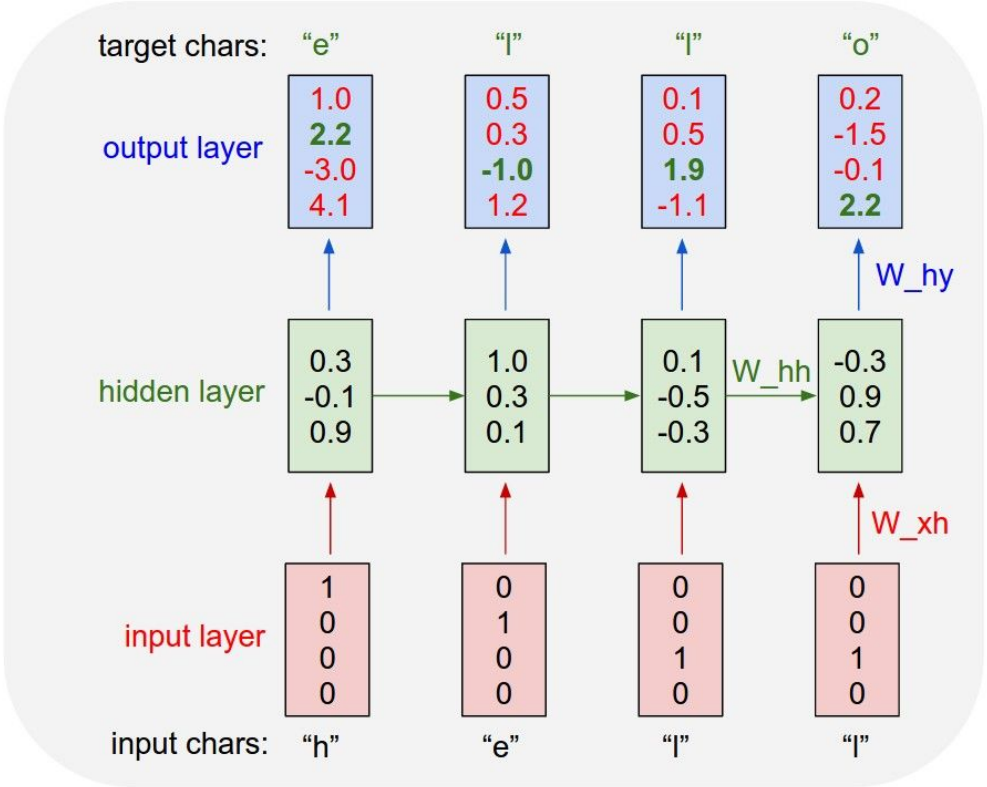
Character Level Representations

(HIGH EFFORT)

['idea', 'complete', 'garbage']

VS

[ide',	'MPL',
'dea',	'PLE',
'eas',	'LET',
'as<space>',	'ETE',
's<space>a',	'TE<space>',
'<space>ar',	'E<space>Gv
'are',	'<space>GA',
're<space>',	'GAR',
'e<space>C',	'ARB',
'<space>CO',	'RBA',
'COM',	'BAG',
'OMP',	'AGE']



Source: [The Unreasonable Effectiveness of Recurrent Neural Networks]

NLP on conversational data is **different**

Standard preprocessing techniques remove **digital paralanguage**

Paralanguage can leveraged using **filtering**, **ensembling**, or **character-level representations**

Thanks!

Katie Bauer
Senior Data Scientist, Reddit
@imightbemary