Tweets Emotion Detection Using Topic Modelling Features

Norah Alkhalifah
Meshal Alamr

Introduction

EDA

Text Preprocessing

Topic Modeling

Classification

Conclusion

Introduction

- **Classify emotions from tweets:**
  - Help businesses analyze their customer feedback through Twitter.
  - Build music and movie recommender based on mood.

Objective

Can Topic Model Distributions Classify Tweet Emotions?

Introduction

Train LDA Model → Get Topic Distributions Per Tweet → Input to a Supervised Classifier

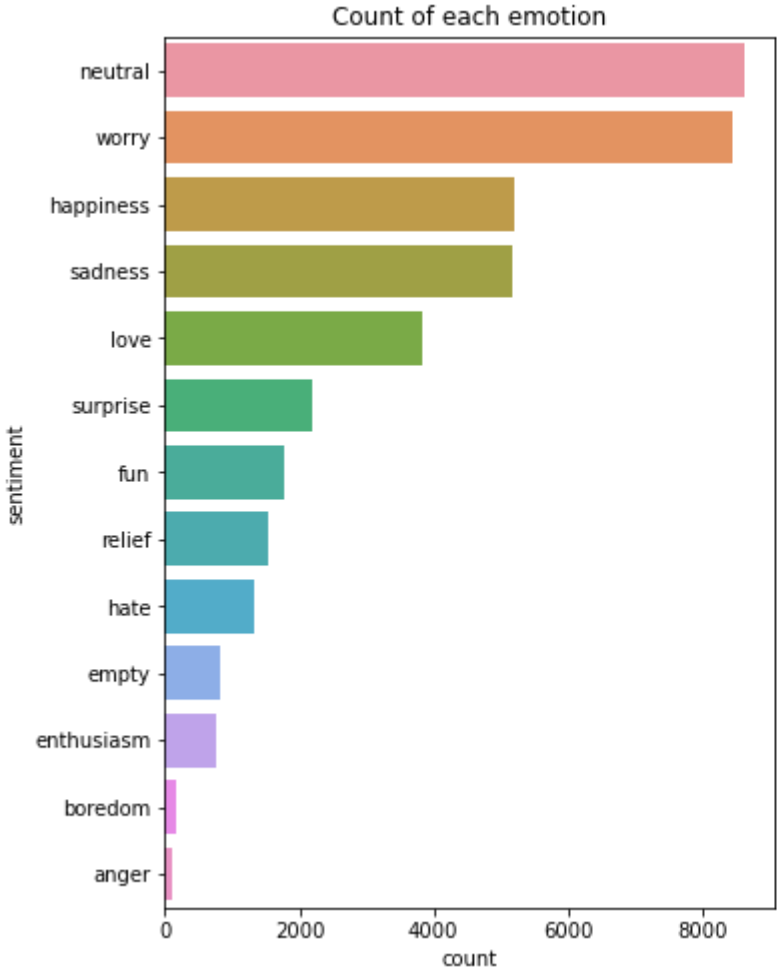
Dataset

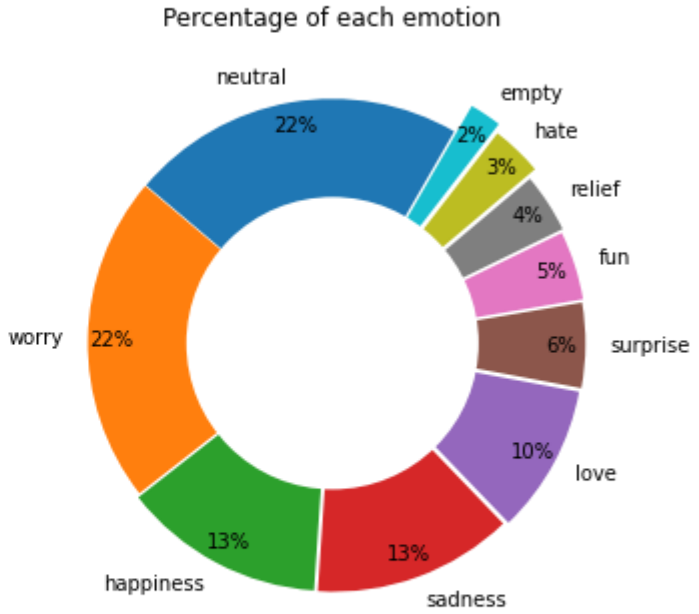
	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

40,000 Rows 3 Columns

EDA

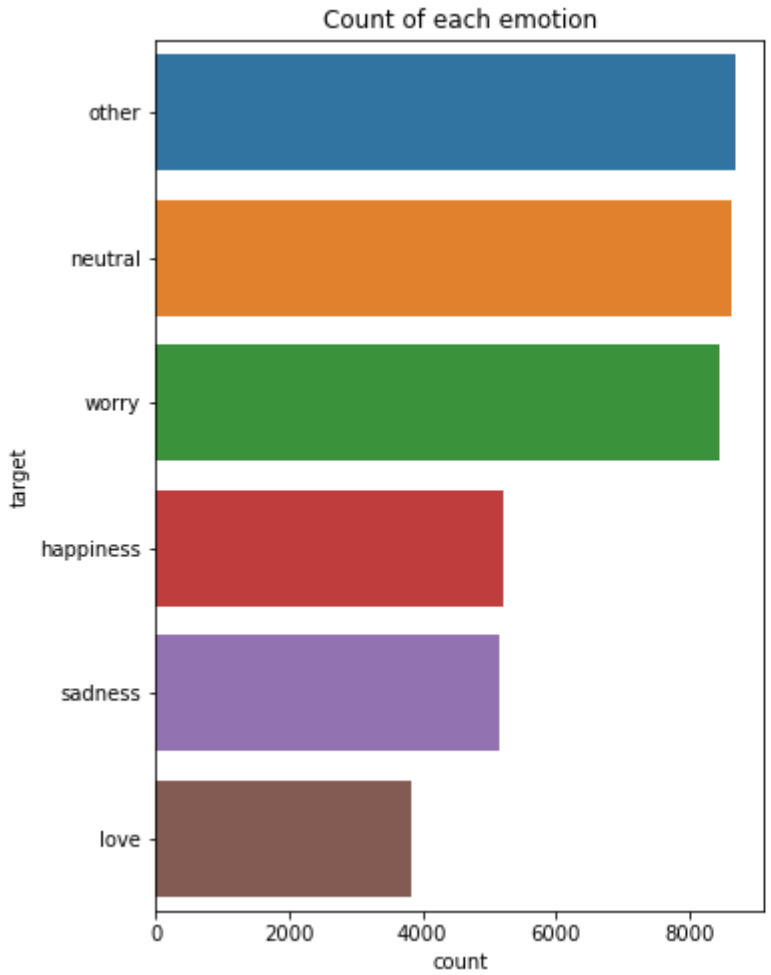
Distribution of Emotions

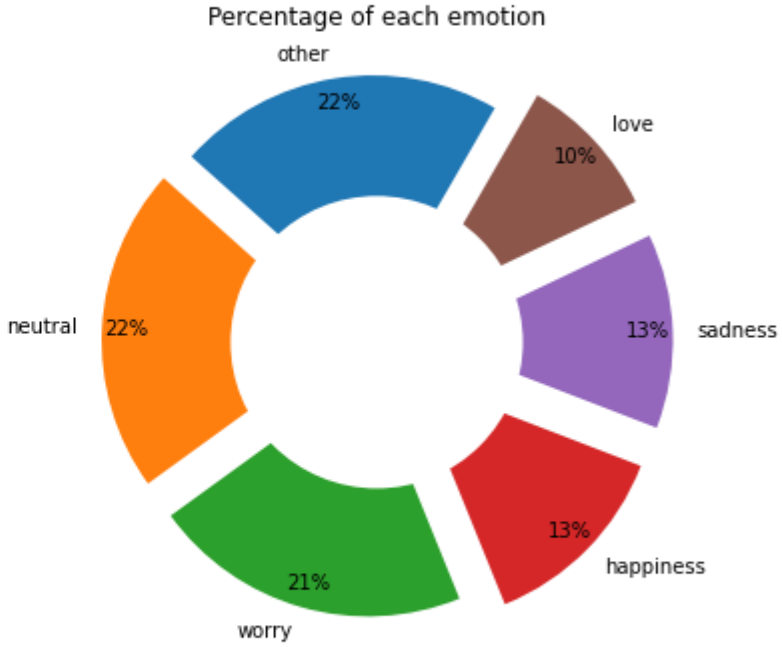




EDA

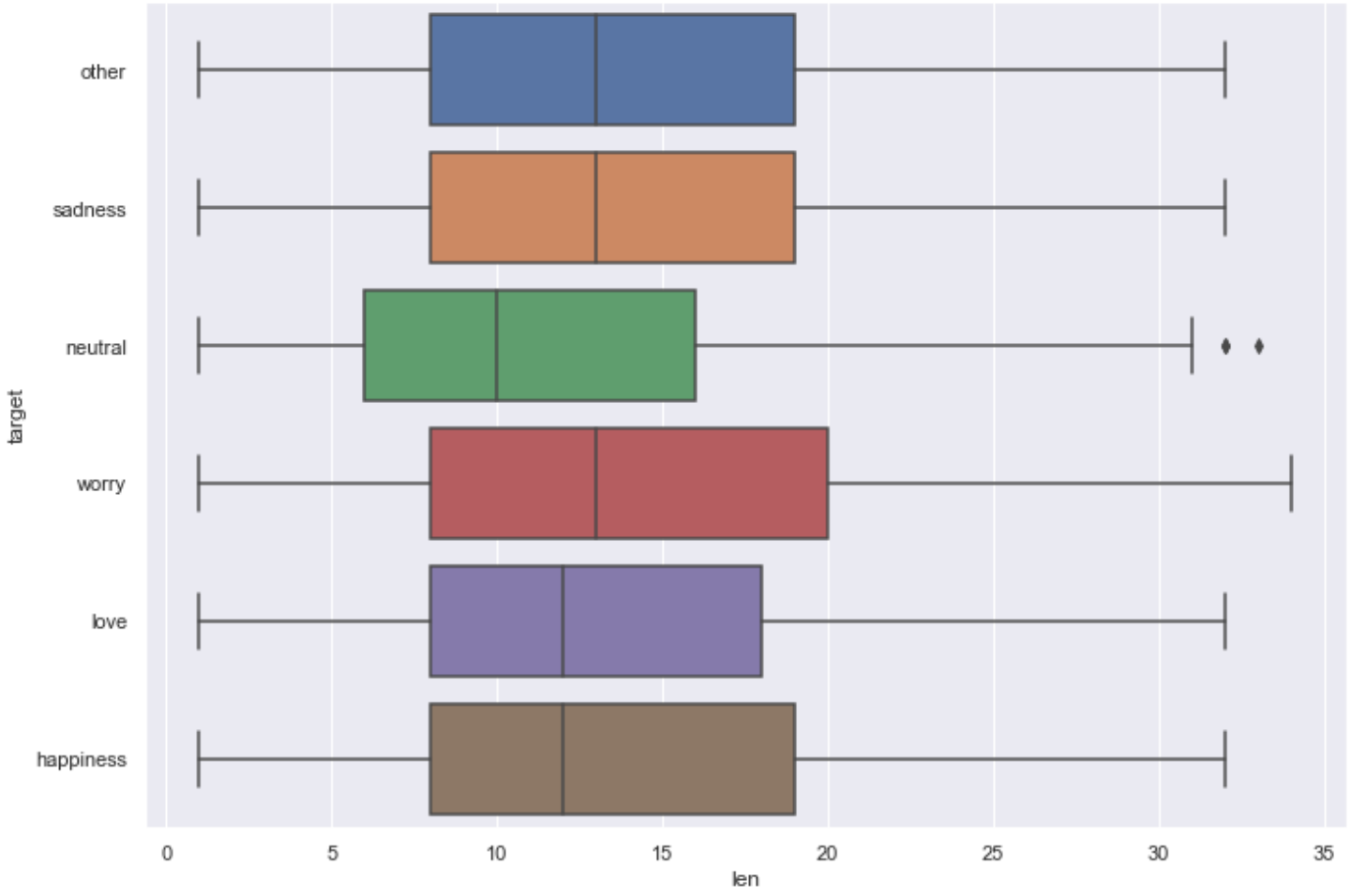
Distribution of Emotions





EDA

Number of Words Per Emotion



Word Clouds

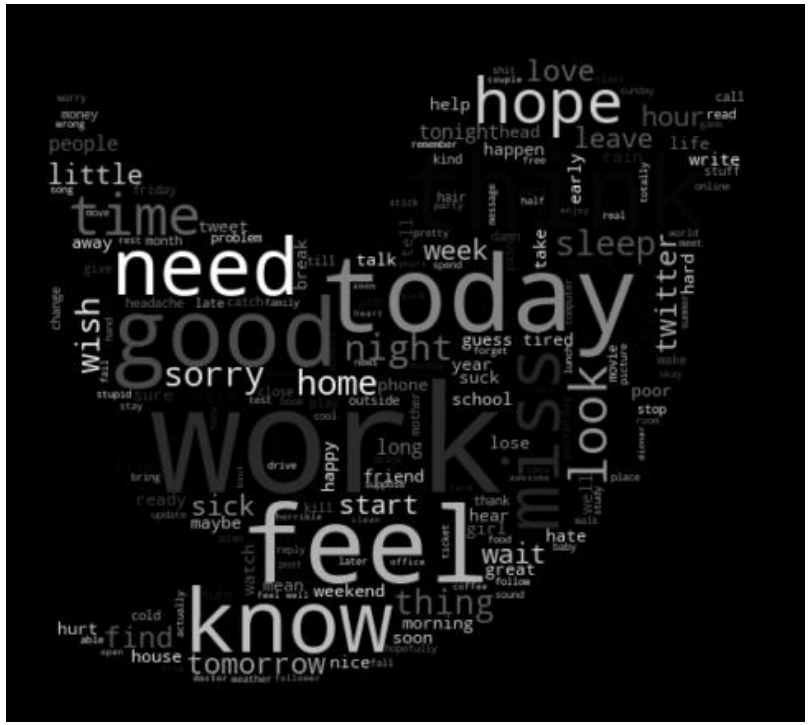
Happiness

Love

Word Clouds

Sadness

Worry



Text Preprocessing

- Remove URLs
- Convert to lowercase
- Remove Twitter Handles (@username)
- Remove Special Characters / Numbers
- Lemmatization
- Tokenization
- Remove Empty Tweets
- Spellcheck

Experiments

- Modelling:
  - LSA
  - LDA
- Embedding:
  - TF-IDF
  - CountVectorizer

LSA / TF-IDF

Topic 0
day, mother, happy, good, mom, work, love, today, thank, morning

Topic 1
thank, good, work, morning, love, know, time, miss, today, night

Topic 2
thank, mother, happy, follow, mom, love, appreciate, followfriday, god, follower

Topic 3
work, thank, mother, happy, tomorrow, hour, ready, wish, link, friday

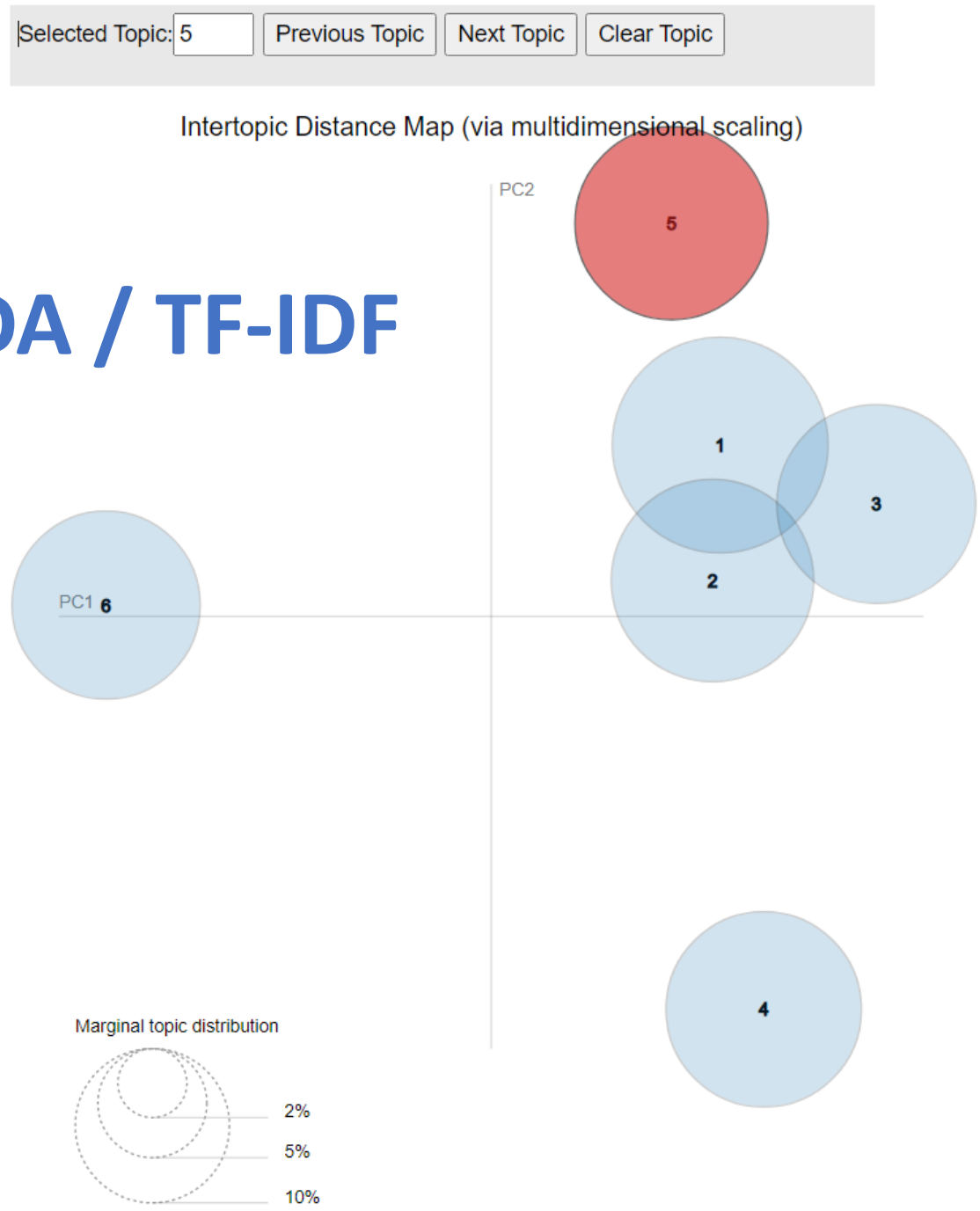
Topic 4
miss, love, know, lol, think, time, home, sad, come, twitter

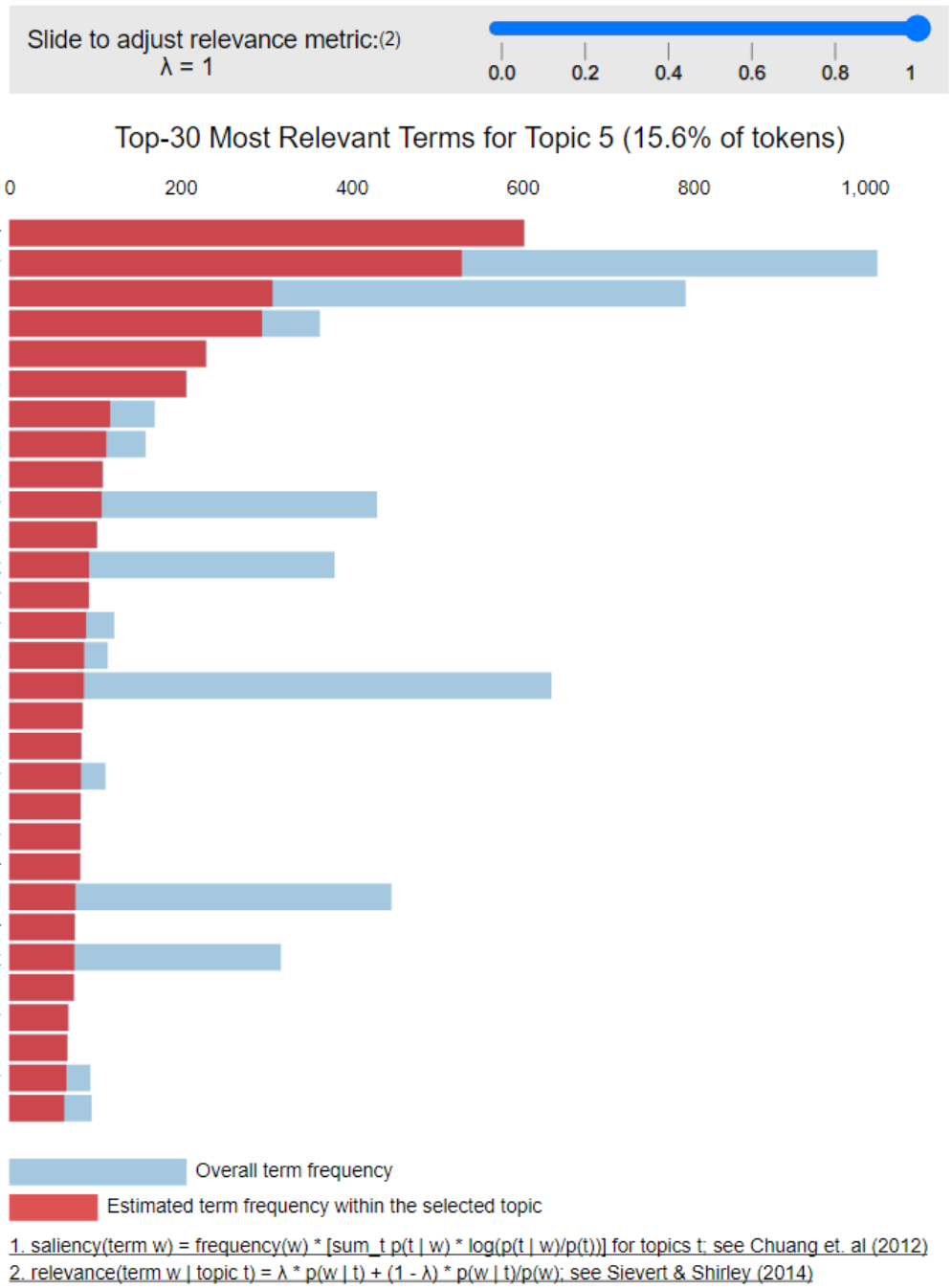
Topic 5
miss, thank, day, today, morning, good, sad, tonight, bad, school

Topic Modeling

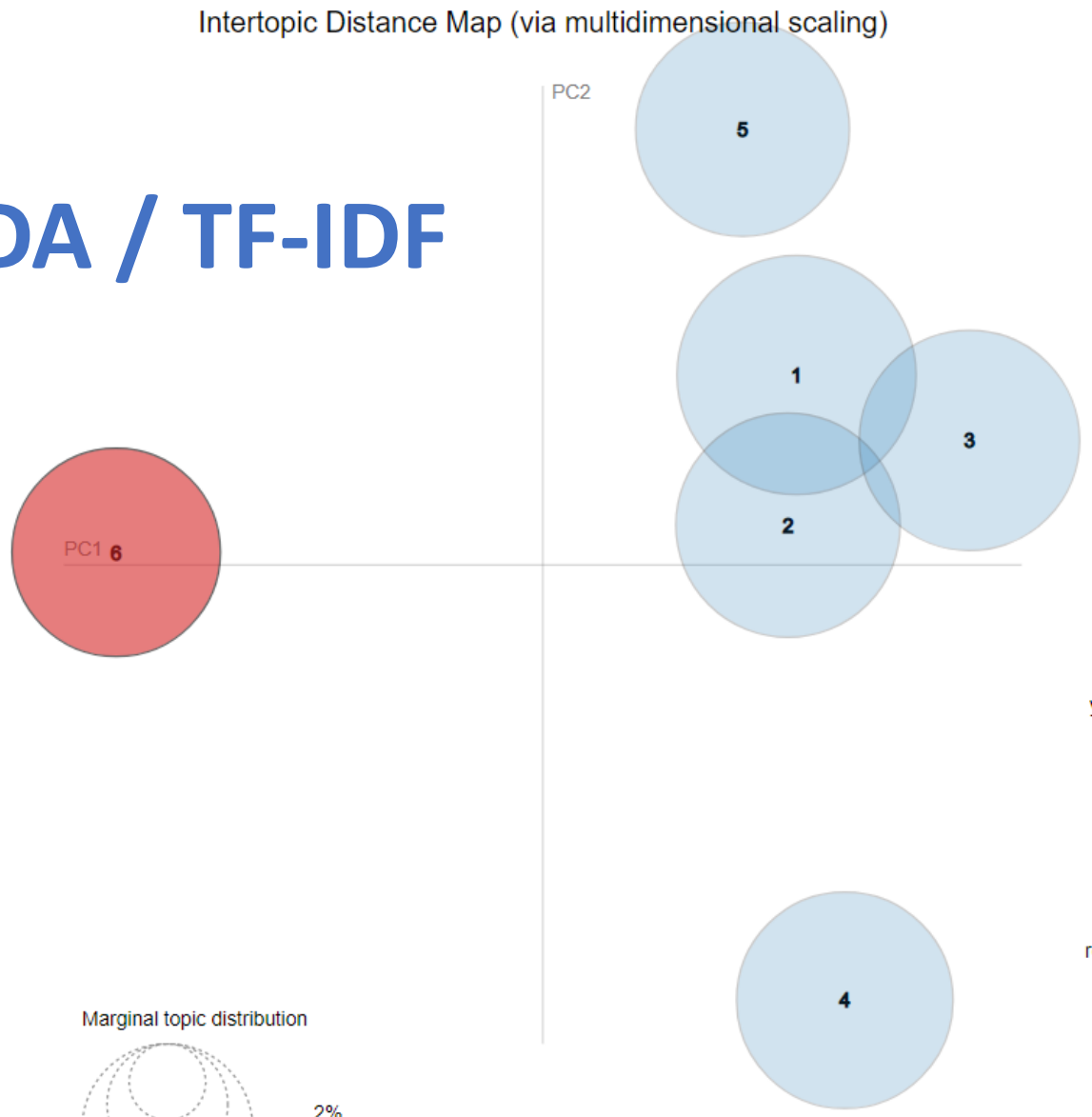
Train LDA Model → Get Topic Distributions Per Tweet → Input to a Supervised Classifier

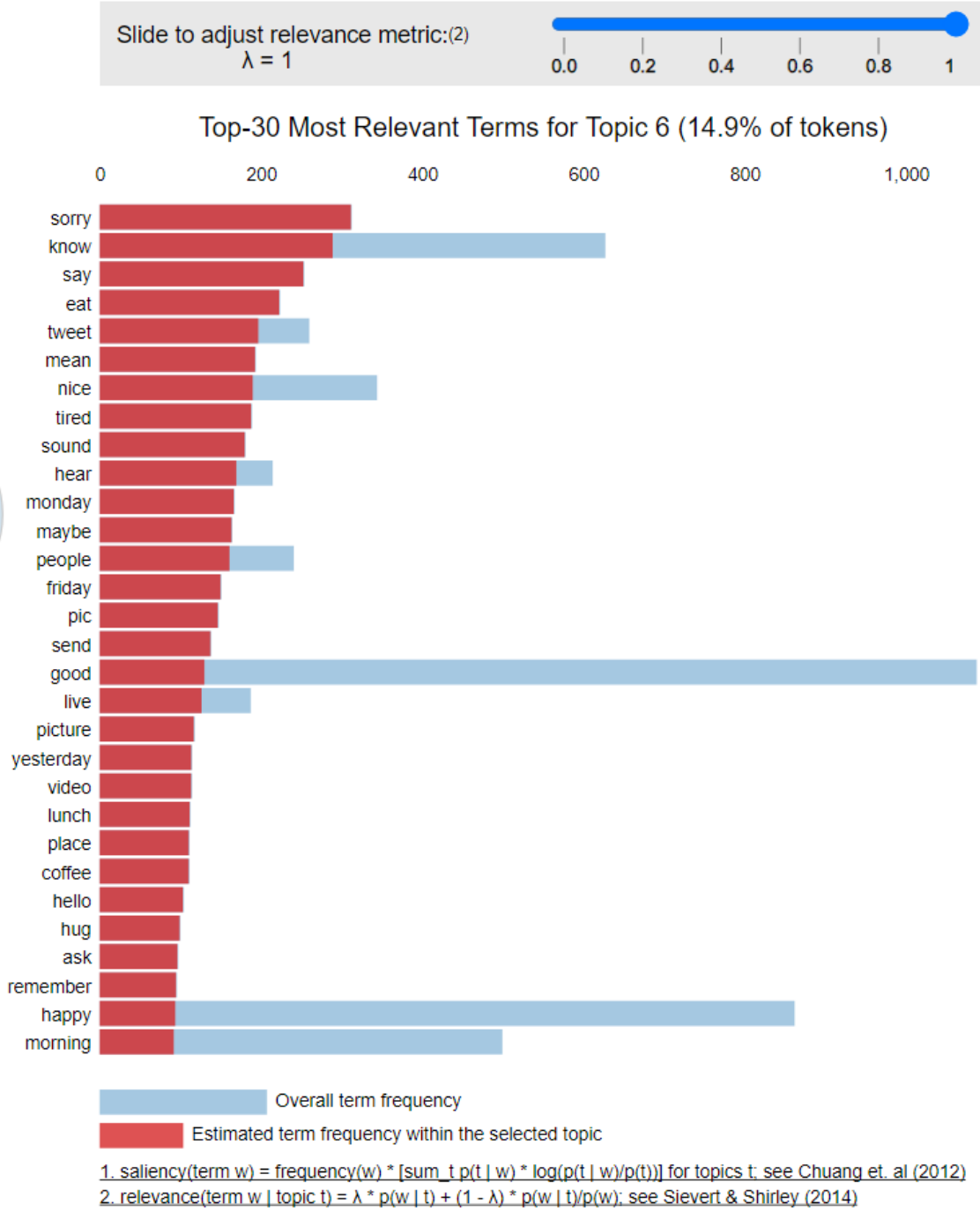
LDA / TF-IDF


Selected Topic: 5
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
1
3
2
6
4
Marginal topic distribution
2%
5%
10%


Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Top-30 Most Relevant Terms for Topic 5 (15.6% of tokens)
0
200
400
600
800
1,000
mother
day
good
morning
yeah
awesome
long
tell
wake
today
kind
night
funny
pretty
house
love
mum
beautiful
early
bored
lovely
weather
lol
hair
great
hahaha
probably
idea
buy
omg
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA / TF-IDF


Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
1
3
2
6
4
Marginal topic distribution
2%
5%
10%


Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Top-30 Most Relevant Terms for Topic 6 (14.9% of tokens)
0
200
400
600
800
1,000
sorry
know
say
eat
tweet
mean
nice
tired
sound
hear
monday
maybe
people
friday
pic
send
good
live
picture
yesterday
video
lunch
place
coffee
hello
hug
ask
remember
happy
morning
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic Modeling

Train LDA Model → Get Topic Distributions Per Tweet → Input to a Supervised Classifier

Classification

Train LDA Model → Get Topic Distributions Per Tweet → Input to a Supervised Classifier

Classification

	Topic1	Topic2	Topic3	Topic3	Topic4	Topic5	target
0	0.020848	0.449249	0.467249	0.020957	0.020864	0.020833	3
1	0.033333	0.233211	0.033467	0.633322	0.033333	0.033333	4
2	0.833333	0.033333	0.033333	0.033333	0.033333	0.033333	4
3	0.033333	0.233090	0.033578	0.633332	0.033333	0.033333	3
4	0.041667	0.041667	0.791666	0.041667	0.041667	0.041667	2

Classification

- XGBoost
- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Random Forest
- Support Vector Machines

Final Model

Model / Metric	F1	Accuracy
Random Forest	23%	24%

Conclusion

- Relying on topic distributions solely is not effective in detecting emotions:
  - Topics are not distinct enough for emotion detection.

Conclusion

- Future Work:
  - Feed more features to the classifiers.
  - Classify directly without topic modelling.


Thank You!