

# BUILDING END TO END TWITTER USER GEO-INFERENCING APP

**WESLEY CHEW**

**DSI 12**

# PROBLEM STATEMENT

To predict which region and country a twitter user is from

WHY?

**Geolocation marketing, Security monitoring, Regional Sentimental Analysis ...**

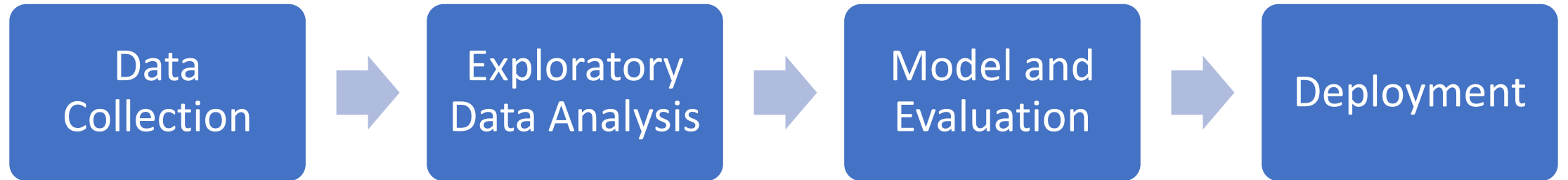
BUT...

15% of users specify their location on profiles

# Approach

- **Predict region of user based on activity patterns**
  - Number of tweets by hour of day
- **Predict country of user based on tweet text using NLP techniques**

# WORKFLOW




# DATA COLLECTION

- Social Bakers
  - Social Media Marketing Platform



## Twitter Profiles Stats in Singapore

		Followings	Followers
1	 Aaron Aziz Official (@AaronDwiAziz)	859	2 078 725
2	 JJ Lin (@JJ_Lin)	58	1 091 270
3	 The Straits Times (@STcom)	128	1 034 809
4	 Kaylani Lei (@KaylaniLive)	555	879 728
5	 Medyo Maldito (@Medyomaldito)	37 233	777 802
6	 Lee Hsien Loong (@leehsienloong)	29	724 367



## Twitter Profiles Stats in Australia

		Followings	Followers
1	 Hugh Jackman (@RealHughJackman)	10	14 146 487
2	 5 Seconds of Summer (@5SOS)	29 825	12 635 812
3	 Luke Hemmings (@Luke5SOS)	20 467	10 466 523
4	 ceo of ceos (@Michael5SOS)	14 841	10 387 883
5	 troye (@troyesivan)	3 474	9 320 363
6	 PRINCE NEPTUNE (@CodySimpson)	73 927	7 254 275

# DATA COLLECTION

	date	time	username	tweet
0	2019-09-25	03:47:30	suarezlegroupe	Good night! #andalusia 🌶️ <a href="#">pic.twitter.com/xPtdm...</a>
1	2018-02-08	22:54:27	suarezlegroupe	En live sur @virginradiofr 🌸❤️👑 <a href="#">https://www.i...</a>
2	2018-01-03	16:37:19	suarezlegroupe	Hahahah
3	2017-11-19	05:30:29	suarezlegroupe	We love Huy ! 🥰🥰💋❤️🌸 <a href="#">https://www.instagram.co...</a>
4	2017-02-12	17:29:14	suarezlegroupe	joyeux anniversaire 🎉🥰

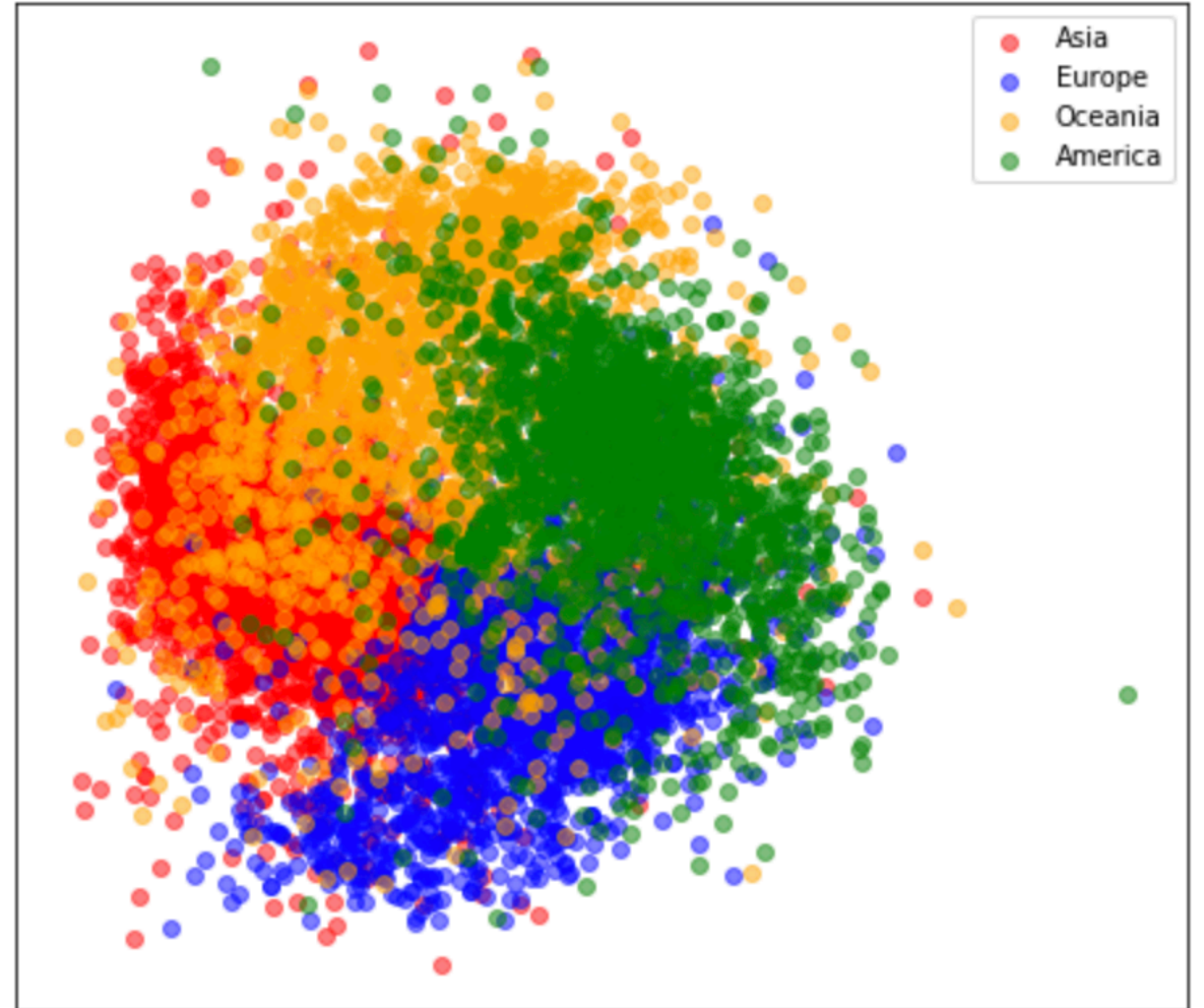


	username	country	region	tweets	0	1	2	3	4	5	...	21	22	23	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
0	electronicalree	belgium	europe	We've just updated our "Secret Weapons & Gems"...	64	39	22	28	14	16	...	50	27	37	44	77	71	94	96	86
1	suarezlegroupe	belgium	europe	Good night! #andalusia 🌶️ <a href="#">pic.twitter.com/xPtdm...</a>	1	3	6	29	3	1	...	3	2	3	18	9	15	16	15	10
2	twezus	belgium	europe	We need to talk about Rudy. <a href="#">https://www.nieuws...</a>	25	39	43	32	34	52	...	27	32	28	74	104	64	80	78	63
3	focusvif	belgium	europe	Marion Cotillard, profession: actrice <a href="#">https://...</a>	36	32	14	15	17	8	...	44	51	34	84	74	83	76	108	48
4	tuiflybelgium	belgium	europe	For long haul flights we advise you to be at t...	22	16	21	19	12	13	...	33	37	24	72	88	63	79	86	60

# EXPLORATORY DATA ANALYSIS

## Principal component analysis

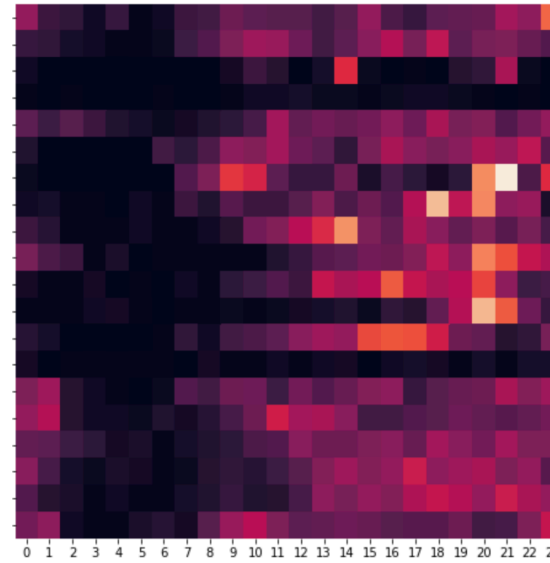
- Reduce dimensionality of dataset
- Transform 24 dimensions to 2 dimensions.



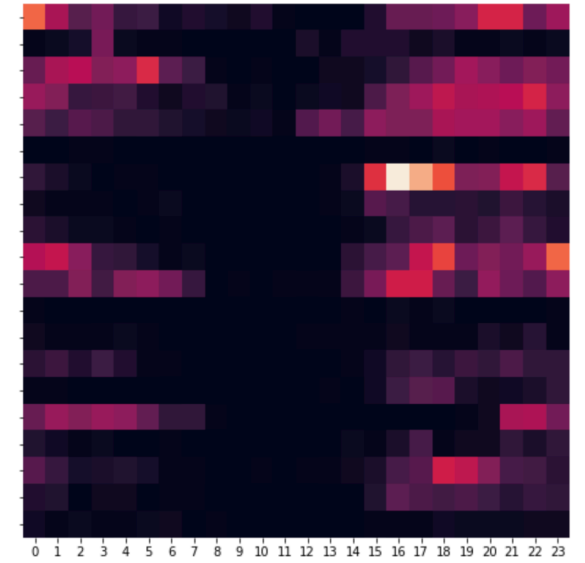
# EXPLORATORY DATA ANALYSIS

## Activity heatmap of users in each region

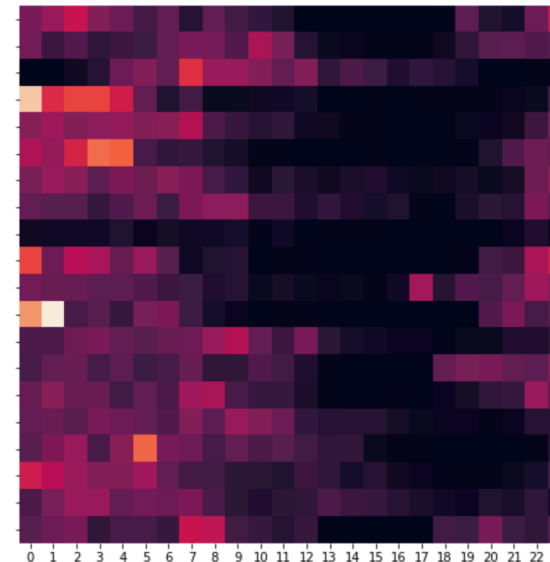
- Each row represents a user
- Columns represent 0 to 23 hours
- Heatmap is with respect to Singapore time zone (GMT+8)



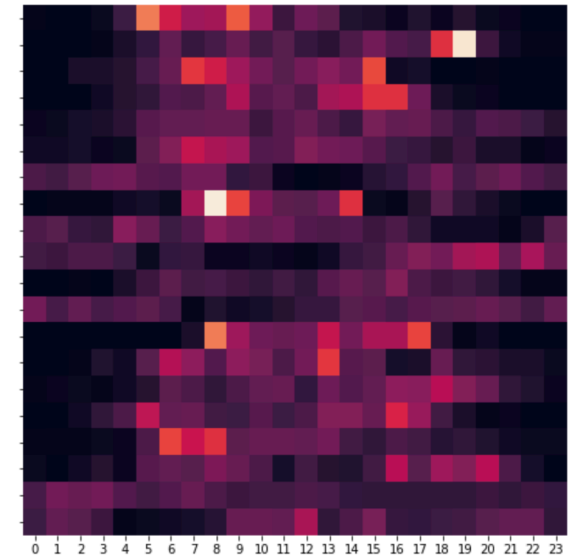
Asia



Europe



America



Oceania



# MODEL (Region)

## Model

- XGBoost and Feed forward Neural Network
- Same accuracy of (85%)

	America	Asia	Europe	Oceania
America	440	9	20	26
Asia	8	644	27	15
Europe	32	8	452	3
Oceania	51	35	27	381

## Possible reasons for the low accuracy

- Overlapping time zone for country in Oceania and America
- Users posting tweets at irregular timings

# MODEL (Country)

## Data Cleaning

- Tokenize, removing stop words, lower case, remove URLs, emojis and punctuation
- Fit into TFIDF vectorizer

## Modeling

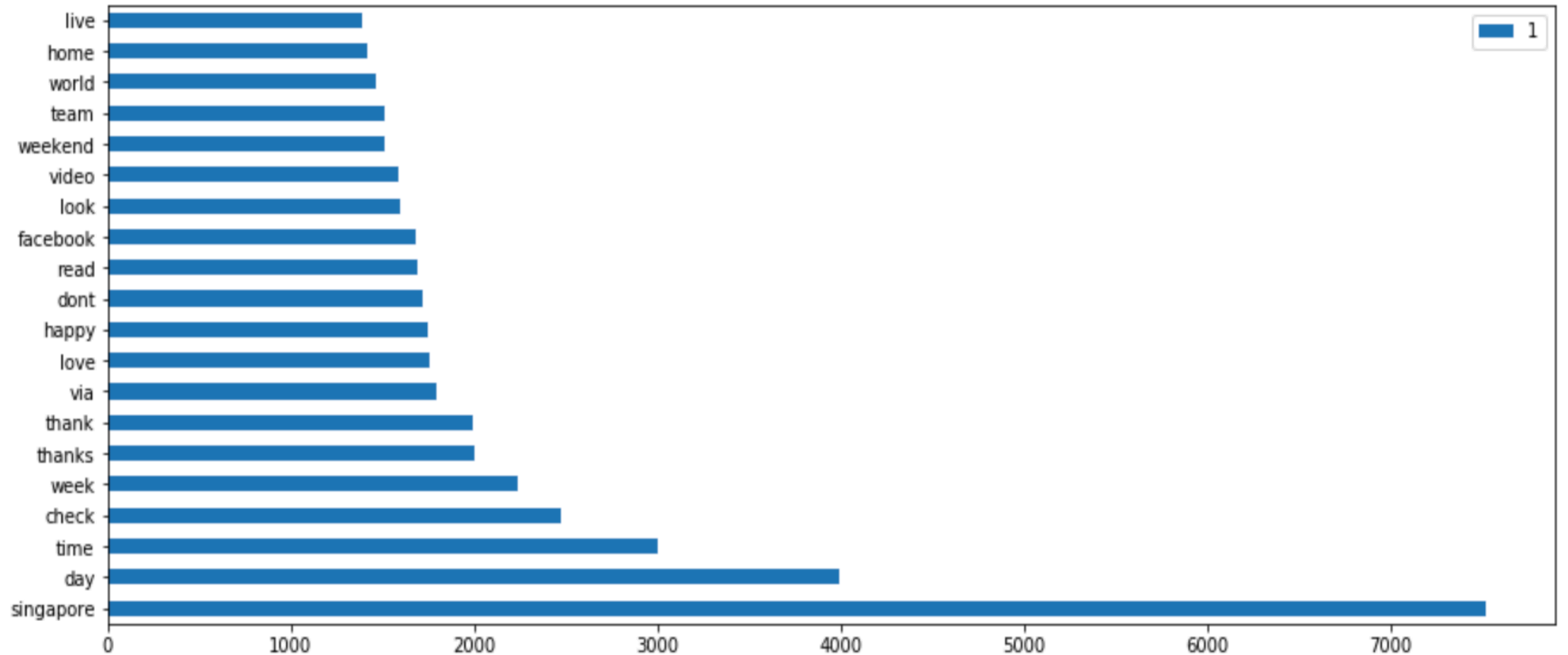
- XGBoost vs Feed forward Neural Network
- Accuracy of (83% vs 66%)
- Neural Network overfitted despite early stopping

	Australia	Belgium	Canada	Indonesia	Malaysia	New-Zealand	Singapore	United-Kingdom	United-States
Australia	202	3	2	4	1	5	9	3	16
Belgium	9	196	25	1	1	6	0	2	1
Canada	5	14	211	1	0	4	2	2	2
Indonesia	4	3	1	198	7	3	15	11	5
Malaysia	2	0	0	11	216	1	4	1	12
New-Zealand	6	8	2	1	1	209	2	1	2
Singapore	2	0	1	5	0	0	212	22	8
United-Kingdom	17	3	3	3	1	2	31	174	11
United-States	7	1	5	7	0	1	9	10	210

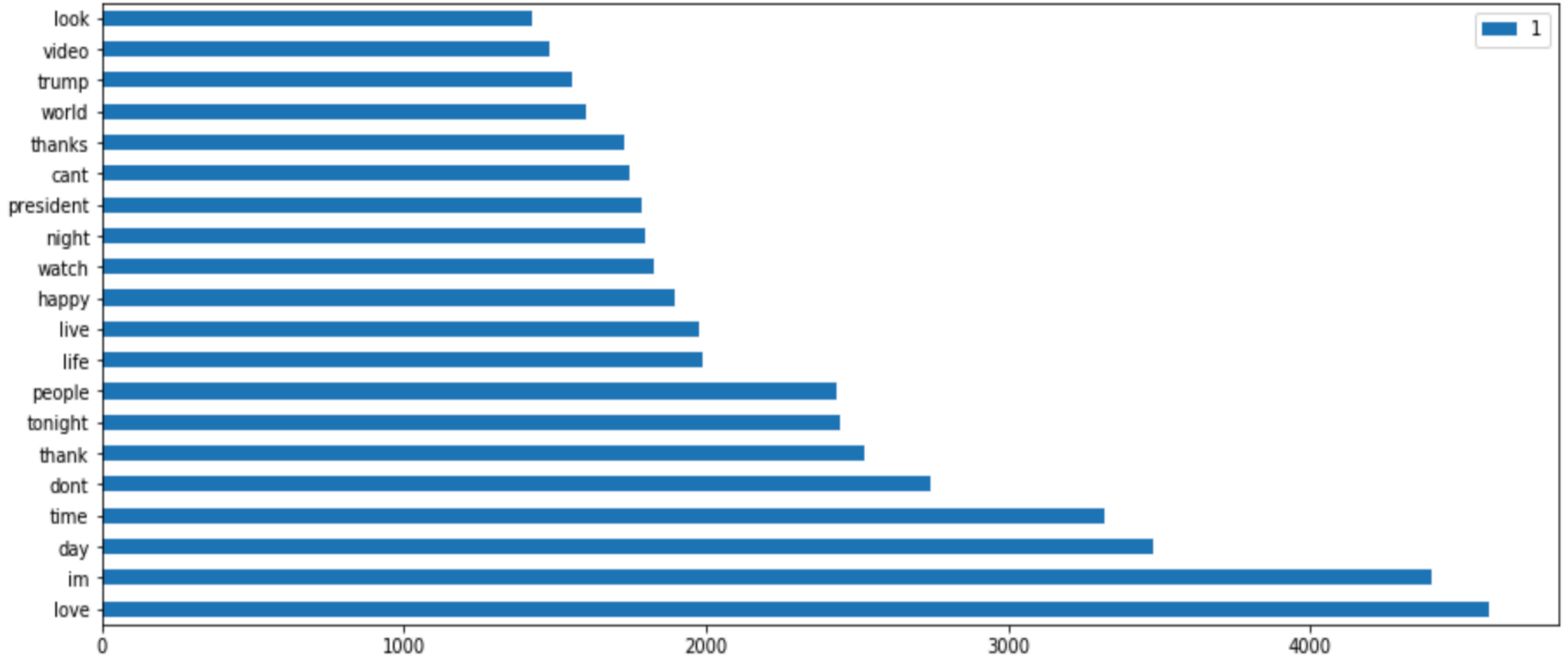
## Possible Reason for the low accuracy

- Over lapping common words for different countries

# Common Words (Singapore)



# Common Words (United-States)



# DEPLOYMENT

- **Built a user analytics dashboard hosted on Heroku**
  - Activity Patterns
  - Region & Country Prediction
  - Text analysis
- **Demo**

# QUESTIONS?

Live URL @ <https://folloing.com>