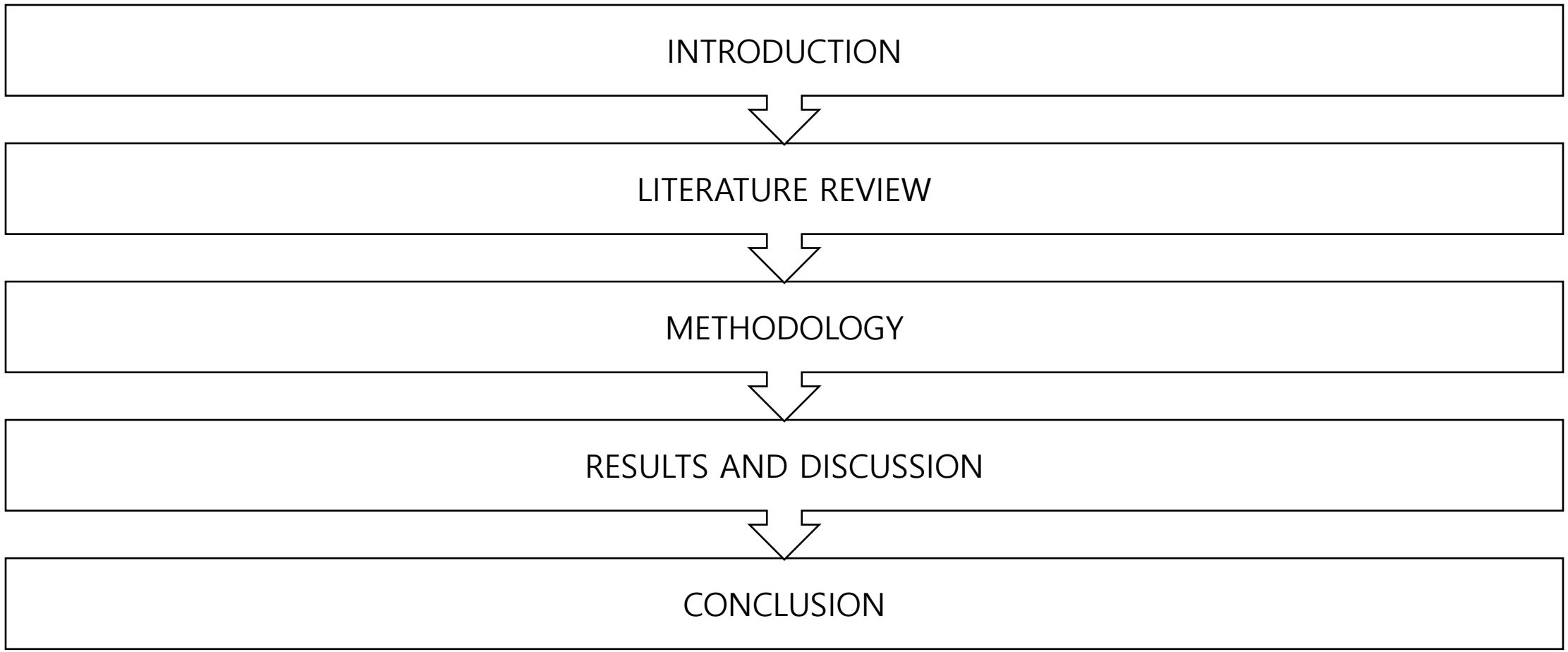
An Unsupervised Misinformation Detection Framework to Analyze the Users using COVID-19 Twitter Data

17102042 김민선


INTRODUCTION
LITERATURE REVIEW
METHODOLOGY
RESULTS AND DISCUSSION
CONCLUSION

INTRODUCTION

- After the spread of Covid-19, social media has become most frequently use source of sharing the news and information
- However, misinformation disseminates in social media Literature
- Earlier works relying on manual or annotation-based approaches posed limitations: 1) time and labor-intensive, 2) requires domain knowledge to identify the context of the news

INTRODUCTION

- Fake news tends to spread significantly faster, farther, deeper and more broadly than truth (high likelihood of retweeting false information by common public, not bots)

- This research propose an unsupervised framework for detecting misinformed content and the users who may be the source or susceptible to sharing misinformation.

- The proposed framework leverages popularity of the tweets and the credibility of their sources (users) to identify the misinformed users and contents.

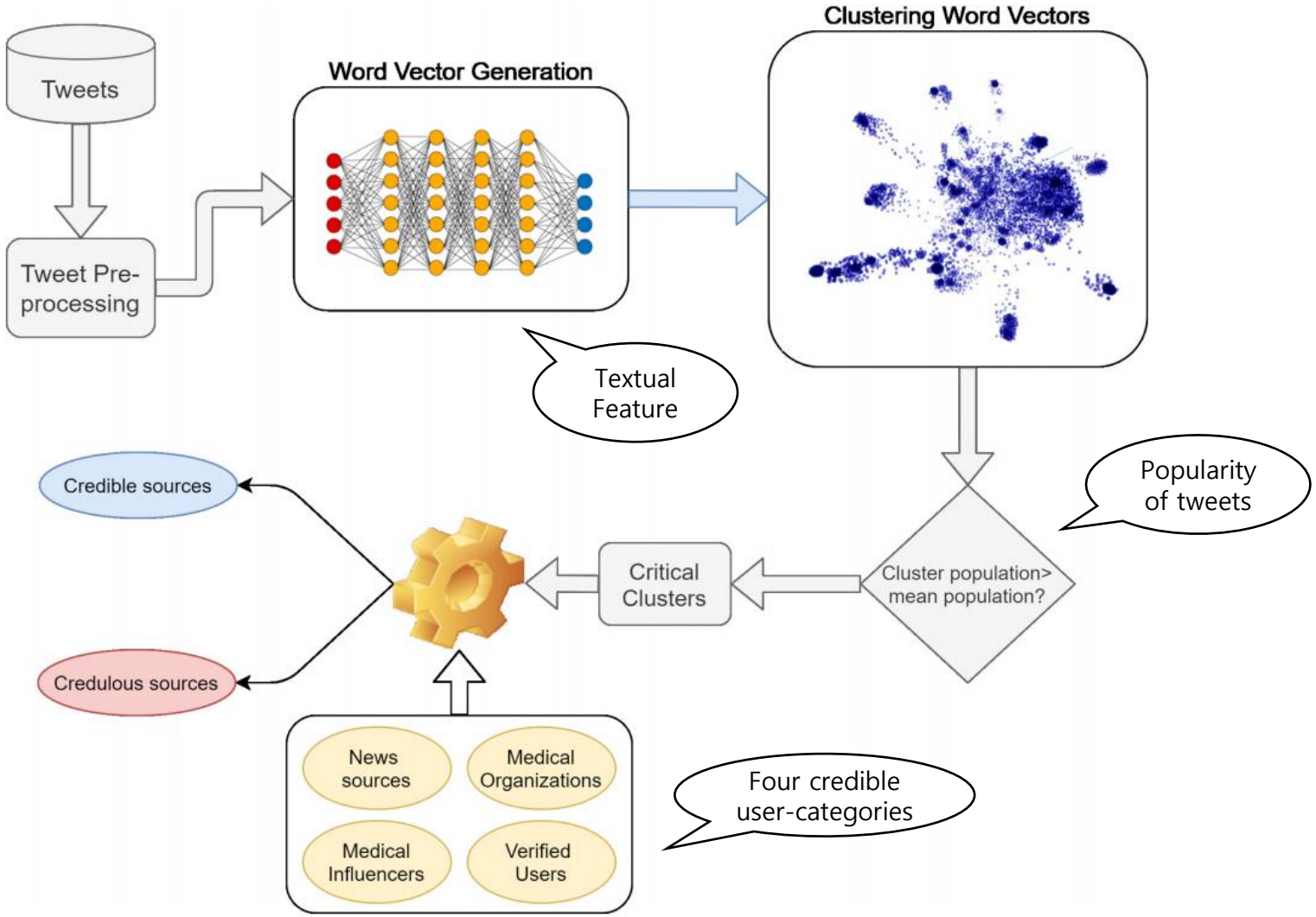
PROPOSED FRAMEWORK

- 1. Identification of popular/influential tweets and then credibility of them
    - a set of semantically similar critical clusters generated
- 2. Backtrack the tweet-retweet connection to identify the original-poster of tweets for each cluster
    - Veracity of a cluster is measured in terms of the proportion of the credible users (news sources, medical organizations, medical influencers, verified accounts)
    - Accuracy validation, Performance validation

LITERATURE REVIEW

- Text-based/message-centric models highly require enough labeled data to identify the fake news writers' textual features
- User-centric approaches is developed
  - Vosoughi et al. revealed that humans were much more likely to spread the misinformation not bot.
  - They also suggested that false news disffuse farther, more in-depth, faster and deeply ->
- Hybrid approaches appears that use text and the user-level information to identify the misinformed content
  - Popularity of the tweets and retweets are also used in this model

PROPOSED FRAMEWORK



METHODOLOGY

1. Data collection and Preprocessing
   - Twitter Streaming API
   - Extract only essential features
   - Remove impurities of text data

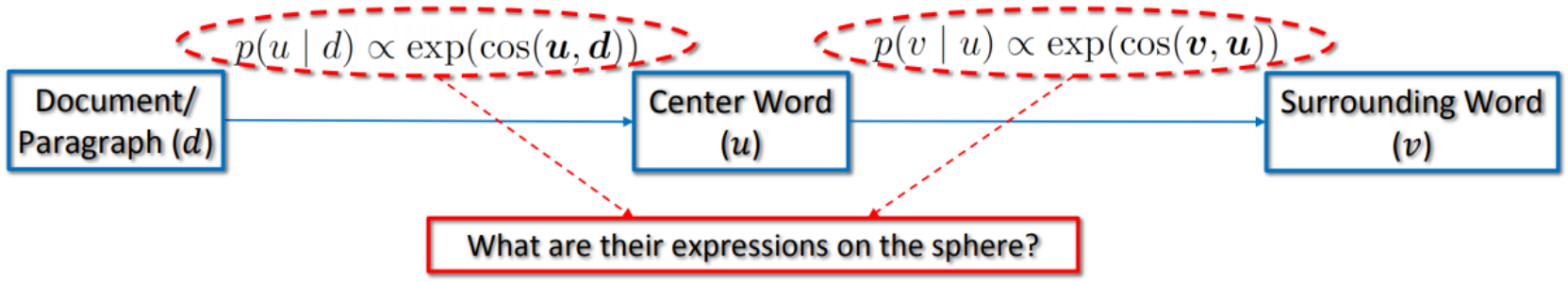
METHODOLOGY

2. Identifying Critical Clusters

- First, the semantically similar tweets are grouped to form clusters
  - JoSe, a spherical text embedding generation deep learing model
  - High performance in clustering vector representations with less run-time requirement (not only local context but also global context co-occurrence)

❑ "**Spherical**": Embeddings are trained on the unit sphere, where vector norms are ignored and directional similarity is directly optimized

❑ "**Text Embedding**": Instead of training word embeddings only, we jointly train paragraph (document) embeddings with word embeddings to capture the local and global contexts in text embedding



JoSE: Spherical Text Embedding

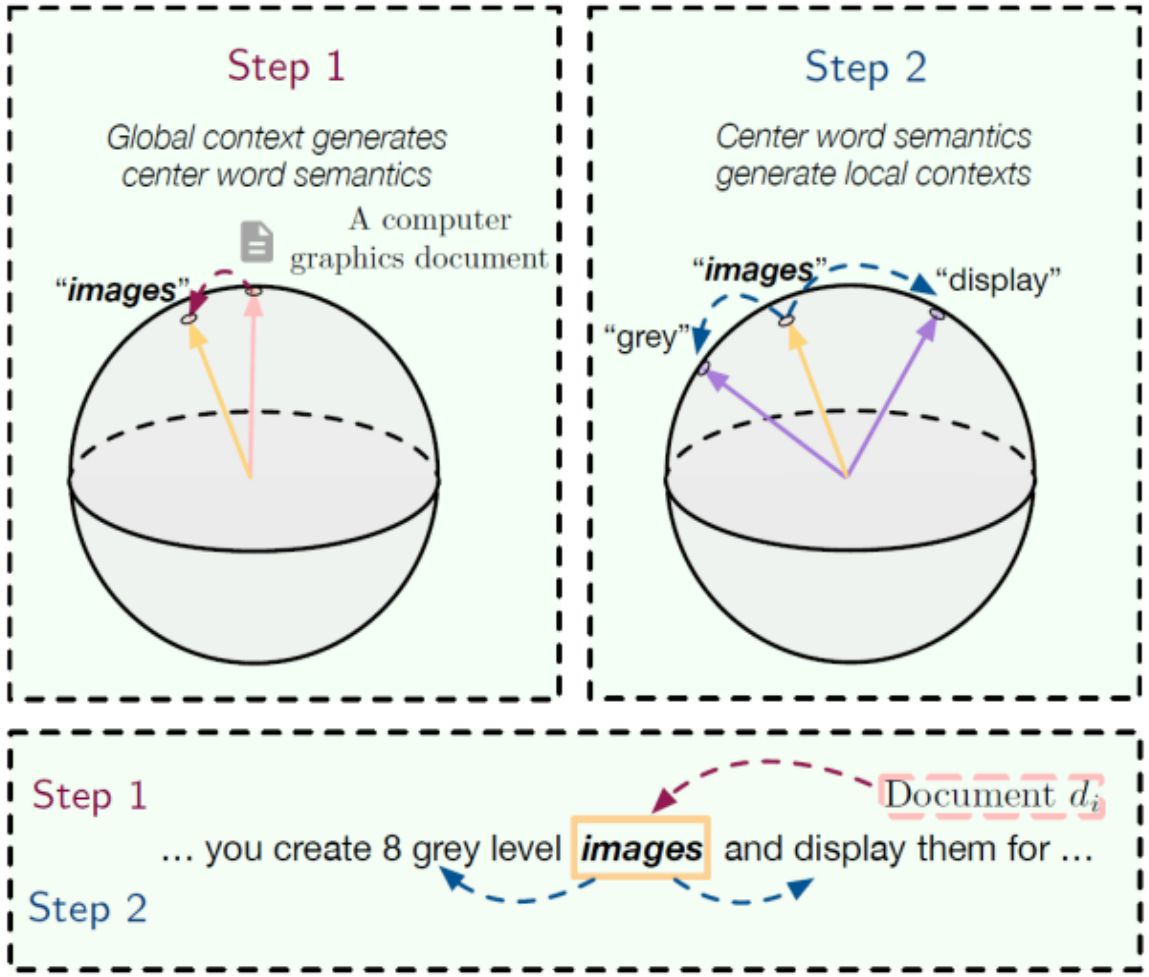


Table 1: Spearman rank correlation on word similarity evaluation.

Embedding Space	Model	WordSim353	MEN	SimLex999
Euclidean	Word2Vec	0.711	0.726	0.311
	GloVe	0.598	0.690	0.321
	fastText	0.697	0.722	0.303
	BERT	0.477	0.594	0.287
Poincaré	Poincaré GloVe	0.623	0.652	0.321
Spherical	**JoSE**	**0.739**	**0.748**	**0.339**

Table 2: Document clustering evaluation on the 20 Newsgroup dataset.

Embedding	Clus. Alg.	MI	NMI	ARI	Purity
Avg. W2V	K-Means	1.299 ± 0.031	0.445 ± 0.009	0.247 ± 0.008	0.408 ± 0.014
	SK-Means	1.328 ± 0.024	0.453 ± 0.009	0.250 ± 0.008	0.419 ± 0.012
SIF	K-Means	0.893 ± 0.028	0.308 ± 0.009	0.137 ± 0.006	0.285 ± 0.011
	SK-Means	0.958 ± 0.012	0.322 ± 0.004	0.164 ± 0.004	0.331 ± 0.005
BERT	K-Means	0.719 ± 0.013	0.248 ± 0.004	0.100 ± 0.003	0.233 ± 0.005
	SK-Means	0.854 ± 0.022	0.289 ± 0.008	0.127 ± 0.003	0.281 ± 0.010
Doc2Vec	K-Means	1.856 ± 0.020	0.626 ± 0.006	0.469 ± 0.015	0.640 ± 0.016
	SK-Means	1.876 ± 0.020	0.630 ± 0.007	0.494 ± 0.012	0.648 ± 0.017
JoSE	K-Means	1.975 ± 0.026	0.663 ± 0.008	0.556 ± 0.018	0.711 ± 0.020
	SK-Means	**1.982** ± 0.034	**0.664** ± 0.010	**0.568** ± 0.020	**0.721** ± 0.029

Spherical embeddings with better performance with a spherical clustering approach

METHODOLOGY

2. Identifying Critical Clusters

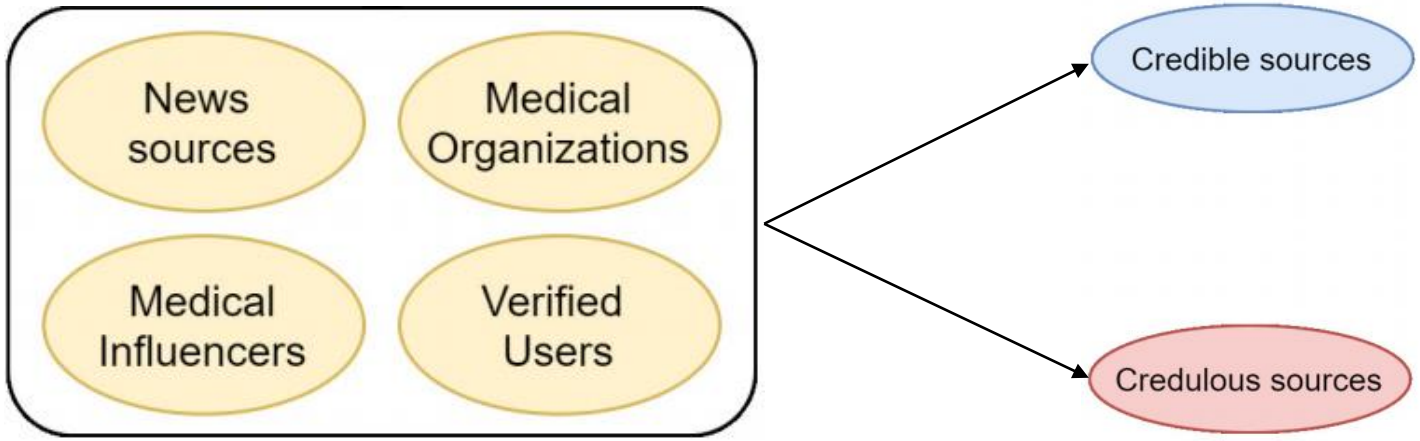
- Second, the clusters with high tweet populations are identified, and labeld *critical clusters*
  - SK-means, the spherical clustering algorithm
  - Elbow curve method to determine the optimal value of *k*
  - Threshold-based approach (mean cluster population)

$$C_R = \{C_i \text{ such that } |C_i| \geq \mu\}$$

* Elbow curve method
1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

METHODOLOGY

3. Identifying the credible and credulous critical clusters
   - Backtrack the tweet-retweet connection -> identify the source
     User_descritpions -> identify credulous users and contents
   - Divide critical clusters into two categories
     - Credible clusters: tweets from credible sources -> four primary categories
     - Credulous clusters: misinformed content not been verified



METHODOLOGY

3. Identifying the credible and credulous critical clusters
   - Keep the *seed* list of users by updating dynamically, using the user-description of verified users
   - Filter out the credible sourced clusters by backtracking all the tweets to their source users: *M[i]*
   - All the retweeted-users are labeled as one of the 4 categories or '*Unknown user*'
   - Clusters categorization into *credible* and *credulous* clusters

$$N[i] = \sum N_{credible}[i] + N_{unk}[i],$$

$$p_{unk}[i] = \frac{N_{unk}[i] - \sum_{s=credible} N_s[i]}{\sum_{s=credible} N_s[i]},$$

$$\mu_{p_{unk}} = \frac{\sum_{i=1}^{m} p_{unk}[i]}{m}$$

Calculate the proportion of unknown users and mean of all clusters

$$C_R[i] = \begin{cases} credible & \text{if } p_{unk}[i] \leq \mu_{p_{unk}} \\ credulous & \text{if } p_{unk}[i] > \mu_{p_{unk}} \end{cases}$$

METHODOLOGY

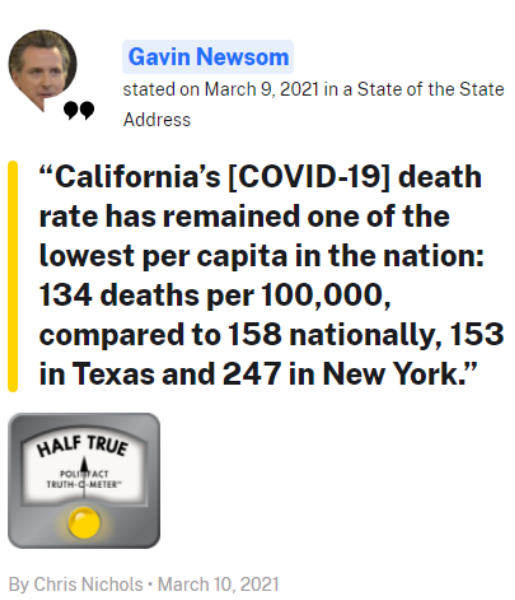
3. Emotion data labeling
   - Emotional coefficient per user categories
     : *Sadness, fear, anger,* and *joy*
   - Use the list of keywords curated by Saif M. Mohammad, which provide the label and score, that defines the intensity of that emotion for a given keyword when it is mentioned.
   - Average of the emotional score of all the keywords for each emotion
   - Tweet is labeled with the emotion of highest value

$$\{es_{anger}, es_{sadness}, es_{fear}, es_{joy}\}$$

RESULTS AND DISCUSSION

A. Ground Truth Data

- Validate model on two public fake new Twitter datasets
  - Annotations collected from Politifact and GossipCop website



RESULTS AND DISCUSSION(추가필요)

B. Accuracy Performance Evaluation

- Compared with baseline fake-news detection models
- Generate the set of events from the credible news articles by using K-means clustering and then perform online clustering to filter out fake news from the existing credible clusters

- K-means Clustering
- Support Vector Machine
- Convolutional Neural Network
- Social Article Fusion

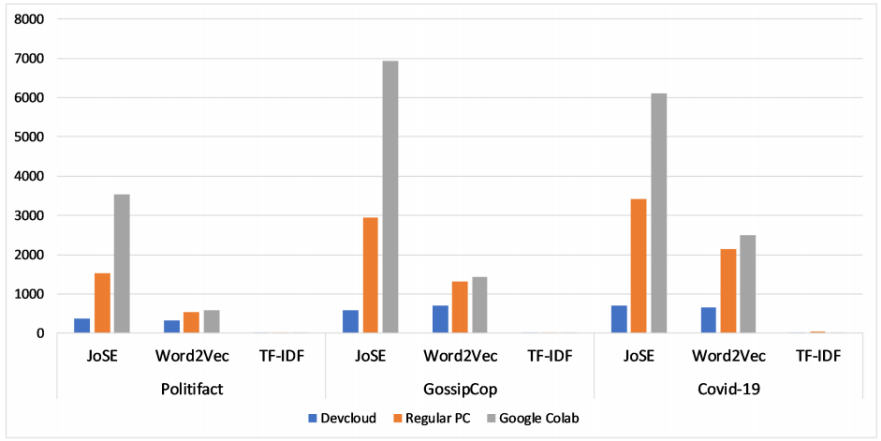
Models	Politifact (477,019 tweets)			GossipCop (1,274,372 tweets)		
	Recall	*Precision*	*F1 score*	*Recall*	*Precision*	*F1 score*
JoSE+SK-means +user-info(proposed)	**0.84488**	**0.96122**	**0.89931**	**0.98228**	**0.94866**	**0.96517**
JoSE + K-means + user-info [23]	0.79439	0.79728	0.795546	0.96090	0.94463	0.95269
SVM (polynomial kernel) + TF-IDF [24]	0.883057	0.90729	0.895009	0.52683	0.65688	0.52102
SVM (polynomial kernel) + Word2Vec [24]	0.9075	0.84649	0.87593	0.874381	0.875522	0.874182
CNN + TF-IDF [25]	0.84147	0.90933	0.87407	0.53709	0.6568	0.52107
CNN + Word2Vec [25]	0.82834	0.78529	0.80225	0.79782	0.80129	0.79829
SAF [20]	0.8875	0.85429	0.87057	0.923581	0.89052	0.90674

Models	Politifact (k=10)			GossipCop (k=15)		
	Recall	*Precision*	*F1 score*	*Recall*	*Precision*	*F1 score*
All clusters	0.465655	0.603849	0.525824	0.779422	0.764906	0.772096
Critical Clusters	**0.844889**	**0.961228**	**0.899312**	**0.98228**	**0.948661**	**0.965178**

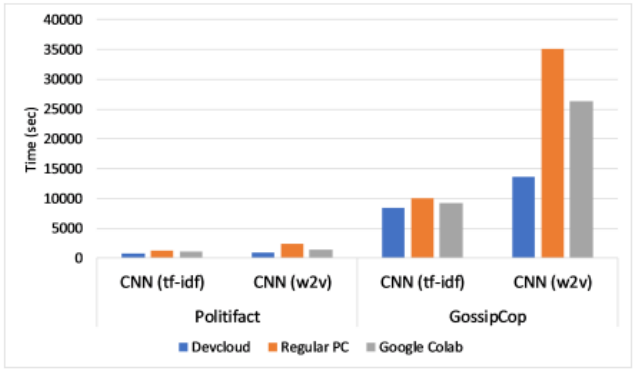
RESULTS AND DISCUSSION

C. Time Performance Evaluation

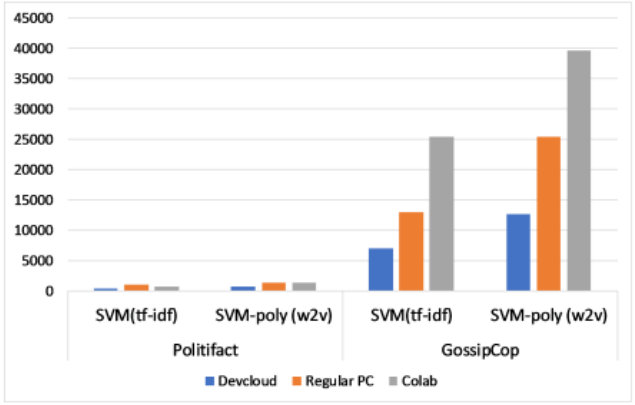
- Embedding generation task (Politifact, GossipCop, COVID-19)
- Clustering classification task



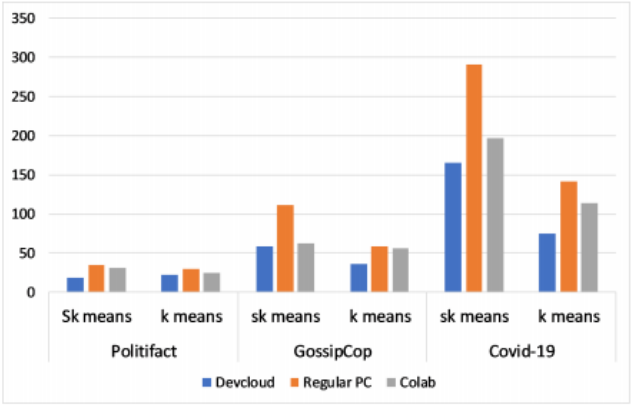
Embedding generation models



(a) Convolutional Neural Network



(b) Support Vector Machine

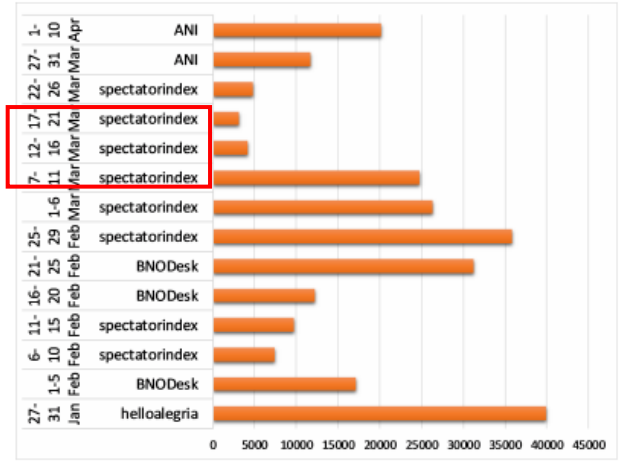


(c) Spherical K-means and K-means

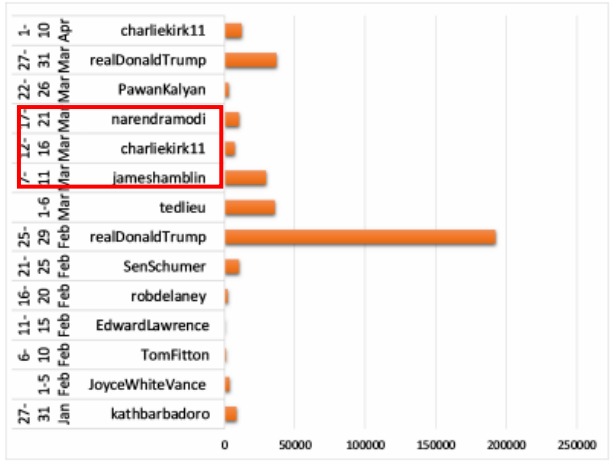
Extensive advantage of using DevCloud in terms of run-time for all tasks with variable sizes

RESULTS AND DISCUSSION

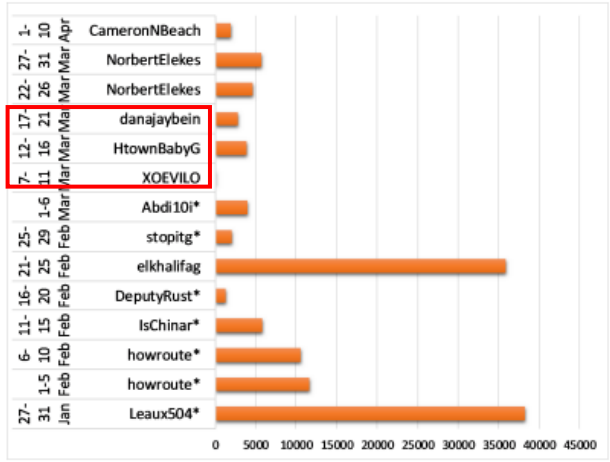
D. Validating the identified credulous *COVID-19* tweet clusters and users
   - Credulous clusters were searched manually to validate the results of proposed model
   - Validation was also performed by checking the status of identified credulous users on Twitter



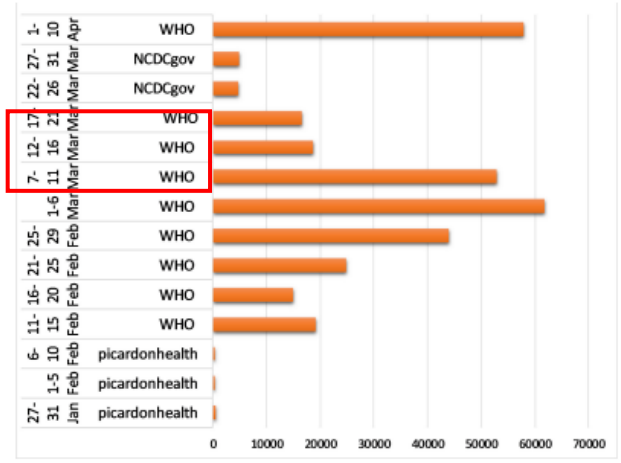
(a) News accounts



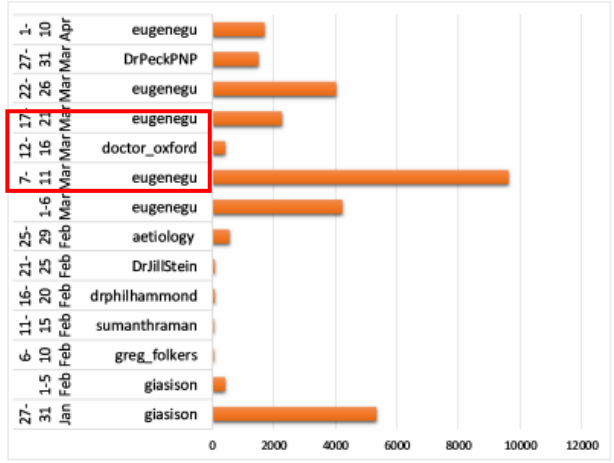
(b) Verified accounts



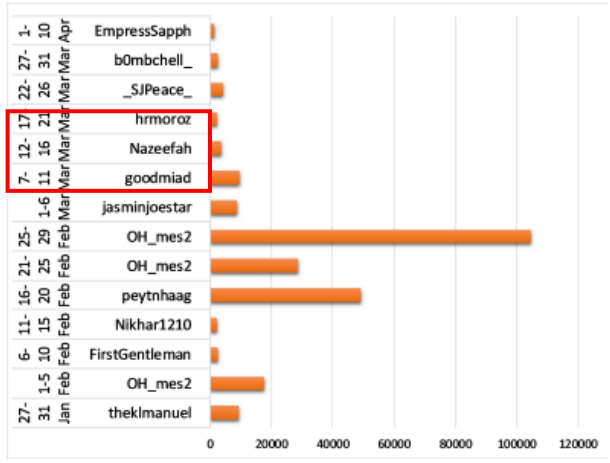
(c) Credulous accounts



(d) Medical organization accounts



(e) Medical influencers accouts



(f) Popular unknown user accounts

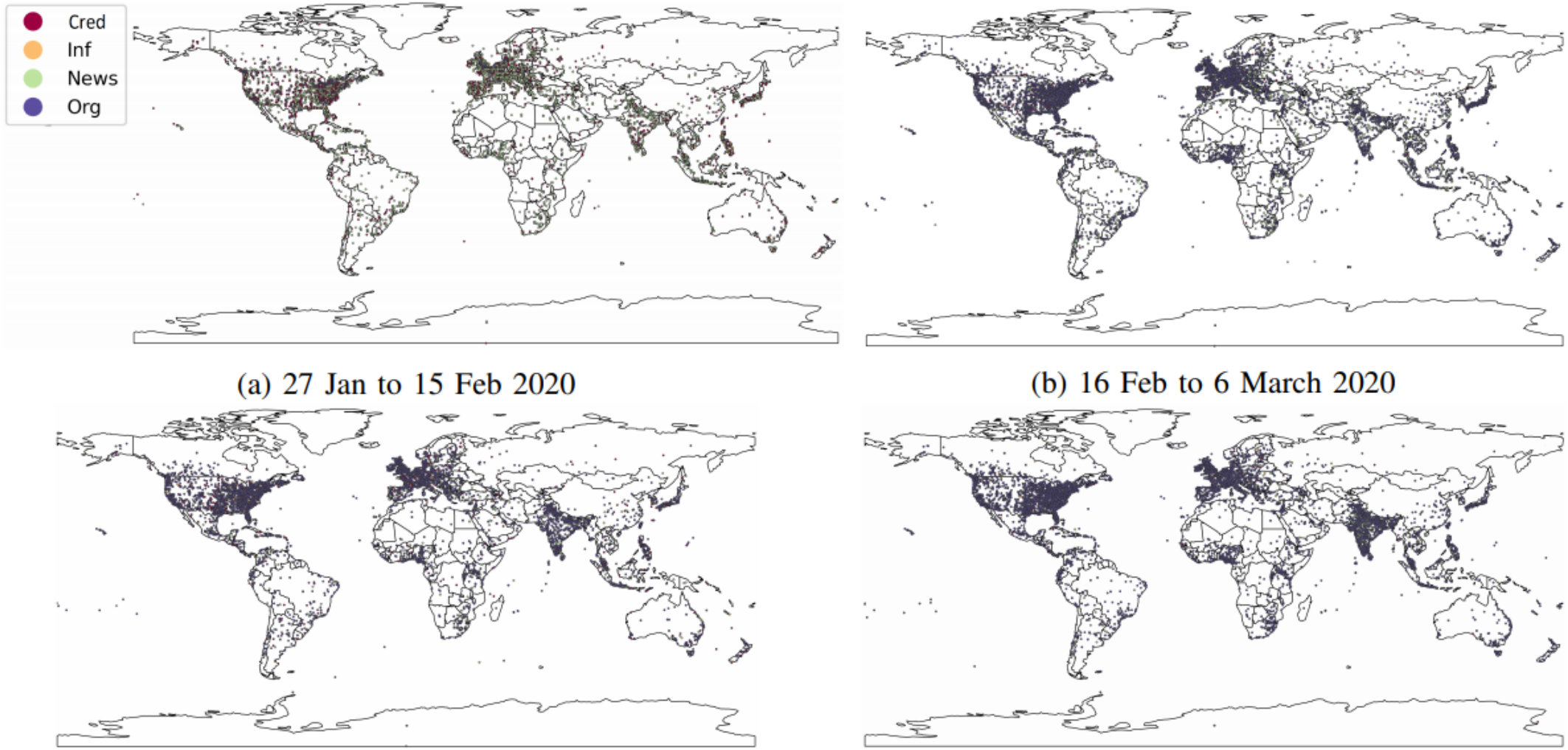
Fig. 4: The number of tweets generated per top retweeted user in the 5-day interval for each user category for 27 Jan to 10 Apr 2020. Asterisk user-names were found to be suspended on Twitter.

During mid-march, Twitter improved the security features

All the identified users belonged to the pre-mid-march duration

RESULTS AND DISCUSSION

D. Validating the identified credulous *COVID-19* tweet clusters and users
   - Identify the geographical location of influenced users (user-locations mentioned in the user-profile)
   - Do not consider verified users' followers and ordinary users



(a) 27 Jan to 15 Feb 2020

(b) 16 Feb to 6 March 2020

(c) 7 March to 26 March 2020

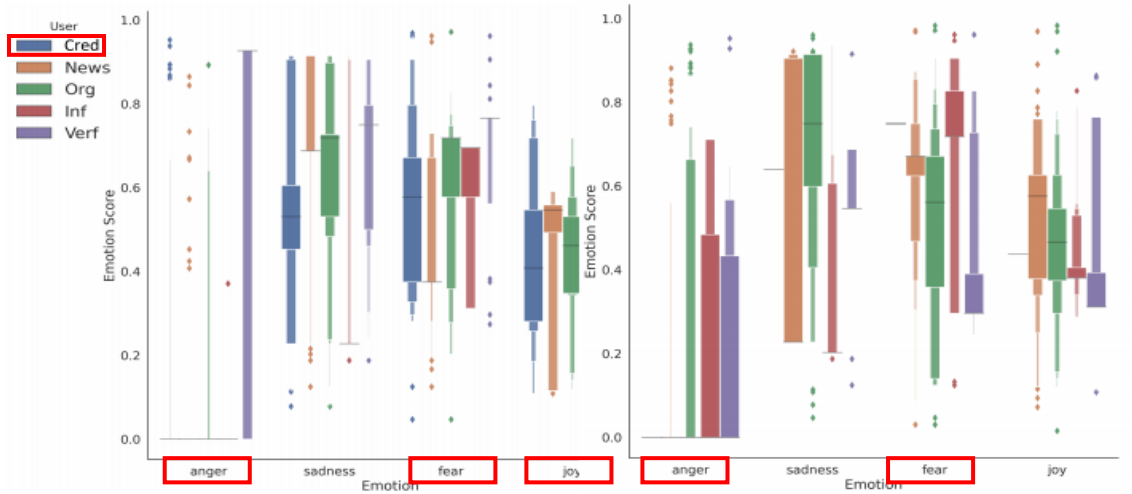
(d) 27 March to 10 April 2020

Fig. 5: The temporal change in the geographic plot of tweets from different types of user types. The abbreviations used in the legend are as follows: Cred:Credulous, News: News, Org: Medical Organization, Inf: Medical Influencers

As the spread of the COVID-19 cases increases, the tweets and retweets also tend to increase

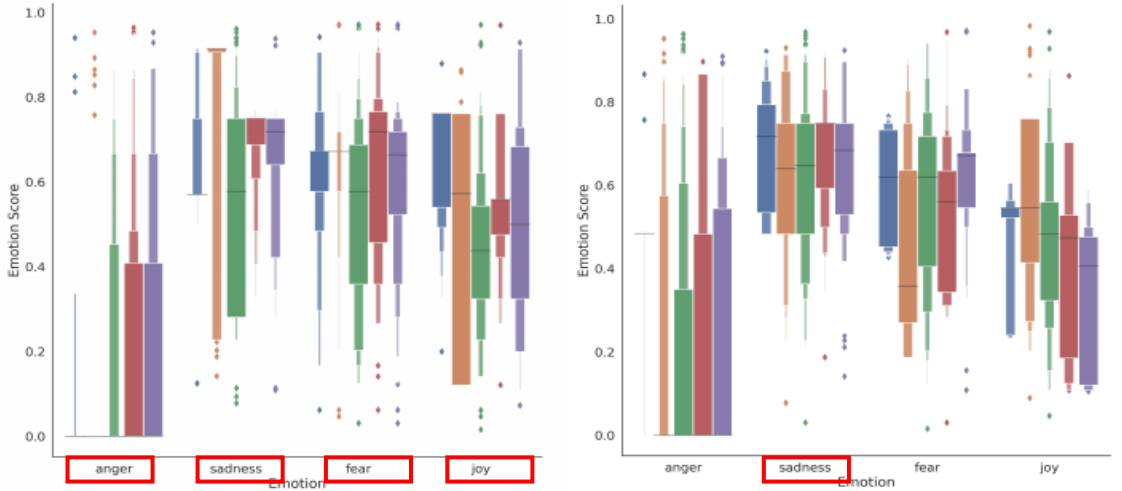
RESULTS AND DISCUSSION

E. Validating the credulous *COVID-19* tweet clusters using Emotional Analysis
   - Emotion intensity scores and number of tweets per emotion for the different user-categories
   - Validates that the content of credulous clusters tend to show high proportion of anger and fear generating tweets with high intensities



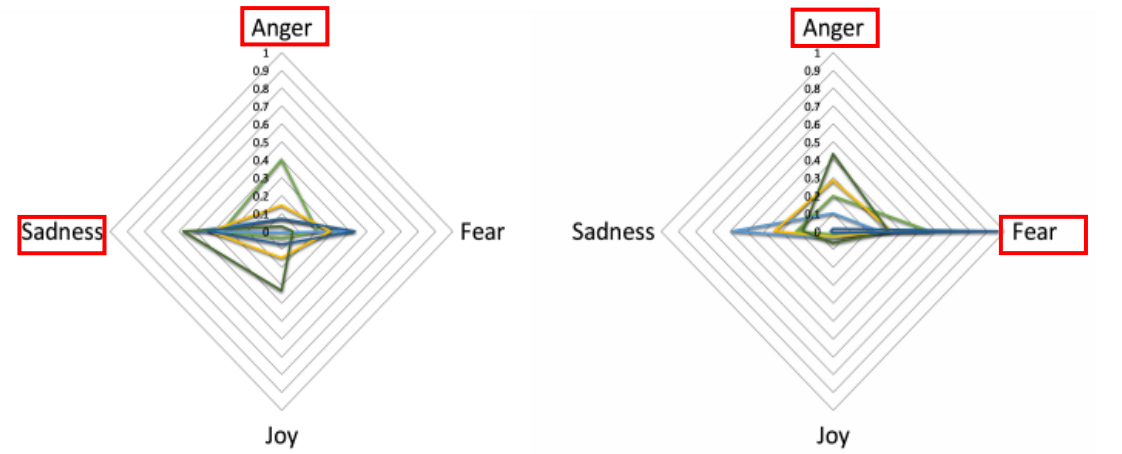
(a) 27 Jan to 15 Feb 2020 (b) 16 Feb to 6 March 2020

(a) 27 Jan to 15 Feb 2020 (b) 16 Feb to 6 March 2020



(c) 7 March to 26 March 2020 (d) 27 March to 10 April 2020

High intensity of sad and fearful tweets in credulous clusters



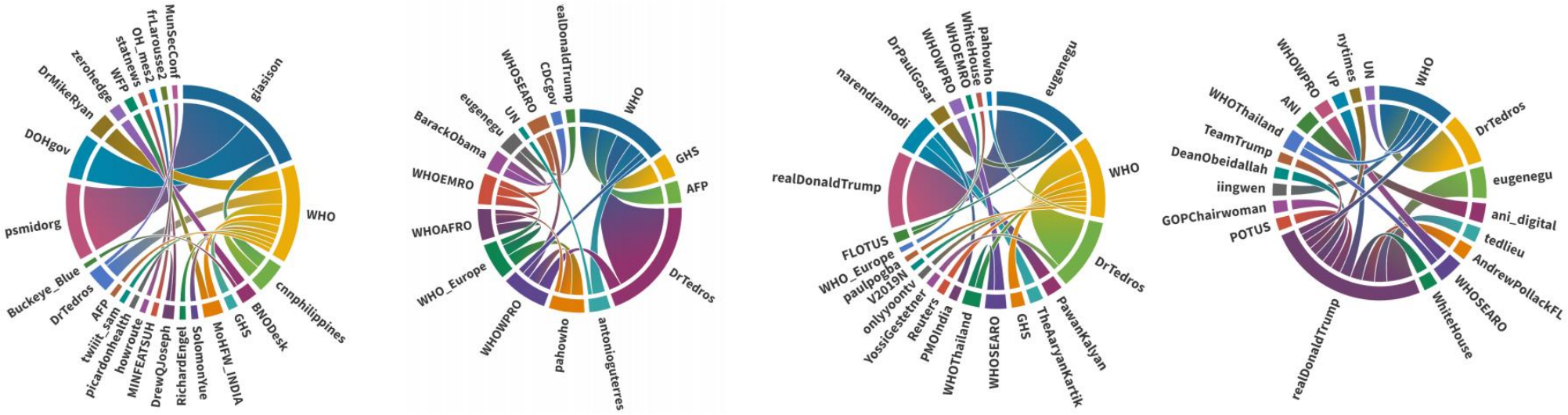
(c) 7 March to 26 March 2020 (d) 27 March to 10 April 2020

High proportion of angry, sad, and fearful tweets in credulous clusters

RESULTS AND DISCUSSION

F. Analyzing User-Interactions on *COVID-19* dataset

- Identify the most popular user-interactions in the whole data by using the poster of the tweet and the mentioned or retweeted users
- Change in the pattern of user-interactions over the whole duration



(a) 27 Jan to 15 Feb 2020 (b) 16 Feb to 6 March 2020 (c) 7 March to 26 March 2020 (d) 27 March to 10 April 2020

Primary portion of major users are medical influencers, organizations, news, and verified accounts

CONCLUSION

- Identify the misinformed content and the credulous users
- Semantically similar cluster to extract the most influential content
- Credible sources help in filtering out the credulous users and related information
- Accuracy performance and Time performance increases upto 5% and 10% respectively
- Higher affinity toward the angry, sad, and fearful tweets in the credulous clusters

FUTURE WORK

I. Performing a comparative analysis of credulous and credible sourced clusters in geographical spread of the type of user-interactions

II. Eliminating the possibility of sharing misinformation in credible sources

Thank you