Natural language processing approaches: Various NLP techniques have been applied in studying the Twitter data from different perspective, such as semantic representations, topic modeling, communities and semantics, and etc. Primary problems, approaches and finds are briefed as follows.

First, though word embedding has been the "standard" approach to represent semantics of texts, a key issue in representing semantics of tweets is that the representation usually trend to becoming confusing when aggregating the embedding vectors for a tweet, and the longer the tweet the more confusing. This issue is illustrated in **Figure NLP#8**, and in fact it has been a commonly agreed drawback of word embedding, and we have been looking for effective and efficient alternatives. A couple of methods based on phrases and algebraic topology have been invented [MF19] [MF20] [KM20]. Also, a number of variants extending the phrase based methods have been proposed utilizing phrase clustering and graph signal processing. **Table NLP#1** shows the performance of several such variants and some of them outperform the provided fastText embeddings and topic vectors. Moreover, **Figure NLP#9** and **Figure NLP#10** show the statistics of the part-of-speech (POS) pairs of the extracted phrases on which our embedding methods are based. In addition, during CP4, the phrase based methods were evolved with more effective and efficient phrase detection and extraction approaches, and further extended them to a new type of knowledge graph (named PKSG short for phrase-based sentiment graph) organizing the phrases by their co-occurrences defined by the scope of tweet. Moreover, this knowledge graph is equipped with phrase-based sentiments modeled by RNTN [SP13]. This knowledge graph can be interfaced to various downstream models such as graph embedding models and socio-cognitive models, and also itself can provide plentiful insights on the data. The construction pipeline of PKSG is illustrated in **Figure NLP#5**, and how PKSG can be projected and aligned to community structures is illustrated in **Figure #6**. An example application of PKSG is illustrated in **Figure NLP#1**.

Second, the community structures on Twitter need to be hierarchical and hybrid on connectivity and influencers. This was learned when trying to detect communities from the message propagation graph built upon retweets, replies and quotes (in which retweets are dominant). **Figure NLP#2** shows such an example extracted from the WhiteHelmet Twitter data. Both hierarchical community detection and influencer detection algorithms are developed. Regarding the hierarchical community detection, the classic Louvain algorithm is extended to a recursive version and the terminal conditions are parameterized by the modularity and the lowest desired cluster size. On the other hand, a modularity-based influencer detection method was also invented, and its core algorithm is summarized as: $\gamma = e^{\frac{e_{io} - \mathrm{a}_i \cdot \mathrm{b}_o}{1 - \mathrm{a}_i \cdot \mathrm{b}_o}}$, where $\mathcal{E} = \begin{bmatrix} e_{ii} & e_{io} \\ e_{oi} & e_{oo} \end{bmatrix}$ and $\mathrm{a}_i = e_{ii} + e_{io}$ and $\mathrm{b}_o = e_{io} + e_{oo}$. The method adapts a greedy search starting with a subset of nodes covering 50% of the edges.

Third, the influencers were particularly studied in CP3 regarding their relations to semantics. Beforehand, the Twitter data was reorganized as a time series. The stride is a week. Then firstly it was explored how persistent are the influencers, and **Figure NLP#3** shows a selection of influencers and their presence over time. It is concluded that a non-trivial number of influencers are relatively persistent, though many more are rather not, and these persistent influencers can help align communities over time as a great portion of community substructures are established centering at influencers. Secondly, how influencers affect semantics propagation was investigated. Topics are computed and utilized to represent semantics of tweets. Moreover, to reduce dimensionality, topics are further clustered by a recursive spectral clustering with size prediction and compensative classification invented by us. Each user then is embedded into the

topic cluster space. Thereby, influencers are compared with all other users as a whole with respect to their semantics. From the results shown in **Figure NLP#4**, influencers may significantly affect the semantics propagation over time. This study of influencers also implies that importance of understanding community structures in depth.

Fourth, it was studied how well semantics and communities are aligned in CP3. The conclusion from us is that typically these two items may not align. **Figure NLP#7** shows this result. The topic representation is computed by utilizing the same method as above. In addition, the persistence of topics over time was also studied and the result is shown in **Figure NLP#7**. In general, many topics are persistent. This observation is quite intuitive as the data was collected regarding a predetermined general topic (e.g. WhiteHelmet and COVID-19), and the fine-grained topics would not dramatically change within a short term. The correlation between the user graph constructed based on message propagation and the user graph based on semantic similarities is rather low in our experiments. However, it may not be rigorous to conclude that these two items do not have any relation or interaction. A potential explanation is that their relations have annihilated in the noise when considering the data as a whole. The community and semantics studies above have provided concrete evidence to the relations between user message propagation and the semantics of their tweets. Thus, exploring the interactions between semantics and community structures requires more sophisticated approaches, and our team has been working on this.

Fifth, in CP4, tweet response cascades also interested us. A pipeline was developed to track how semantics change along cascades. The semantics are represented by semantic units including noun phrases and core clause graphs. The core clause graphs are extracted from parse trees with further pruning and adjustment. An example is demonstrated in **Figure NLP#11**. Such a cascade is a data structure that is very intuitive to understand (compared to embeddings), which is important to downstream modeling such as socio-cognitive models. Also, since phrases can be extracted from those semantic units and can be further represented as real-valued vectors, this data structure is also friendly to downstream modeling. **Figure NLP#12** shows a concrete example of utilizing this data structure to capture semantic distribution over a cascade. The semantic distribution provides a ground truth point of view of how semantics vary along a cascade.

Sixth, another work centered at the phrases aforementioned is to compare distances between communities by utilizing phrase-clustering-based embeddings. The phrases extracted from the CP4 Twitter data are clustered. Then each phrase is represented as a one-hot vector, and each tweet can be represented as an aggregation of a set of such one-hot vectors. The distances between communities can be computed by utilizing such resulting phrase vectors. To evaluate the performance, the given narrative labels, represented as a narrative vector, are utilized to compute distances between communities. The two similarities are then compared to examine if they align to each other. The distances are all computed by Jensen-Shannon divergence. **Figure NLP#13** shows the comparison results, and it is safe to conclude that the phrase vectors basically align to the narrative vectors. And in **Figure NLP#14** a sample community distance graph is visualized by utilizing the phrase vectors. Additionally, in this study, it is learned that there does exist a non-trivial amount of community pairs that are relatively similar to each other, and also in between some of the pairs are very similar.

Seventh, the Sixth work was further extended to studying how communities behave over time and their relations to external events. This study was proceeded in two directions. One is to examine how communities are similar regarding semantics in each time interval. And the other one is to examine how

each community's semantics varies over time. **Figure NLP#15** shows two examples in the CP4 Twitter data of the first direction. It is straightforward to observe from the results that the distances between communities (if we assume that the communities are relatively stable) can vary dramatically over time. And such variation is led by the external events. **Figure NLP#16** shows the other direction with three example communities reflecting three different types of community behaviors. The first type is that there is no large semantic variation the whole time, the second type is that the semantics trend to have lower variation when significant external events happen, and the third type is that the semantics keep constantly and significantly varying over time. Another interesting point is that it was not observed yet any community trends to have more divergent semantics when significant external events happen. The observations in this direction imply that not all communities would be affected by external events, even though their semantics are still relevant to a given general topic (e.g. Venezuela protests), which provides another strong evidence to the importance of the community structures.

Eighth, following the seventh work, the correlation between external events and the semantics of communities over time was also studied. **Figure NLP#17** demonstrates how the three different types of communities described above respond to the external events. The texts of the external events are extracted manually from Wikipidia [VP19], and their semantics are represented by utilizing the phrase-clustering-based embeddings, the same as the tweets. And JS-divergence is utilized to compare the semantics. It is learned from the results that the first type communities keep being correlated to the external events with respect to semantics over time yet without explicit perturbation of semantic variation. In other words, the talks in those communities are always highly relevant to the general topics but are not affected by the external events too much. The second type communities also keep their semantics highly relevant to the general topics, and also the semantics are affected by the external events significantly (i.e. the people in those communities are more "updated"). The third type communities may not always keep their talks relevant to the general topics, compared to the other two types, and their talks may not be affected by the external events either.

Ninth, another important work extending the community structures is modeling the replies and quotes for each community. An Autoencoder model was developed for this purpose, and its architecture is shown in **Figure NLP#18**. The data is taken from the CP4 Venezuela Twitter data. The input and the output are all semantics represented by the phrase-clustering-based embeddings. The input is derived from a tweet, and the output is derived from a response (i.e. a reply or a quote). Only the tweets with non-trivial texts are considered. From our experiments, for some communities, this model can perform very well. In other words, the model can predict the semantics, represented by the phrase vectors; however, it is not universal for all communities, and for some cases, it could hardly learn any meaningful model. The key reason causing this issue is that in some communities responses to one tweet can have greatly divergent semantics. In other words, in those communities, the users may not have concord in their semantics and attitudes (though attitudes reflecting sentiments their concrete representations are still the uses of tokens), this phenomenon is more explicit in larger communities. Nonetheless, the quality of the community detection can also significantly impact the performance of this model. In our experiments, the communities are derived from a plain community detection method rather than a more sophisticated one such as the hierarchical community detection method aforementioned. And typically the Autoencoder model does work well for many small-scale communities. Thus, it may not be rigorous to conclude that the model could not perform better, but more extensive studies are surely needed.

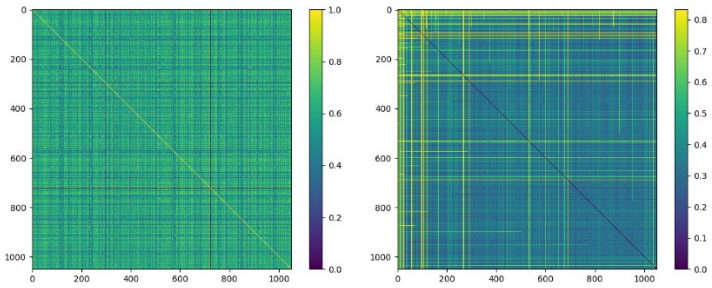

Figure NLP#8. The left figure shows the pairwise cosine similarities between popular tweets (i.e. no less than 10 responses) sampled from the WhiteHelmet Twitter data in CP3. The average is about 0.6, which causes trouble in distinguishing semantics from the tweets. The right figure shows the Jensen-Shannon divergences between token distributions of response tweets. The average is about 0.4, which also demonstrates the same issue.

8-Category CP4 Venezuela Twitter Sampled Data							
avg_node	**avg_edge**	**s_avg_edge**	**w_avg_edge**	**sp_avg_node**	**sp_w_avg_edge**	**ft_vec**	**topic_vec**
0.64	0.64	0.64	0.63	**0.65**	0.63	0.63	0.62
0.65 (9)	0.65 (9)	0.65 (9)	0.65 (9)	0.65 (8)	**0.67 (9)**	0.65 (9)	0.62 (8)

Table NLP#1. Document clustering task (on the CP4 Venezuela Twitter data) performance comparisons between various embedding methods. “avg_node” only consider tokens in extracted phrases. “avg_edge” considers the average embedding of each phrase (which is typically an edge in the parse tree). “s_avg_edge” considers phrases with specific part-of-speech (POS) tags. “w_avg_edge” further assigns each specified POS tag a weight. “sp_avg_node” applies graph signal processing techniques to compute the embeddings and focuses on nodes. “sp_w_avg_edge” instead focuses on edges. “ft_vec” is the fastText embeddings provided in the CP4 data. “topic_vec” is the topic vectors provided in the CP4 data. The first numeric row shows the normalized mutual information score for each method when the number of desired clusters is set to 8 exactly. The second numeric row shows the best score in a range of numbers of clusters.

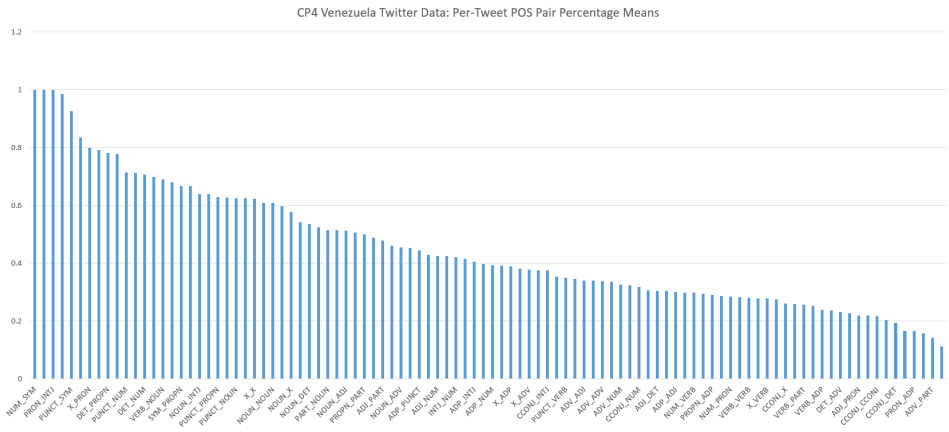

Figure NLP#9. Per-Tweet POS pair percentage means, i.e. in each tweet, in average, what is the percentage of each POS pair of the extracted phrases.

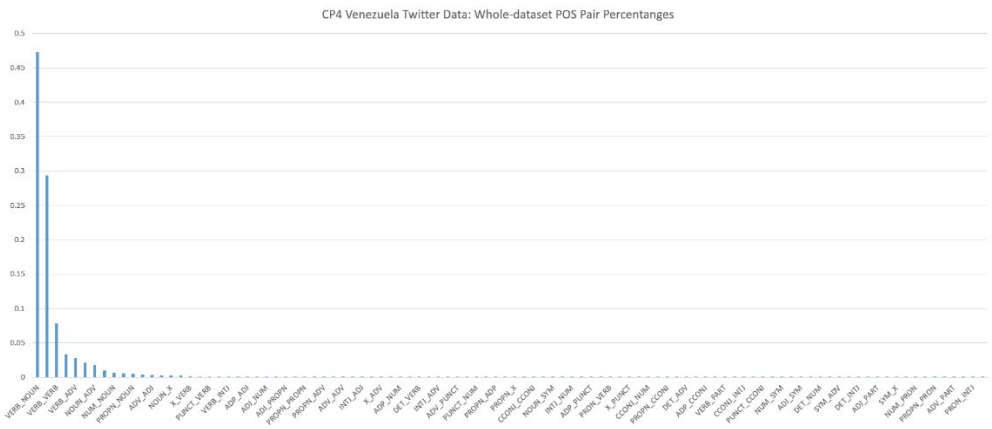

Figure NLP#10. In the whole CP4 Twitter data, what are the percentages of the POS pairs of the extracted phrases. It is straightforward that “VERB NOUN” and “VERB VERB” are dominant in the data, which can significantly help reduce the input size when computing the embeddings.

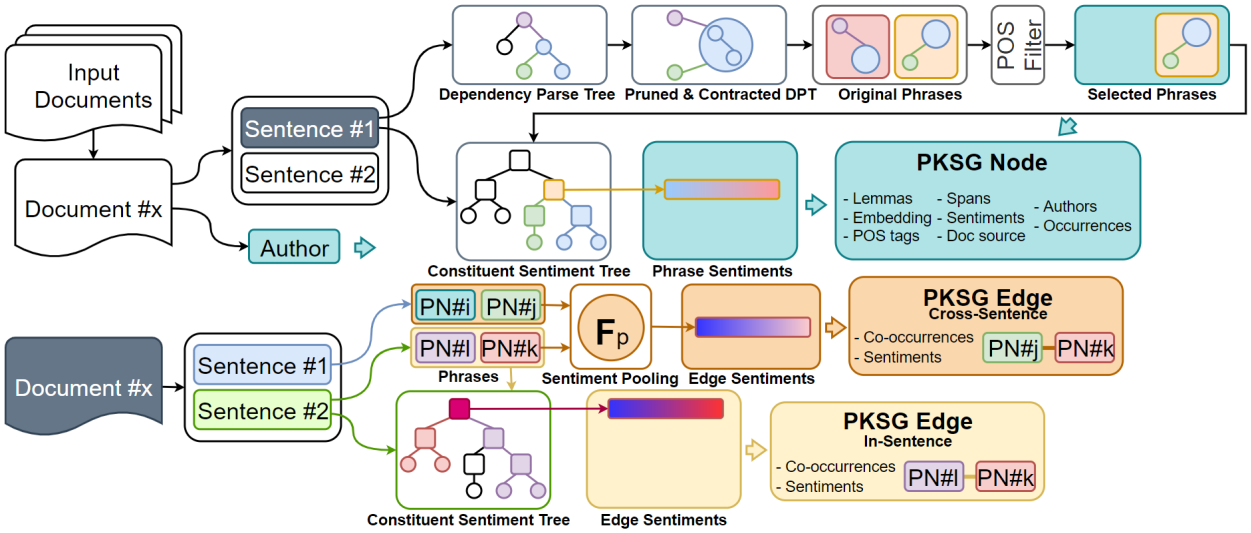

Figure NLP#5. The construction pipeline for PKSG. Each node in a PKSG is a phrase extracted from a parse tree, and assigned a sentiment vector by the RNTN model. Each edge is also assigned a sentiment vector. If the nodes on an edge co-occur in a sentence, then the RNTN model is utilized, otherwise a polarized aggregation is used instead.

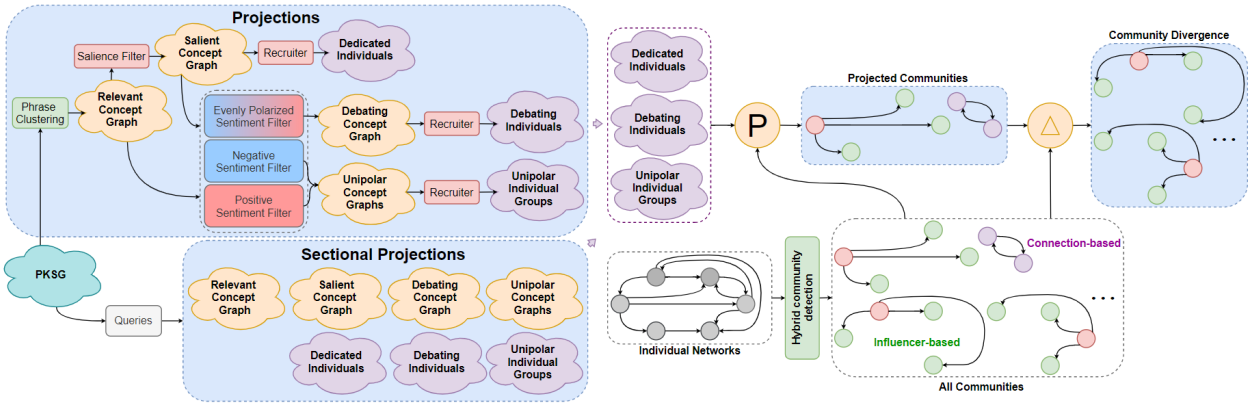

Figure NLP#6. A PKSG can be projected to various spaces, and aligned to community structures. For example, the debating concept graph containing phrases carrying highly polarized sentiments can be mapped to the corresponding user groups, and the user groups can further be aligned to communities detected with different inductive biases.

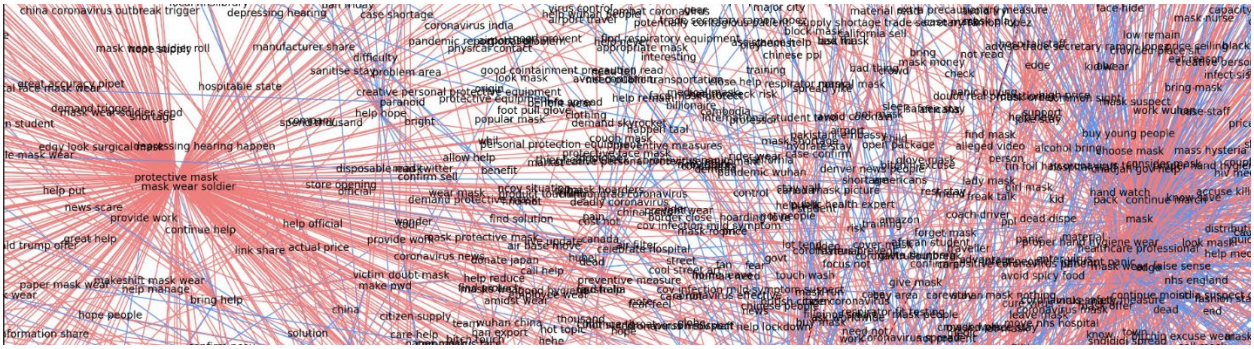

Figure NLP#1. A part of a PKSG built on the Jan.2020 Twitter data for COVID-19 (as a branch of SocialSim) showing the sentiments on co-occurring phrases with respect to “wearing mask”, where red is positive and blue is negative.

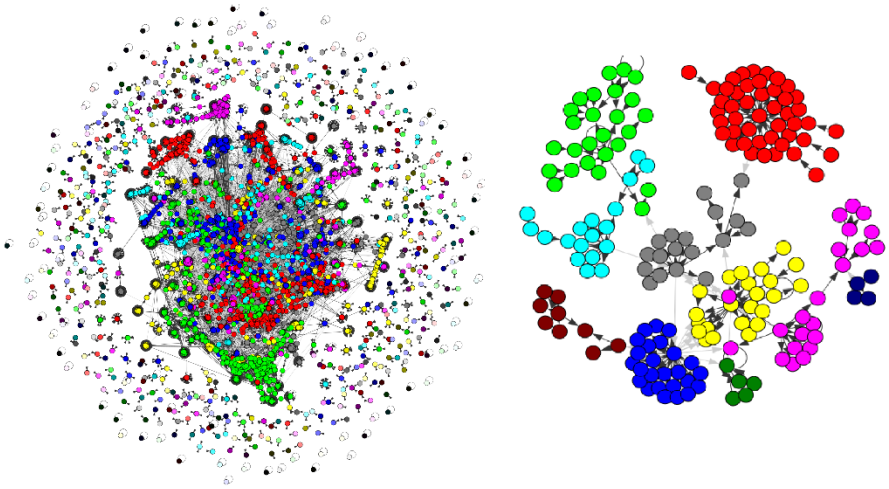

Figure NLP#2. The community structure detected by the Louvain algorithm for the WhiteHelmet Twitter data from Apr.13 to Apr. 19 2018 is shown in the left. A sub-community structure detected from one of

the communities by a recursive hierarchical community detection algorithm is shown in the right. In addition, it is straightforward to see that the sub-community structure is influencer-based.

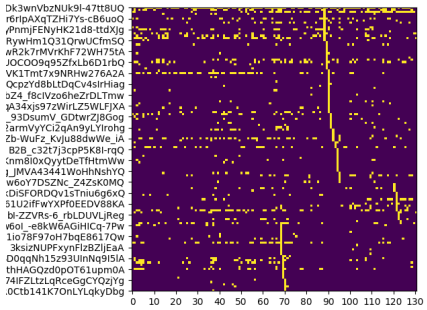

Figure NLP#3. The persistence of influencers. The x-axis is the time intervals, and the y-axis is the influencer IDs. A light spot indicate the presence of an influencer in a time interval.

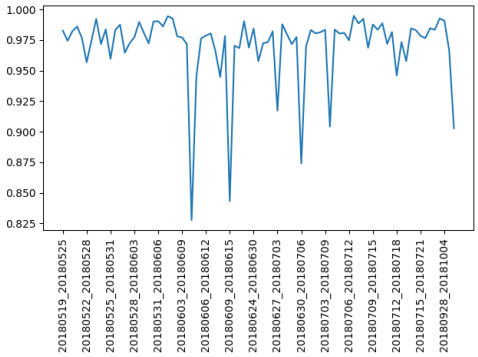

Figure NLP#4. Semantic similarities between influencers and all other users over time. X-asix is time intervals and y-axis is the cosine similarities.

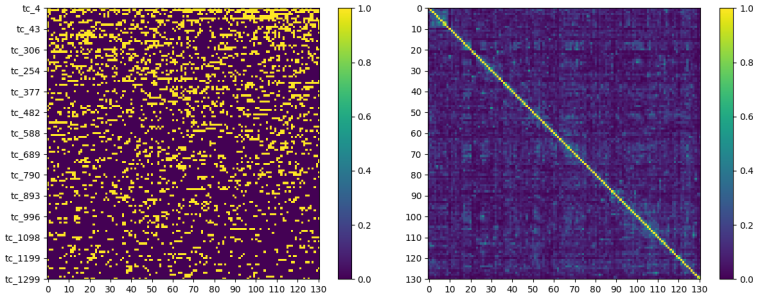

Figure NLP#7. Topic clusters over time are shown in the left, and the spearman scores between the user graph built based on topic similarities and the user graph built based on message propagation are shown in the right.

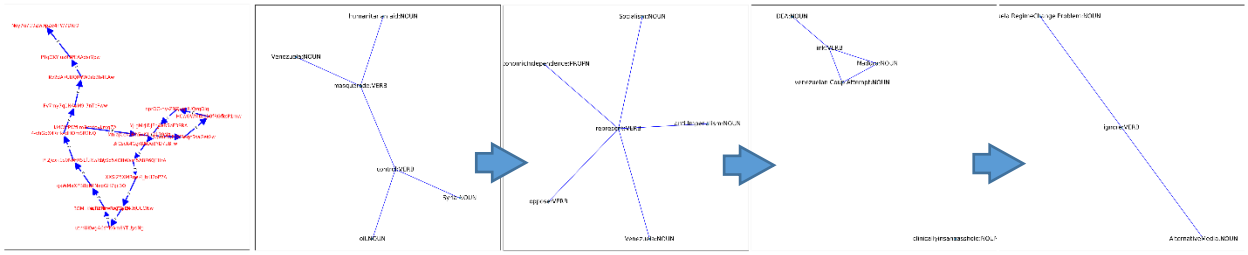

Figure NLP#11. A sample tweet response cascade and the first four core clause graphs along the cascade. The left most figure illustrates the cascade. The other four illustrate the core clause graphs along the cascade.

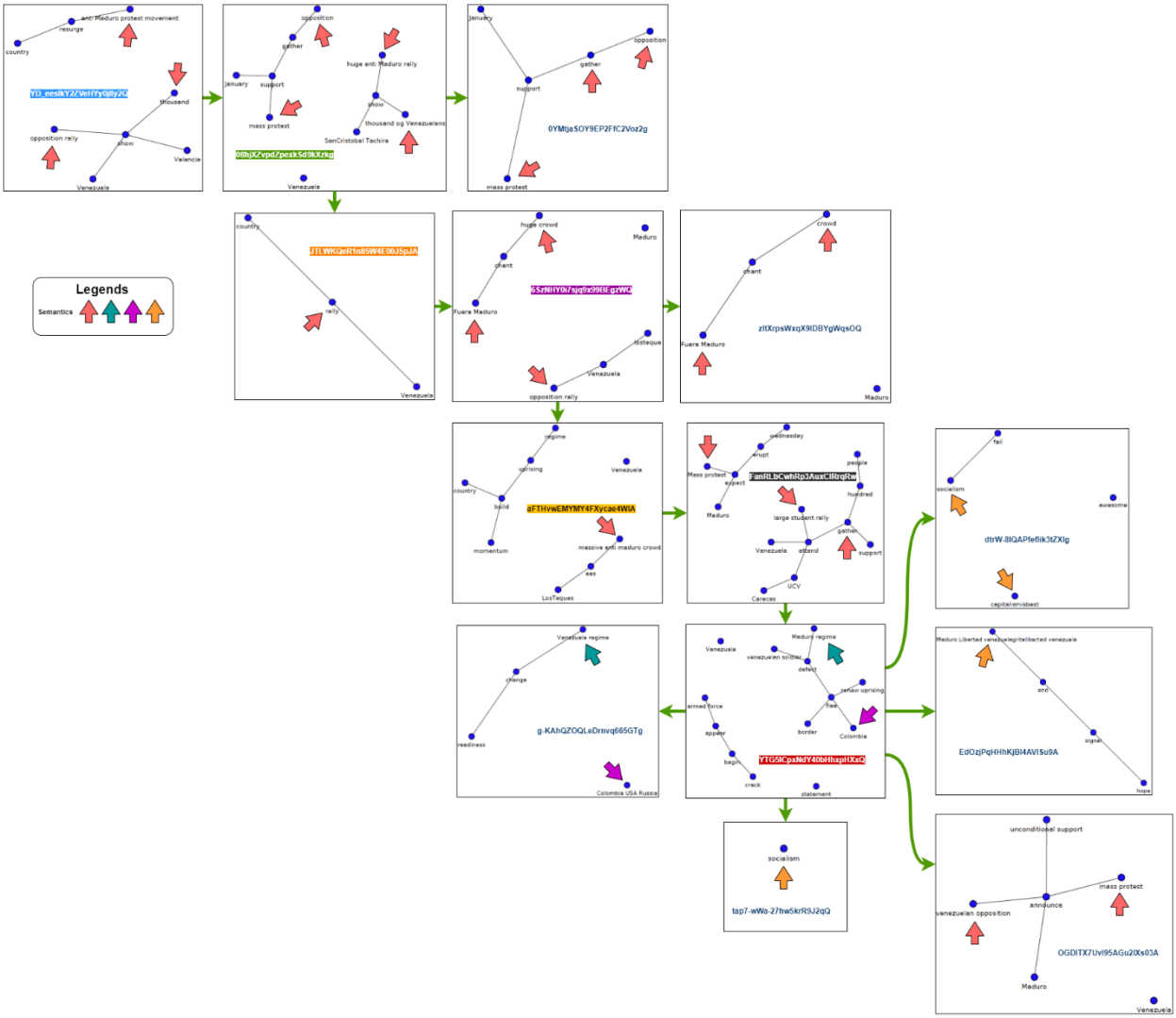

Figure NLP#12. A tweet response cascade in CP4 and the major semantic distribution over the cascade. The bold arrows inside the boxes point out the salient tokens with respect to a semantics. Each color indicates a different semantics.

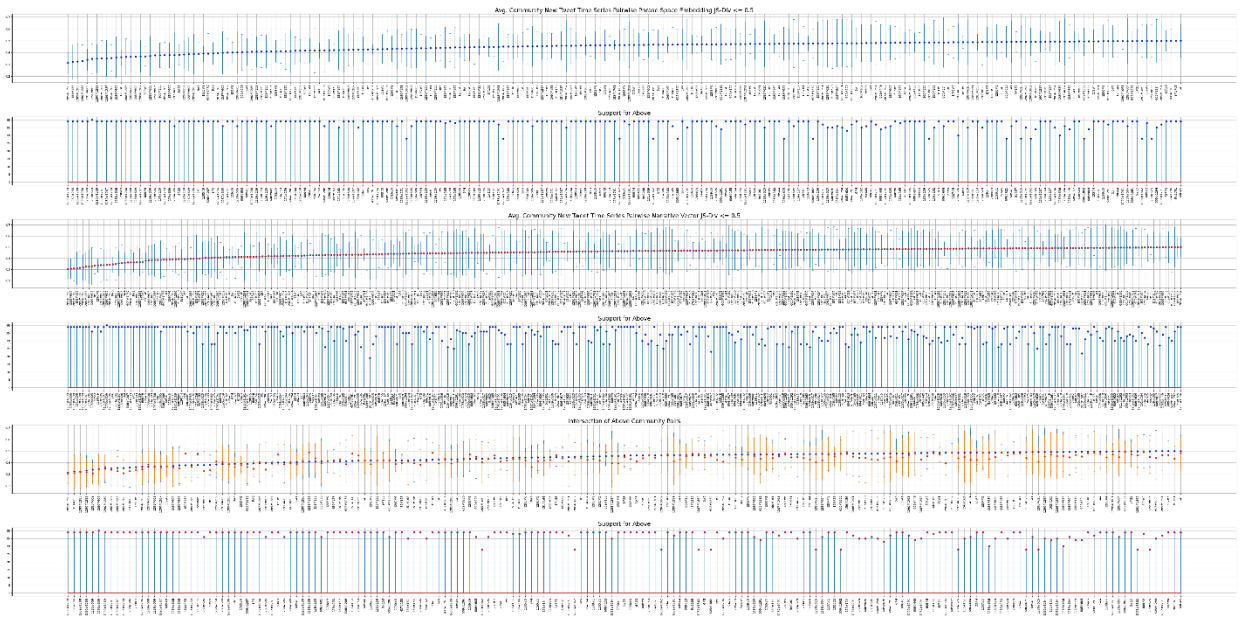

Figure NLP#13. The JS-divergences over phrase-clustering-based embeddings and narrative vectors, and their alignment at the fifth row. The two results basically match to each other. Only top 200 communities are considered in this example. And only comparisons with JS-divergence less than 0.5 are shown. The community pairs shown above can be considered relatively similar. Furthermore, among them, a small set of such pairs are very similar.

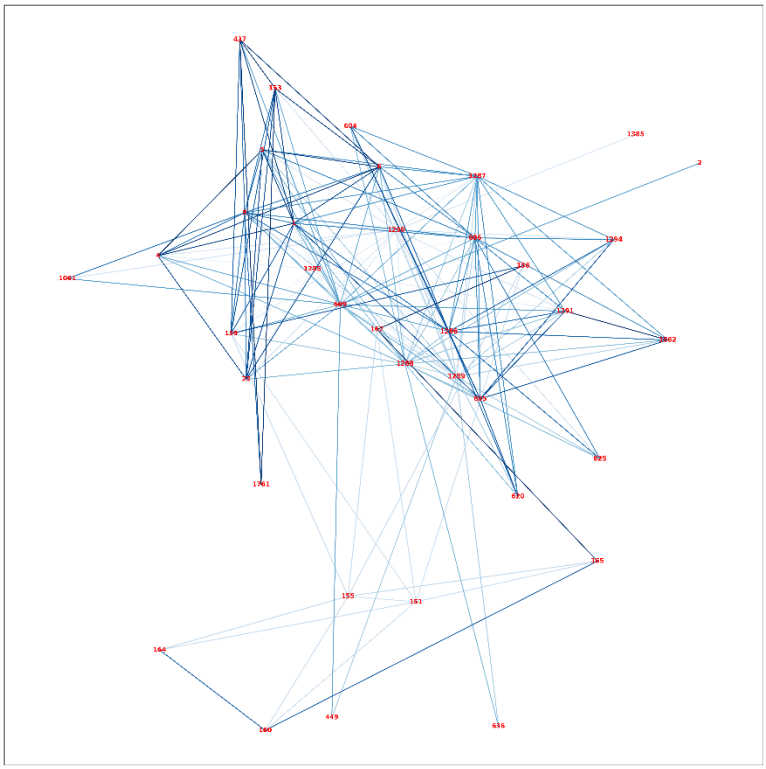

Figure NLP#14. The community similarity graph based on the phrase-clustering-based embeddings. Darker edges mean higher JS-divergence.

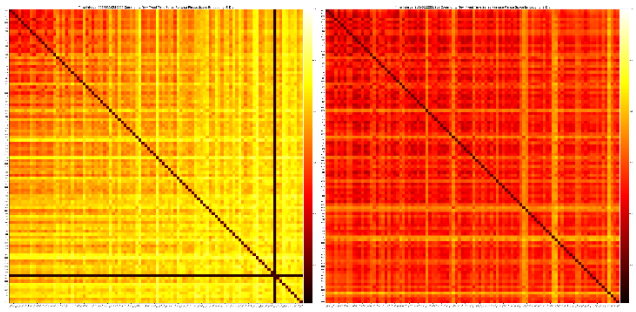

Figure NLP#15. Community JS-divergences in two time intervals. The samples are taken from the CP4 Venezuela Twitter data. The left figure shows the divergences in Jan.10 2019, and the right figure shows Jan. 22 2019. Both axes are communities (top 100). Red indicates similar, and yellow indicates dissimilar regarding semantics. On Jan.10 2019, Nicolás Maduro is inaugurated for his second presidential term by the Supreme Tribunal of Justice. [VP19] And on Jan. 22 2019, Protests throughout Caracas from the previous evening continue into the morning, resulting in the National Guard and National Police being deployed, with reports of tear gas being fired into streets and residential facilities. [VP19]

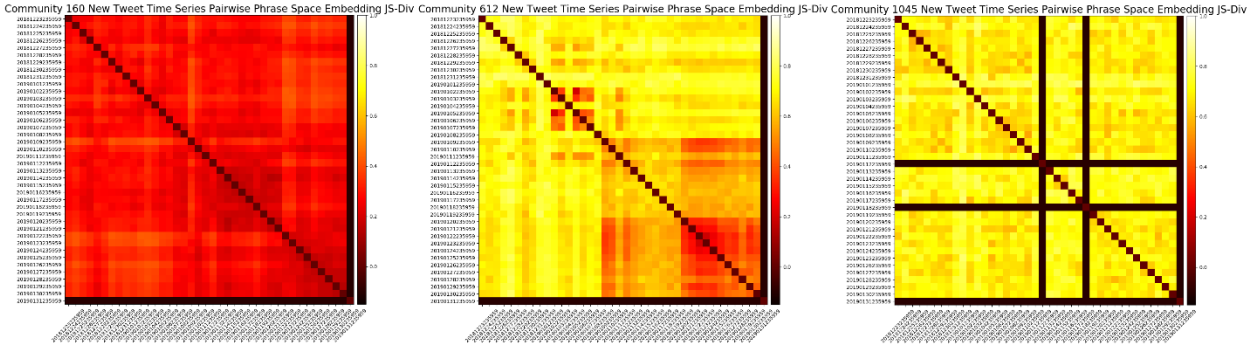

Figure NLP#16. Three communities and their semantic variation over time respectively. The three communities (all non-trivial) are taken from the CP4 Venezuela Twitter data. They are representatives of three distinct types of communities from our observations. And they exhibit nearly completely different behaviors over time. The leftmost one does not have large variation in its semantics over time. The one in the middle does not have large variation before Jan.10 2019, the day Nicolás Maduro is inaugurated for his second presidential term, and trends to have similar semantics thereafter. The rightmost one has constantly varying semantics the whole time regardless of the external events.

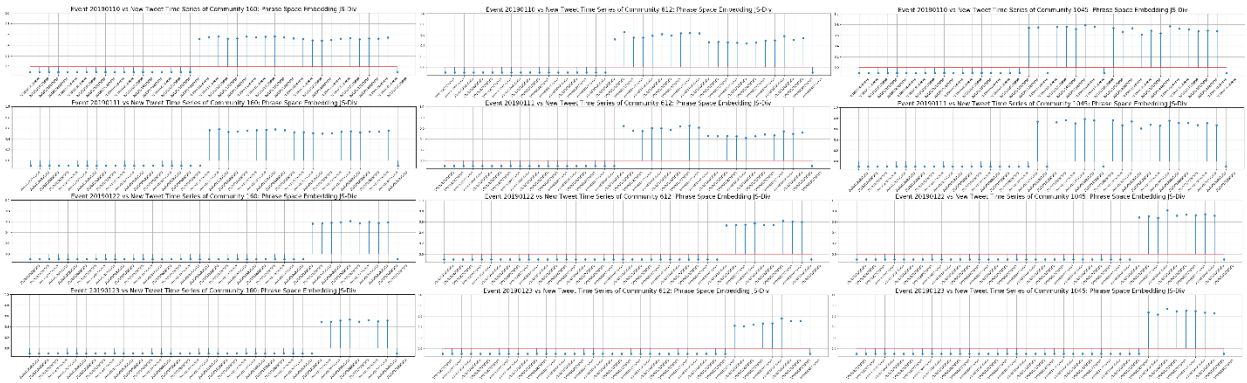

Figure NLP#17. The alignment between communities and external events. The three representative communities shown in **Figure NLP#16** are used again in this demonstration. Each row in the figure shows the JS-divergence between the semantics of the primary external events (relevant to the Venezuela protests) in a time interval and the semantics of the community in the same time interval and after that. Four typical time intervals are taken: Jan.10, Jan.11, Jan.22 and Jan.23. The consecutive dates are used for better tracking on the responses from the social network. It can be observed that the leftmost one, with few semantic variation over time, regarding semantics, is relatively correlated to the external events, but does not have much perturbation of semantic variation the whole time. The one in the middle is also correlated to the external events, and it does have explicit perturbation of semantic variation. The rightmost one is not related to the external events compared to the left and the middle, and it does not have explicit perturbation of semantic variation.

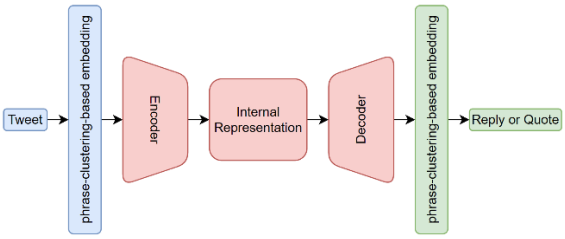

Figure NP#18. The Autoencoder architecture for reply-and-quote modeling. This architecture is simply the typical Autoencoder. In our experiments, the dimension of the interval representation can have a relatively wide range of settings, and typically in the range between 50 and 100 the Autoencoder would not have great behavioral change (when the input and output dimensions are 150). The activation function is SeLU. Other activation functions such as ReLU and ELU have also be tested, and they do not make great change to the model. The optimizer in our experiments is Adam, and the objective is the KL-divergence between the predicted output and the true response.

Reference:

[MF19] Meng, F. (2019). Document semantic representation: an algebraic topological approach (Doctoral dissertation, University of Delaware).

[MF20] Meng, F. (2020). A Topological Approach to Compare Document Semantics Based on a New Variant of Syntactic N-grams. arXiv preprint arXiv: 2103.05135.

[KM20] Kong, Y., Meng, F., & Carterette, B. (2020). A Topological Method for Comparing Document Semantics. arXiv preprint arXiv:2012.04203.

[SP13] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[VP19] https://en.wikipedia.org/wiki/Timeline_of_the_2019_Venezuelan_protests