The overall aim of this scientific research reading review is to aid predictive analytics project of developing social graphs and topic models. Hence, the research papers reviewed in this paper review are related to the either social graphs or ways to discover abstract topics that occur in a collection of posts/comments among different users.

Since most of the data required for topic modelling are generally present in short text formats (in terms of comments or short updates/posts), it is important to extract topics from these short texts while making sure the sparsity and dynamicity of these short texts are captured. To do so Graph-Based Dynamic Topic Models [1] is a single-pass graph-based topic model which addressed the scalability, dynamicity, and scalability of short texts to extract topics.

Another proposed model for topic modelling is User Graph Topic Model [4], it is more inclined on finding relationship between users given that most of the users are do not make multiple post and one post is made only by one user. Using hashtags, user mentions and comments the user graph is developed which displays the related user information.

Dynamic Exploratory Graph Analysis (DynEGA) [6] is a new approach for estimating latent topics in social media texts. It is an approach which deals with determining the latent topics from documents as well as automatically estimating the number of important simulated topics. The proposed algorithm was applied on large political tweet dataset and revealed some interesting topics that were important to several events in the political posts.

One of the main challenges to use GDTM [1] for topic modelling is that is tested and used only for short texts, it may not perform that well with the linguistic data. It does provide automatic feature extraction that can be represented in discrete level, but the execution time and coherence score might not be that great.

UGTM [4] works well in determining the semantic relationships between tweets, and outperforms several clustering models, but faces challenges while performing on smaller datasets. Algorithms like HGTM perform better as they produce better results when the data is rich in hashtags.

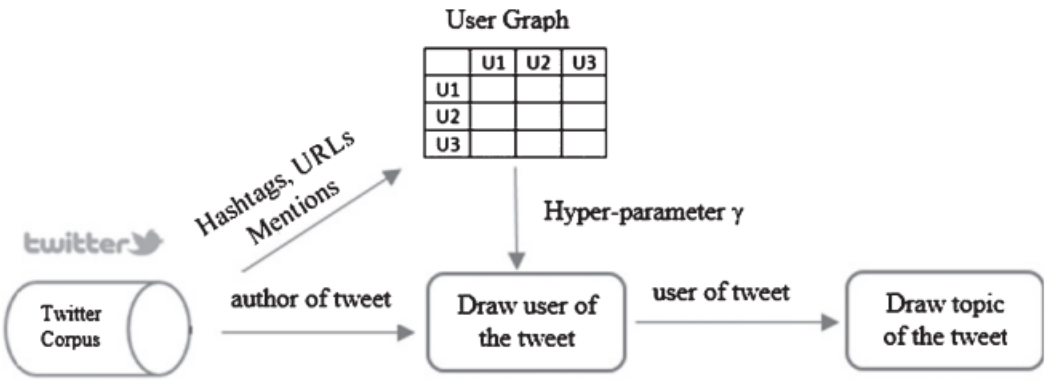

Fig1: UGTM model diagram [4]

While finding the simulated topics from any documents using topic modeling, we need to specify the number of simulated topics that we are interested in (for example in LDA the number of latent topics is specified by the user). One of the challenges that DynEGA [6] resolves is by producing the optimum number of simulated topics based on the document rather than being user defined.

As mentioned, the previous proposed models the existing work lacked in scalability, sparse and dynamicity. For example, the most common Latent Dirichlet Allocation algorithm [2] uses iterations hence the expectation maximization algorithm hinders the scalability. It also assumes that there are fixed number of topics which limits the adaption of the algorithm to dynamicity. Other research like BTM [3] address sparsity by using complex language models but uses a fixed number of partitions which again hinders the dynamicity. The proposed GDTM method [1], achieves great results on short texts by clearly distinguishing between feature representation and extraction of topics, this is done by using single-pass algorithms where each post passes only once to extract semantically rich and low dimension feature representational vectors of the documents passed. By doing this it makes sure their proposed algorithm does not rely on iterative optimization and variable partition numbers. Hence scalability and dynamicity issue are eliminated. Though it takes care of the scalability and dynamicity issue it is not guaranteed it will produce the best results or consistent results when used with the linguistic data or data with large texts.

	Sparsity	Scalability	Dynamicity
LDA	×	×	×
BTM	✓	×	×
GDTM	✓	✓	✓
CDTM	×	×	✓
PYPM	×	×	✓

Table1: Comparison between various algorithms and their features with respect to GDTM[1]

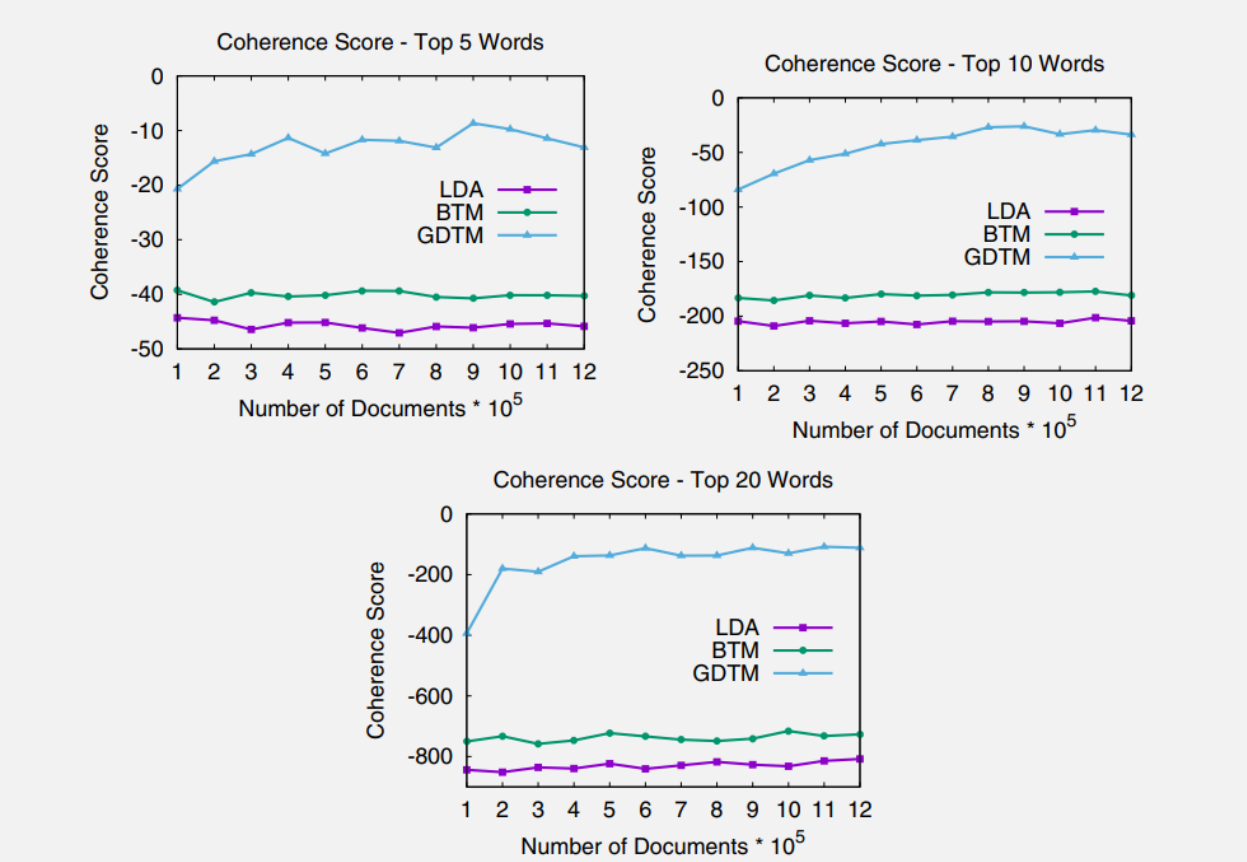

Fig2: Coherence score of GDTM with respect to LDA and BTM for different number of top words [1]

Other approaches for dealing with short texts/posts, use text extension by appending additional words from other sources (CoFE [5] or by aggregating the short texts to make a longer text to work on [6] refer 11. But these techniques are only useful when there is availability of contextual data from other sources or similar texts. Since tweets and replies usually consists of informal and sparse texts, UGTM [4] is better as it can model documents with only one author per document by obtaining Dirichlet hyper parameters from developing relationships between all the user posts. By doing so it outperforms the

traditional methods like ATM in tweet clustering (based on H-score evaluation) and topic coherence (PMI score evaluation) but fails to outperform HGTM as HGTM works better even on smaller datasets.

In the previous related studies (Ghanem et al. [7], Zannettou et al [8]) LDA [3] was used for quantitative analysis of texts to interpret the content of social media political troll posts/comments. Using LDA as the basis of topic modeling is well known but comes with few demerits. User needs to specify the number of latent topics, the interpretability is low as probabilities for every topic needs to be interpreted, and it goes with the assumption that topics are not short term and are uncorrelated. In DynEGA [6] study, a new method for estimating latent dimensions (e.g., factors, subjects) in multivariate time-series data. The DynEGA technique may be used to estimate the latent structure of subjects published on social media (using time series of word frequencies), allowing us to better understand the strategies employed by accounts intended as information warfare instruments. Unlike LDA, DynEGA can describe temporal dynamics in both stationary and non-stationary time series and can automatically identify the number of subjects and the distribution of variables (words) per topic (although the simulation only focused on stationary time series). It is possible to calculate network loadings and network scores for better interpretability of topic modelling. The output was also represented in mean topic scores for different topics.

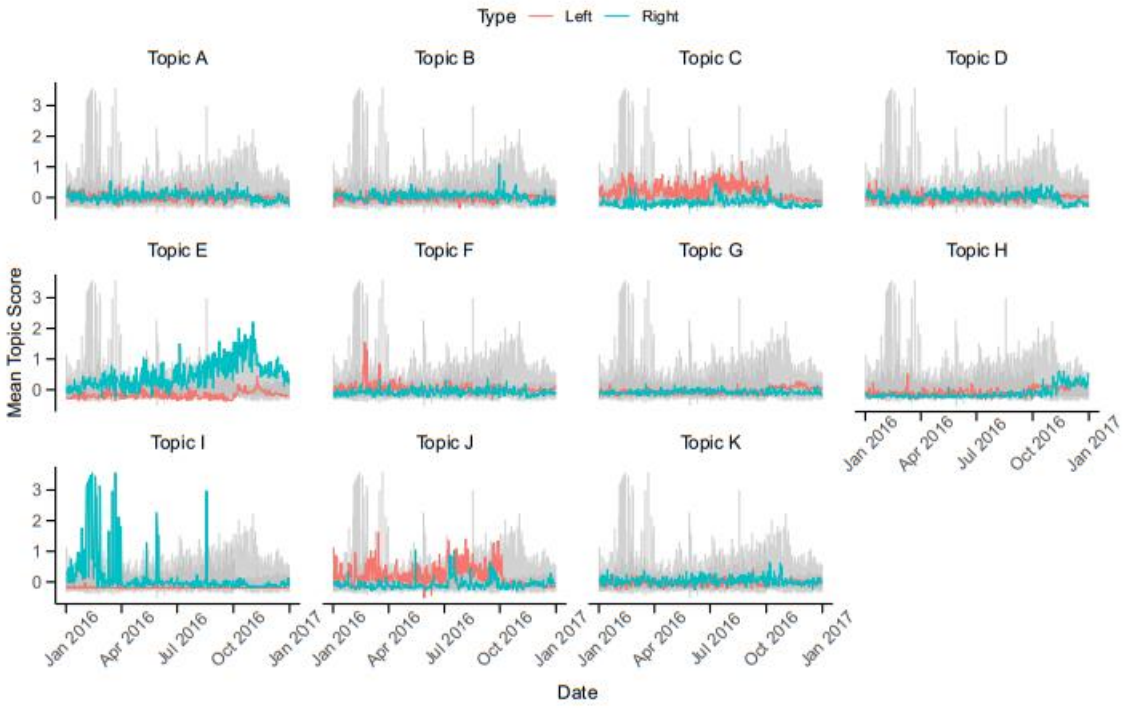

Fig3: Latent trends of the topics in time-series [6]

Since the GDTM [1] demonstrated to be a really good algorithm for topic modelling on large documents with short texts, it could be of great help to our project as Reddit posts data contains several posts with short word limits as well. By using GDTM we could accurately topic extract with vector feature representation with better scalability.

Our first requirement is to cluster similar posts on reddit, the proposed model UGTM [4] performs well with noisy short data as well in clustering as well as topic coherence evaluation. This could be one more algorithm that could be checked to see if it provides better clustering results on our Reddit dataset. Here we can also use H-score evaluation for testing the clustering and PMI score for evaluation the topic coherence.

Finally, we can also try DynEGA for social graphs as it also represents the second order derivatives from the most common topics from a document. In that way we would be able to see how not only how each topic are related, but also how are the most common words in those topics related to different words in other topics.

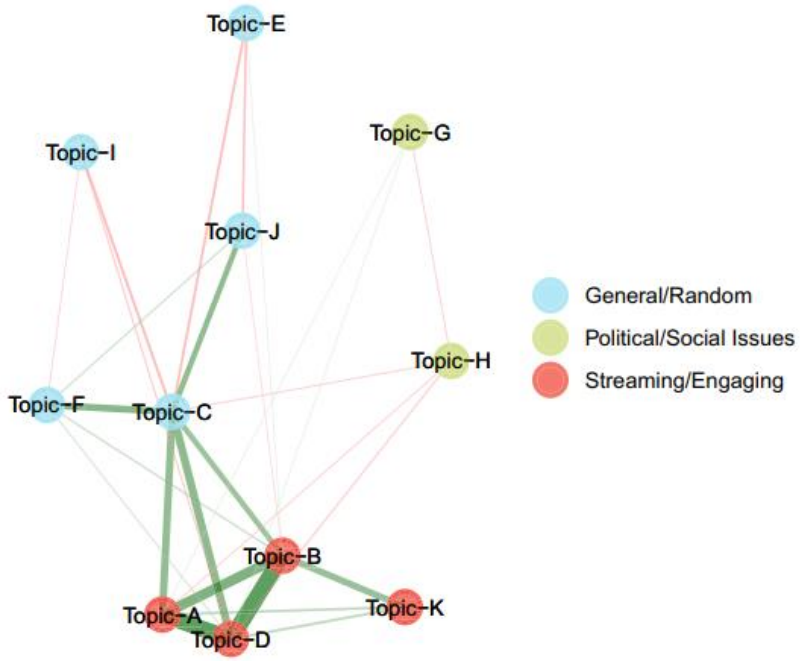

Fig4: Clusters of topics from political twitter dataset represented in network structure using DynEGA for first order topics. [6]

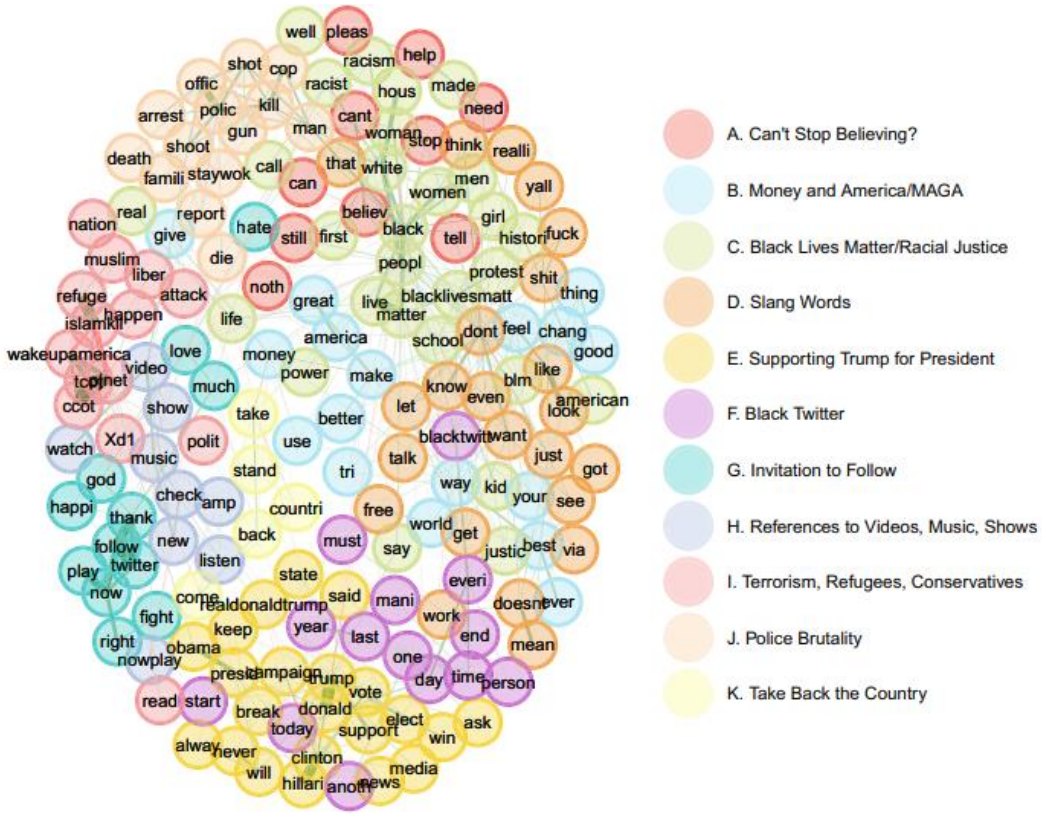

Fig5: Clusters of trolls from political twitter dataset represented in network structure using DynEGA for second order topics. Right side color index represents the topic heads, and left side shows nodes of all subtopics. [6]

References:

[1] Ghoorchian, K., Sahlgren, M. GDTM: Graph-based Dynamic Topic Models. Prog Artif Intell 9, 195–207 (2020). https://doi.org/10.1007/s13748-020-00206-2

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J.Mach. Learn. Res. 3, 993–1022 (2003)

[3] Yan, X., Guo, J., Lan, Y., Cheng, X.: A Biterm topic model for short texts. In:Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456, (2013)

[4] Akhtar, Nadeem and Beg, M.M. Sufyan. ‘User Graph Topic Model’. 1 Jan. 2019: 2229 – 2240.

[5] G. Pedrosa, et al., Topic modeling for short texts with cooccurrence frequency-based expansion, Intelligent Systems (BRACIS) 2016 (2016), 5

[6] Golino H, Christensen AP, Moulder R, Kim S, Boker SM. Modeling Latent Topics in Social Media using Dynamic Exploratory Graph Analysis: The Case of the Right-wing and Left-wing Trolls in the 2016 US Elections. Psychometrika. 2021 Nov 10. doi: 10.1007/s11336-021-09820-y. Epub ahead of print. PMID: 34757581.

[7] Ghanem, B., Buscaldi, D., & Rosso, P. (2019). TexTrolls: Identifying russian trolls on twitter from a textual perspective. arXiv, (1910.01340). Retrieved from arXiv:1910.01340

[8] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In Companion proceedings of the 2019 world wide web conference (pp. 218–226).