Examining Graph Attention Neural Network for Short-Text Topic Modelling using Document Neighbourhood Affinity: A Hydrogen Energy Discourse on Twitter

Si Man Kou

Abstract—Twitter is a well-known social media platform that allows users to express their thoughts on current events through micro-blogging, garnering the attention of text analysis practitioners in employing the topic modelling technique to unveil public perceptions. Despite its importance, the discourse around hydrogen energy on Twitter has remained largely unexplored. Furthermore, the exploration of high-order inter-document relationships within Graph Neural Networks, harnessing contextual tweet attributes beyond text content, to address the sparsity issue in text representation, has been rather limited. Motivated by these gaps, our study proposes a novel neighbourhood-assisted graph convolutional neural network topic model that capitalizes on the rich interaction features of Twitter data to capture high-order document-document, document-word, and word-word relations in topic learning. Moreover, it represents the first discourse study on Hydrogen Energy, examining the evolution of topics between 2013-2017 and 2018-2022. Experimental results with Hydrogen Energy data demonstrate that our method demonstrated certain limitations in generating interpretable and relevant topics compared to the benchmark model, NaNMF, despite that it produced a higher NPMI topic coherence, shedding light on the importance of using a larger dataset with reduced sparsity of word-document representations while choosing the appropriate topic coherence measure suitable to the dataset. The topic analysis underscores a notable shift towards global acceptance and recognition of Hydrogen Energy as a promising renewable energy source in the context of carbon reduction initiatives, highlighting the sustained interest in Hydrogen-powered technologies, particularly fuel cells and transportation. More importantly, the research has demonstrated great potential for future improvements with its novel approach in the fusion of document affinity into topic learning via attention network.

Index Terms—hydrogen energy, green energy, twitter data, topic modelling, graph convolutional network, document relationship

1 Introduction

As social media has entrenched itself as the mainstream platform for hosting a myriad of online interactions, Twitter stands out as one of the prominent avenues through which people express their thoughts towards current events and engage in interactive online social networking [1]. Given its wide adoption with over 300 million monthly active users globally and real-time opinion-sharing capabilities [2], researchers and text mining practitioners have shown keen interest in exploiting the technique of topic modelling to disseminate key themes, representable by a list of top keywords identified for each topic, from the pool of short-text tweets data, in turn contributing to the discovery of public opinions on previously unexplored subjects. From gaining insights into the public discussions on the COVID-19 pandemic outbreak [1] and global warming [3] to understanding the media framing of vehicles on Twitter [4] to collecting topics shared on Twitter in the aftermath of terror attack [5], topic modelling with tweets has continued to yield interesting insights for shaping decision-making across a wide range of domains.

As global energy demands surge in parallel with unprecedented population growth, the imperative for renewable energy is pressing as the shortcomings of fossil fuels and the abundance of sustainable energy garnered significant public attention [6]. Above all, hydrogen energy stands out as a promising sustainable energy source [7]. Hydrogen, the most abundant chemical element in the universe, is the foundation for generating emission-free energy, with applications ranging from fuelling vehicles and aviation to electricity storage [8]. Recent studies have examined public discourse on Twitter concerning sustainable energy sources using topic modelling, adopting an exploratory perspective to shed light on the role and implications for their acceptance and adoption [4], [9]. Given its importance for decarbonization, the specific discourse around Hydrogen Energy remains largely unexplored. A discourse on the evolving public perceptions towards Hydrogen Energy can therefore provide insights into their future role in the energy landscape and potential implications for acceptance and adoption [10].

Unlike traditional documents with lengthy content, extracting meaningful topics from Twitter is challenging owing to the short, unstructured, and informal nature of the text, given the character constraint posed. It creates the well-known sparsity problem, describing when there are limited word co-occurrences derived from extremely sparse text representation, for many state-of-the-art topic models reliant on the frequency of overlapping terms [11]. Contemporary topic modelling studies have emerged to develop methods targeted at microblogging environment retrieval with techniques like

Corresponding author: Si Man Kou (email: siman.kou@connect.qut.edu.au).

document pooling for tweet content augmentation and incorporating external semantic resources [12], [2], [13]. However, these techniques are heavily based on content which can still be prone to suffer from sparsity problem, consequentially leading to poor characterization of topics. Several studies in the literature have gone beyond content to integrate tweet relationships into topic modelling for assistance-based learning from document neighbourhood. Athukorala and Mohotti [14] and Nugroho et al. [2] both coupled document affinity matrix with non-negative matrix factorization, with the former using Jaccard similarity and the latter going beyond content to consider both tweets interaction and content for finding document links.

With the rise of deep learning, the notion of graph neural networks (GNNs) has been employed by many works in topic modelling to move away from treating documents uniformly to drive deeper collaborative learning of hidden document semantics using knowledge graphs of words and/or documents [15], [16], [17], [18]. The use of convolutional layers in Graph Convolutional Networks (GCNs) facilitates the learning of deeper and indirect relationships within the document corpus through information propagation. However, when representing networked documents, researchers often treat documents as connected or not during graph construction, negating the degree of proximity among documents [18], [19]. But in practice, it is intuitive to consider tweets with higher affinity to bear more similar topics than the others. Leveraging the connections between neighbouring documents could be a potential strategy for short-text topic modelling as neighbouring documents can provide additional information to compensate for the limited content in the host tweet.

Motivated by the fact that there has been a scarce focus on exploiting high-order inter-document relationships in Graph Neural Networks using contextual tweet attributes beyond textual content, this paper aims to investigate how enhanced document relations can be integrated into the attention mechanism used in many GNNs to learn the relative importance of neighbour links [20], and its effectiveness in short-text topic modelling in Twitter, in conjecture that integrating richer document similarity for graph topic model will enhance topic quality. Finally, this paper, to the best of our knowledge, is the first Hydrogen Energy discourse study to characterize the most salient topics discussed on Twitter around and topic shifts comparing 2013-2017 and 2018-2022.

We ask three research questions: (1) How can social features of tweets be integrated into document neighbourhood representation in Graph Neural Network to facilitate topic derivation? (2) How effective is the proposed method in capturing distinct topic semantics in Hydrogen Energy dataset compared to other benchmarks? (3) What are the key topics discussed and shifts in Twitter regarding Hydrogen Energy between 2013-2017and 2018-2022?

The rest of the paper is organised as follows. Section 2 covers related work on topic modeling. Section 3 details the proposed model and experimental procedures. Section 4 includes empirical study and benchmarks on Hydrogen Energy dataset for topic modelling and the topic analysis. Finally, Section 5 concludes with closing remarks.

2 Literature Review

In pursuit of our paper's goal to enhance GCNs for short-text topic modeling using rich semantic tweet features, this section will encompass three literature themes: 2.1 development of short-text topic modeling techniques; 2.2 studies focusing on using document relationships for topic model with a focus on how tweets interactions can be exploited for a comprehensive evaluation of connections, and 2.3 topic modeling with graph neural networks.

2.1 Short-Text Topic Modelling

Topic modelling has been a successful technique for text analysis for nearly two decades and remains a subject of ongoing scholarly attention due to knowledge acquisition in unstructured textual data [17]. However, inherent to short text is the sparsity problem which heavily penalized the topic interpretability and coherence in early techniques reliant on word co-occurrences such as LDA [21] and NMF [22], primarily due to the constrained word count within each document.

Recently, much of the focus of contemporary topic modelling research has been on tackling the sparsity issue by extending popular topic models [21], [22] with various document augmentation techniques [2], [11]. A study exploited external contextual assistance such as Wikipedia to extend tweet content; however, relying on external resources can add noise to the original document and depend on the availability of auxiliary data [13]. Whilst researchers have exploited the intuitive pooling technique to merge documents with shared links like hashtags or authors, Albanese and Feuerstein [12] recently proposed a community-based pooling technique to aggregate document contents within the shared retweet network to create a denser word co-occurrence matrix. Despite considering social tweet features, the established document relationships were solely used to improve the training corpus without further exploitation in the topic sampling space. Moreover, these pooling techniques often assume that linked documents inherently share a single topic, limiting document-level topic derivation.

2.2 Topic Modelling with Document Relations

Other short-text extensions considered integrating tweets' neighbourhood information to enhance topic quality. Athukorala and Mohotti [14] captured document affinity by Jaccard similarity to assist information loss resulting from dimensionality reduction in NMF, but only considered lexical similarity by word co-occurrences. Another recent study introduced a document relation matrix into NMF by incorporating content, social and temporal features to overcome extreme sparsity in short text and dynamics of the social media environment and showed the outperformance of the proposed method on tweet data compared to other methods using few to no contextual features beyond content [23], [2]. Despite the novel ideas, a common limitation is that only first-order relation is considered, but in Twitter networks, unconnected tweets can also be topically related. Inspired by Nugroho et al. [23], this study incorporates document interactions like replies, retweets, and mentions in computing document similarity, thereby exploiting the rich semantic features available on Twitter which have been overlooked in many studies.

2.3 Graph Convolutional Networks for Topic Modelling

The recent advances in Graph Neural Network have drawn an emerging research direction towards Graph Convolutional Networks (GCNs) for topic modelling, harnessing semantic relation graphs and message passing to capture richer contextual and semantic document information [24]. GraphBTM [15] adopts a variational-based inference approach that uses GCN in the encoder for biterm graphs to encode high-order word co-relations and learns a decoder to reconstruct input graphs. Moreover, the stacking of GCN layers allowed for learning of information from its immediate and higher-order neighbourhoods. However, the initial merging of sampled documents constrained topic generation to a mini-corpus.

Other studies integrated attention mechanisms into GCNs to address the fixed propagation weight inherent in most GCNs. Yang et al. [20] constructed a document-word bipartite graph for GCN by introducing a weighted (self-attention) graph convolution operation to learn propagation attributions for each connection, while Zhang and Lauw [19] learned topics from networked documents via neighbour competition and reconstruction, allowing documents to share similar representations in topic space with their close neighbours. Xie et al. [25] proposed a novel high-order graph attention topic model to efficiently explore correlations among networked documents by using the shortest distance between two nodes as input into the graph attention network. These studies did not take into account all document-document, document-word, and word-word relations in a unified framework of topic and graph modelling.

On the other hand, Xie et al. [18] proposed the first unified graph topic model, GTNN combining all document-document, document-word and word-word relations as knowledge graphs. It introduces inter-domain (document-word) and intra-domain (document-document, word-word) message passing to allow high-order semantic learning between words and documents. However, this study only gives equal contributions to neighbouring nodes in message propagation, negating the notion that documents of higher proximity should contribute more. Nevertheless, it failed to capture the degree of connection within documents as the dataset pertains to a citation network in which two documents are connected depending on a reference link, motivating our study to extend this method to the Twitter context.

Due to a lack of focus in the Twitter context observed in this line of research, there is a motivation to integrate the tweets' deeper semantic relations as neighbourhood proximity in the inputs for the GCN. Our proposed method, **G**raph **A**ttention **T**opic **N**eural **N**etwork (**GATNN**), builds upon the work of GTNN to exploit document-document, document-word and word-word relations with the introduction of the novel attention mechanism for the dynamic propagation of information amongst neighbourhoods. Moreover, it differentiates itself from other works by factoring tweet interactions and content for document proximity calculation for fusion into GCN for short-text topic derivation in the Twitter context.

3 Proposed Experimental Methodology

The proposed experimental methodology is outlined in Fig. 1 which graphically explains the experimental flow for this study, from data collection, data preparation, model input preparation, model application, model evaluation against baseline models using both an empirical and qualitative assessment of topic quality, to topic exploratory analysis.

3.1 Hydrogen Energy Dataset

One primary dataset is used for evaluation experiments and topic exploratory analysis and was collected via the Academic and Research Application Programming Interface (API). Since the objective of the research is to reveal the historical landscapes of Hydrogen Energy, a decade's worth of data from 2013 to 2022 amassing 30 million tweets were collected for the analysis using the Hydrogen Energy related keywords. The dataset, however, does not contain any ground truth labels on the associated topics. Along with the body of each tweet (i.e. tweet content), other semantic features were also employed for topic modelling for the derivation of document neighbourhood affinity based on the tweets' interactions and contents, which are summarized in Table 1.

3.2 Relevance Annotation

To address the issue of noise and unrelated content in the original dataset, a random sampling approach was further used to select 5000 tweets from the corpus for further evaluation by human annotators to eliminate noise. Although this reduced the overall dataset size, it allowed for a more direct examination of the impact on the topic quality extracted by the method, aiding the generation of interpretable and pertinent topics for the exploratory analysis.

The tweet annotation task entails a manual review of the content of 5000 English tweets, conducted by three annotators in accordance with annotation guidelines. These guidelines provide clear instructions on how to objectively assess the entire tweet content to minimise the rating biases, offering examples of relevant and irrelevant topics (e.g., fuel cells and hydrogen transportation considered relevant, while hydrogen bombs and hydrogen chemicals considered irrelevant).

Thereafter, all three annotated datasets were concatenated to perform inter-annotator agreement evaluation. As opposed to Cohen's kappa statistic which is used for two raters, Fleiss's Kappa agreement score was adopted and implemented using 'statsmodels' [1] package as it can assess the inter-rater reliability for categorical ratings given by any number of raters [26]. Overall, there is a substantial agreement with a Fleiss's Kappa score of 0.85 for the annotation task, with 88.86% (4443/5000 tweets, 2442 relevant, 2001 irrelevant) agreed across all annotators and the remaining 413 and 144 tweets marked as relevant and irrelevant respectively by the majority. However, to maximise the size of the final dataset for this study, the 557 discrepant tweets were reannotated, giving a final dataset of 2736 relevant tweets for topic modelling and analysis.

3.3 Tweet Content Preprocessing

Since the original tweet is represented in the raw JSON format, before text preprocessing, a separate procedure was needed to extract and map the required tweet features to the CSV format. After then, any tweet duplicates with identical

1. https://github.com/statsmodels/statsmodels/blob/main/statsmodels/stats/inter_rater.py

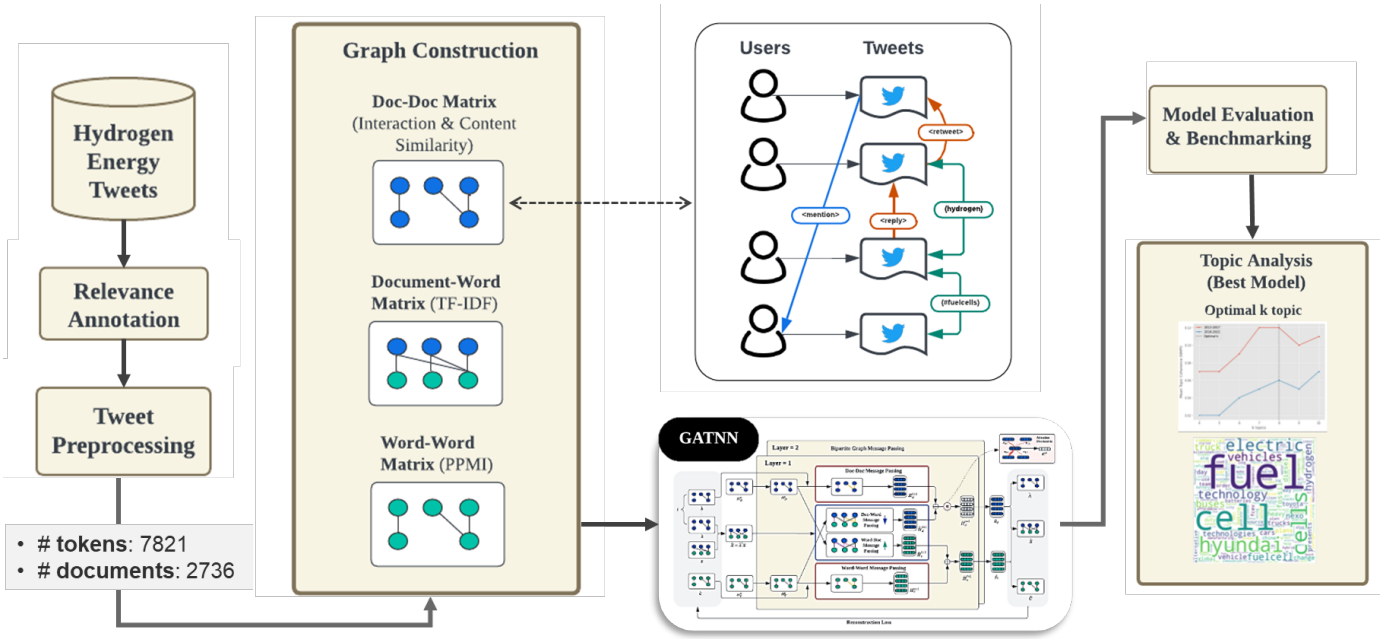


Figure 1: Experimental Workflow

Feature Name	Description
tweet id	ID of a tweet
tweet type	'Original', 'Reply', 'Retweet'
original tweet id	0 if tweet type belongs to 'Original' otherwise it contains the id of the tweet it retweeted or replied to
mentions	a list of mentioned usernames
username	the name of the user of the tweet

Table 1: Tweet Feature Columns

tweet IDs and text body were removed, with retweets being an exception to this rule, as they are retained due to their characteristic 'RT' prefix, distinguishing them from original tweets. Additionally, any non-English tweets are filtered out using language tags before the random sampling process. The tweet content is then subjected to a sanitization process, which involves the removal of usernames, special characters, punctuation marks, single quotes, digits, hyperlinks, emojis, emoticons, words with a length of less than three characters (except for 'ev', short for electric vehicles), excessive spaces and newline characters, non-English words, and stopwords derived from combined lists sourced from NLTK [2] and WordCloud [3]. The sanitized tweet messages are subsequently converted to lowercase and tokenized, resulting in a vocabulary consisting of 7821 distinct tokens.

Despite its popularity in topic modelling studies, stemming was not adopted in this study based on the research findings that stemmers do not improve topic coherence and can instead hinder interpretation [27]. Stemmers failed to consider the semantic meaning of words (e.g. 'Apple' vs 'Apples') by assuming the same stems should necessarily pertain to the same topic, disregarding the genuine semantic relationship [11], [27].

2. https://www.nltk.org/search.html?q=stopwords
3. https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html

3.4 Measuring Tweet Similarity

Beyond the content of the tweets, our study exploits various social interaction features to consider both lexical content similarity and interaction similarity in measuring the similarity of tweets, based on the similarity definitions put forth by Nugroho et. al. [28].

To illustrate the type of tweet interactions, Figure 2 pictorially demonstrates how two tweets can be connected through shared interactions, such as replies, retweets, and mentions (depicted by the red and blue arrows), as well as shared word occurrences, including hashtags and the content within the tweet message (indicated by the green arrows). For example, even though t1 and t3 do not share any common words or hashtags, they are assumed to demonstrate some degree of connection due to their shared mention of the same user, implying a potential association with a common topic. Similarly, t3 and t2 are closely related, not only because they discuss the same keyword, but also because t3 has interacted with t2 by replying. The ability to capture both type of relationships also suggests that t3 could be more strongly related to t2 than a tweet that only replied to t2 without any shared terms. Moreover, since t2 is a retweet of t1, it is evident that they should be categorized under the same topic. And the relationship between t3 and t4 shows a pure content-based relationship as both tweets capture the same hashtag (#fuelcells).

According to the definition by Nugroho et. al. [28], a tweet can be measured based on three components – (1) people-

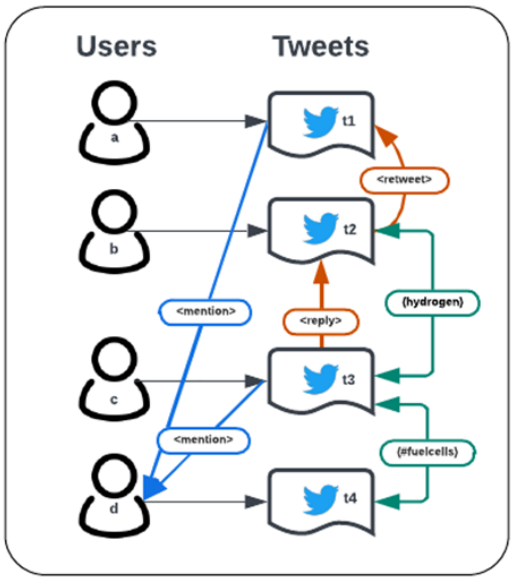


Figure 2: Characterization of Tweet Document Similarities

based interaction determined by the mention of users indicated by blue arrows in Fig. 2, (2) action-based interactions determined by retweet and reply activities indicated by the red arrows, and (3) tweet content indicated by the green arrows.

Mathematically, a tweet t can be defined as a tuple of $< P_t, RTP_t, C_t >$, where P_t denotes the set of people mentioned in the tweet, RTP_t represents the existing reply or retweet interaction with another tweet, and C_t represents the tweet content. The overall similarity can be formulated as Eq. 1. Generally speaking, a $R(t_i, t_j)$ of 0 means there exists no interaction between two tweets, which are the same as *self-contained* tweets.

$$R(t_i, t_j) = po(P_{t_i}, P_{t_j}) + act(RTP_{t_i}, RTP_{t_j}) + sim(C_{t_i}, C_{t_j}) \quad (1)$$

$po(P_{t_i}, P_{t_j})$ represents the people interactions and is defined as the number of common mentioned people between tweets i and j, normalised by the union of common people shared between both tweets (Eq. 2).

$$po(P_{t_i}, P_{t_j}) = \frac{|P_{t_i} \cap P_{t_j}|}{|P_{t_i} \cup P_{t_j}|} \quad (2)$$

$act(RTP_{t_i}, RTP_{t_j})$ quantifies the action-based interactions based on the retweet and reply activities, as defined in Eq.3. It will equal 1 if it meets any of the conditions: (1) tweet i retweeted or replied to tweet j, (2) tweet j retweeted or replied to tweet i, or (3) both tweets commonly retweeted or replied to a third-party tweet. Otherwise, the tweets will be considered unrelated with a value of 0.

$$act(RTP_{t_i}, RTP_{t_j}) = \begin{cases} 1, & \text{if } (RTP_{t_i} = t_j) \\ & \text{or } (RTP_{t_j} = t_i) \\ & \text{or } (RTP_{t_i} = RTP_{t_j}) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Lastly, due to the dominance of self-contained tweets, tweet content will be factored for similarity calculation based on the term occurrences. Different from Nugroho et. al. [28] which only calculates the cosine similarity based on the term frequency, we employed three different similarity metrics for evaluation purposes on model performance, as defined below:

1) **Jaccard Similarity** (Eq. 4): evaluates the the number of overlapped terms amongst two tweets, as a ratio of the union of two term sets.

$$sim_{jaccard}(C_{t_i}, C_{t_j}) = \frac{|C_{t_i} \cap C_{t_j}|}{|C_{t_i} \cup C_{t_j}|} \quad (4)$$

2) **TF-Based Cosine similarity** (Eq. 5): measures the similarity between two document term count vectors C_i and C_j where each dimension in a vector corresponds to a unique term in the document collection with the value representing the term frequency in a document vector.

$$sim_{cosine\_tf}(C_{t_i}, C_{t_j}) = \frac{C_{t_i} \cdot C_{t_j}}{||C_{t_i}|| \cdot ||C_{t_j}||} \quad (5)$$

3) **TF-IDF-Based Cosine similarity** (Eq. 6): instead of using term frequency, this metric applies a TF-IDF weighting scheme to represent each dimension in the tweet vectors. Term Frequency-Inverse Document Frequency (TF-IDF) measures the term importance not just based on the occurrences in a document but also its rarity across the document corpus, weighting higher on terms that are uniquely meaningful to the documents with less appearance over the document corpus (document frequency).

$$sim_{cosine\_tfidf}(C_{t_i}, C_{t_j}) = \frac{C_{t_i} \cdot C_{t_j}}{||C_{t_i}|| \cdot ||C_{t_j}||} \quad (6)$$

Finally, a sigmoid transformation was applied over $R(t_i, t_j)$ to standardize the score between 0 to 1, with 0 indicating absence of document relation and 1 as perfect correlation.

3.5 Graph Construction

Before applying **GTNN** and **GATNN**, graph construction was a crucial step in representing document-to-document, document-to-word and word-to-word connections for neighbourhood learning. We define D as a collection of N (2736) documents, a vocabulary, V, of size M (7821) tokens.

The normalised pairwise tweet correlation derived from Section 3.4 forms the document affinity matrix S which will only be employed by GATNN as it fuses the connection weight into the attention mechanism. A cut-off threshold was applied on top of the weighted matrix to derive a binary adjacency matrix A, of size N x N, as inputs into both models to capture the existence of a document relation.

The document-to-word matrix is denoted as matrix X of size N x M where each value in the matrix represents the TF-IDF of a word in the document. It can alternatively be represented as a bipartite graph with connections only passing between nodes of two different domains (word and document).

Finally, the word-word matrix, C of size M x M, represents a weighted symmetrical matrix with each edge representing the Positive Pointwise Mutual Information (PPMI), which measures the likelihood of co-occurrence of two words given the prior of the individual word.

3.6 GATNN

The proposed GATNN extends the existing GTNN [18] and implements an attention mechanism that fuses the document affinity matrix S, as visualised in Figure 3.

3.6.1 GTNN Backbone

Before applying the GCN layer, a multi-hop (l-lop) diffusion process on binary document network A was used as a low-pass filter $\hat{A}^l$ on the document content X to preserve the low-frequency components. This operation is equivalent to applying Laplacian smoothing to X to make the node feature more indistinguishable leaving just the features of interest, as defined by the equation:

$$\hat{X} = \hat{A}^l X \tag{7}$$

where $\hat{A} = D^{-0.5}(A+I)D^{-0.5}$ is a symmetrically normalised matrix with self-added connections, I is identify matrix of A, D is the degree matrix of A. The l-hop diffusion across X and A essentially facilitates capturing indirect doc-doc and doc-word correlations, allowing the network to learn similar latent representations for documents sharing common words or connections. The 'filtered' networks $\hat{A}$, $\hat{X}$, along with the normalised matrix $\hat{C}$, were then passed into the Graph message-passing mechanism for a unified topic and graph learning, which ultimately projects the learnt document and word representation in the topic space as $\theta \in \mathbb{R}^{N\times K}$, defined as the document-topic matrix, and $\beta \in \mathbb{R}^{M\times K}$, defined as the word-topic matrix.

An integral part of the GTNN is the Bipartite graph message passing (BGMP), which facilitates information propagation through the immediate and higher-order neighbours. There are two types of learning. First is the learning amongst nodes of the same nature, known as intra-domain message passing (IAMP). The hidden representations of word node v and document d after one GCN layer are computed as follows:

$$\begin{aligned} H_v^t &= \sigma(\sum_{v'\in N(v)} \hat{C}_{v,v'} H_{v'}^{t-1} W_v^{t-1}) \\ H_d^t &= \sigma(\sum_{d'\in N(d)} \hat{A}_{d,d'} H_{d'}^{t-1} W_d^{t-1}) \end{aligned} \tag{8}$$

where $N(v)$ is the set of word neighbours of word node v, and $N(d)$ is the set of document neighbours of node d, W_*^{t-1} is the node-specific and layer-specific trainable weight matrix of size $*\times K$, σ as the nonlinear activation function and H as the hidden representation of document or word node at previous layer t-1.

Following IAMP is the inter-domain message passing (IDMP) proposed to explore nodes of different natures leveraging the indirect relationships between documents and words captured in the filtered document content $\hat{X}$ in two separate word-to-document and document-to-word message-passing processes defined as follows:

$$\begin{aligned} \hat{H}_v^t &= \sigma(\sum_{d\in N(v)} \hat{X}_{d,v}^T \hat{H}_d^{t-1} \hat{W}_v^{t-1}) \\ \hat{H}_d^t &= \sigma(\sum_{v\in N(d)} \hat{X}_{d,v} \hat{H}_v^{t-1} \hat{W}_d^{t-1}) \end{aligned} \tag{9}$$

The combination of Equations 8 and 9 forms the final hidden document $\tilde{H}_v^t = H_v^t + \hat{H}_v^t$ and word representation $\tilde{H}_d^t = H_d^t + \hat{H}_d^t$ at layer t, where the final layer representations are equivalent to the θ and β.

To preserve the underlying graph network structure in the projected representations, Xie et. al. [18] further proposed an online graph decoding process that reconstructs graphs C and A and the filtered documents $\hat{X}$ separately based on the learnt embeddings θ and β, with the objective of minimizing the total divergence between the original and reconstructed graph structures. Full details can be found in [18].

3.6.2 Attention Mechanism

Note that the backbone model does not leverage the precomputed document affinity matrix S, as it only captures the absence and presence of document connection with A where all neighbouring nodes, as long as they are connected, give equal attribution to the host. To factor in the degree of similarity for information propagation, an attention network is proposed to fine-tune the learned aggregated document embeddings $\tilde{H}_d^t$ using the importance between two nodes i and j defined as attention coefficient $\alpha_{i,j}$, as inspired by [25], as below:

$$\begin{aligned} e_{ij} &= relu(\phi(S_{ij})(\tilde{h}_{di}^t \cdot \tilde{h}_{dj}^{t\,T})) \\ \alpha_{i,j} &= softmax(e_{ij}) \end{aligned} \tag{10}$$

where $\tilde{h}_{di}^t$ and $\tilde{h}_{dj}^t$ are the learned 1-D hidden document embedding for a document node, $\phi(s_{ij}) = \frac{exp^{(-lns)^2/\sigma^2}}{s\sigma\sqrt{2\pi}}$ is the probability density function of the log-normal distribution, σ as the variance of log-normal prior, s_{ij} as the document connection weight between nodes i and j, relu as the ReLU activation function. The final attention coefficient will be used to update the $\tilde{H}_d^t$ based on the derived pairwise attention of the neighbours using the equation 11:

$$\ddot{h_{ij}} = \sigma(\sum_{d'\in N(d)} \alpha_{d,d'} \tilde{h}_{dj}^t) \tag{11}$$

as illustrated in pink parts in Fig. 3.

3.7 Model Evaluation

Two models, GTNN [18] (our model backbone) and NaNMF, a neighbourhood-assisted NMF using Jaccard Similarity, served as the baseline models for GATNN. Transductive learning was used for all models, meaning all samples are used for training, and an arbitrary number of 20 was used as the number of topics for this benchmarking study. For GATNN and GTNN, the degree (number of propagation layers) was configured to 2, whereby 50 was used consistently as the number of epochs across all the experiments. As mentioned, a further optimisation study was performed across the three content similarity metrics at different threshold values used to define

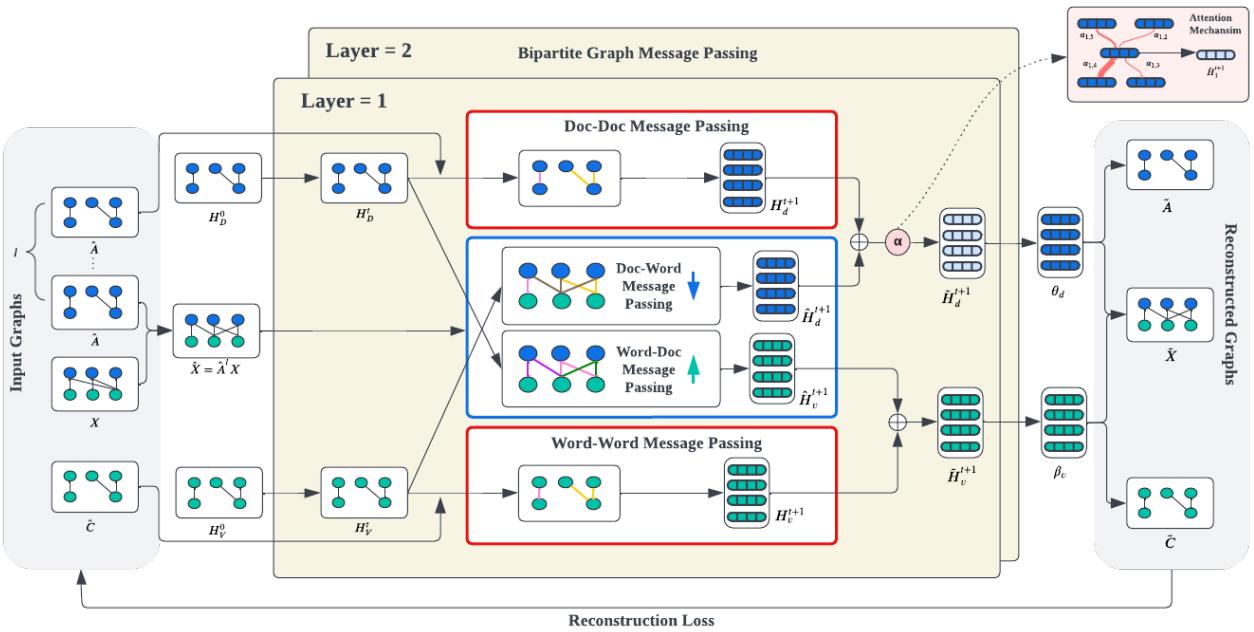


Figure 3: Model Architecture of GATNN - An extension of GTNN with attention mechanism using document similarity score

the adjacency matrix. To maintain consistency between matrices S and A, the same metric was applied to derive both matrices across all the experimental runs for GATNN. On the other hand, topics were learnt at default settings for NaNMF.

Three types of analysis were used to evaluate the topic learning effectiveness and topic quality empirically and qualitatively:

1) **Topic Coherence**: measured by Normalised Pointwise Mutual Information (NPMI), defined in Eq. 12, as it was proved to be the best-performing metric that demonstrated the most alignment to human judgement [29]. We calculated topic-level coherence which is the average pairwise NMPI of the top 10 words topic-wise, and each model was assessed on the average and median NPMI amongst the 20 topics.

$$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \qquad (12)$$

2) **Topic Interpretability**: this involved a human interpretation of 10 keywords extracted for each topic, rating on the coherence of words to each topic, the topic uniqueness, and the interpretability of keywords.
3) **Learned Document-Topic and Word-Topic Embeddings**: we visualised the projected word-topic and document-topic representations in a heatmap to understand the distinctiveness of topics by word and document compositions. Moreover, t-SNE dimensionality reduction reduces the projected document representations in 2D dimensional space to aid visual assessment of how capable the model is of distilling common patterns and topics amongst the corpus.

3.8 Qualitative Topic Analysis

Following model evaluations, the most effective model was employed to examine the topic. Word clouds were used for visualizing the relative distribution of words within each topic, emphasizing the words with the most significant contributions. Nevertheless, selecting an appropriate number of topics k relevant to the dataset is crucial. Opting for too few topics might lead to masking of distinct topics, while an excessive number of topics could destruct any cohesive topics. For this reason, an iterative process was adopted to assess the average topic coherence, using NPMI, for each k across a range of k values from 4 to 10, and the k with the highest average NPMI was selected. To periodically study the shifts and persistence in topics, the dataset was split into two subsets of 2013-2017 and 2018-2022, comprising 1163 and 1573 documents respectively, followed by the aforementioned data preparation steps.

4 Results and Discussion

4.1 Comparing Content Similarity Metrics and Adjacency Matrix Cut-off

Determining an optimal cutoff threshold is crucial to balance excessive connections and lack of connections. As evident in Fig 4, cut-off thresholds between 0.5 to 0.6 exibit a significant downturn in the resulting density in matrix A and have been used for evaluation against different content similarity measures to observe topic quality and model characteristics in the pursuit of model optimisation. Density statistics are shown in Table 3, encompassing thresholds within the optimisation range as well as 0 to observe the extremity where 90% of the matrix is connected, with Table 2 showing the corresponding mean and median NPMI and Fig 5 showing the resulting

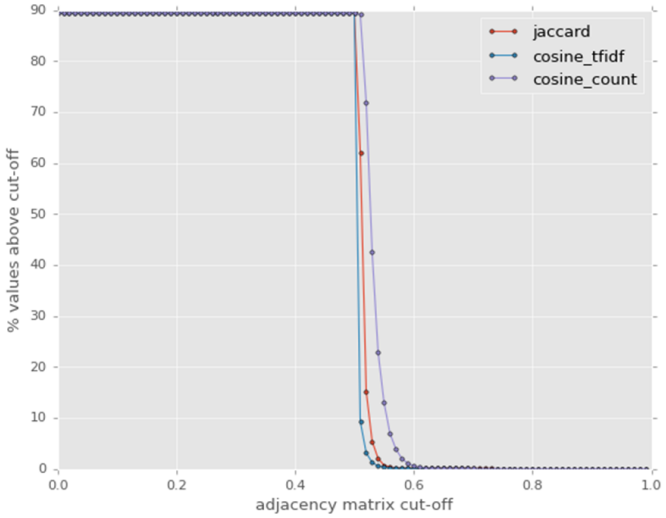


Figure 4: Comparison of Density of Adjacency Matrix A between Jaccard, Cosine with TF-IDF, and Cosine with TF as the content similarity measure across different cut-off thresholds.

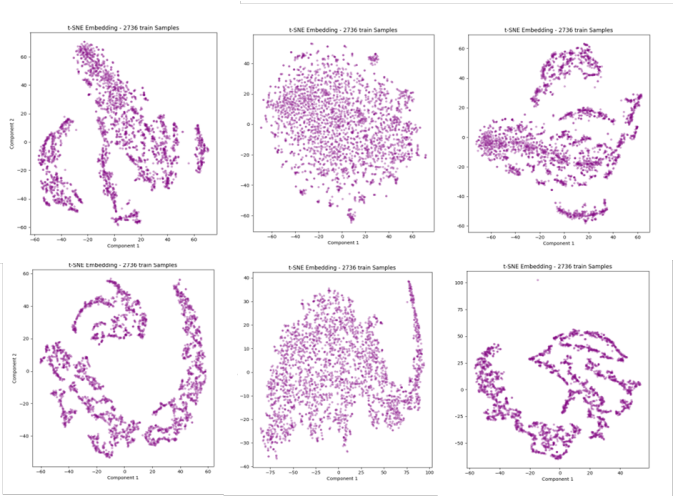


Figure 5: Comparison of document embeddings between GTNN (First row) and GATNN (Second row) at the optimal thresholds for each similarity metric (left: Jaccard, centre: Cosine TF-IDF, right: Cosine TF as found in Table 2).

document embeddings between GTNN and GATNN across different similarity measures.

Table 2 highlights consistent results across two models in terms of the respective optimal cut-off thresholds per measure - 0.52 for Jaccard, 0.53 for Cosine TF-IDF and 0.55 for Cosine TF, as evidenced by the highest mean and median NPMI. Comparing this against Table 3, we can further observe that generally for Jaccard and Cosine TF, the optimal cut-off points fall between the density range of 13-15%, in contrast to the low adjacency matrix density of 1.2% for Cosine TF-IDF. Overall, Cosine TF emerges as the best-performing metric considering both mean and median NPMI, at its respective optimal threshold of 0.55; hence, it has been chosen to represent GTNN and GATNN for model evaluation alongside NaNMF.

The low-dimensional TSNE document embeddings (5) further shed light on the difference between the dense adjacency matrix by Jaccard and Cosine TF and the sparse adjacency matrix by Cosine TF-IDF. It shows that the denser adjacency matrix results in distinct clusters with the final document-topic embeddings, suggesting the model's effectiveness in capturing document relationships in topic modelling so that connected documents are more likely to share similar topic representations than other distant nodes. In contrast, the low density of document affinity matrix using Cosine TF-IDF led to document embeddings that exhibit a more random distribution, evidenced by the absence of well-defined groupings. In the extreme case where the density is as high as 90%, the model suffered from poor topic coherence (see Table 2). This underscores the critical nature of threshold fine-tuning in our study, as opposed to the original study which adopted a deterministic approach based on the empirical evidence of reference links in the citation network.

4.2 Model Evaluations

4.2.1 Topic Coherence

Table 4 shows the overall NMPI for each model. Higher NMPI means higher topic coherence. It shows that GTNN emerges as the best-performing model with the highest mean and median NMPI of 0.20 and 0.22 respectively, followed by GATNN and NaNMF.

4.2.2 Word-Topic Representations

To complement our examination of topics beyond their coherence, we employed a qualitative assessment approach akin to [30] to gauge the distinctiveness of topics. This assessment explored how words were distributed within each of the 20 word embeddings produced by the three models as illustrated in Figure 6. It highlights a striking resemblance between GTNN and GATNN, which exhibit a more monotonous word distribution across different topics, with the majority of the word probabilities per topic gravitating towards the two extremes of -1 (dark brown) and 1 (dark green). In contrast, NaNMF shows a more distinct and discernible word pattern amongst the topics, with each topic comprising a rather distinct set of words. Additionally, the observation that most words have a probability of 0 suggests that the NaNMF can differentiate the important words for each topic, and hence, the ten most probable words are more representative of the topic than the other two models. Despite more discernible patterns observed in NaNMF, it is worth noting that there are still topics that failed to display any distinct words (e.g. topics 2, 6, 8, 13, 17) which indicate that the data might have fewer inherent topics than 20.

Figure 7 provides a detailed exploration of the word embeddings pertained to GATNN, specifically focusing on topics characterized by extremely high and low word probabilities along with the extracted topics. Generally, we observed that topics represented by equally high contribution from all words (topics 3 and 9) tend to generate more noisy topic words than topics at the low extremes (topics 12, 17) which appear to capture interpretable and relevant words, despite the presence

	Cut-off Thresholds	Jaccard		Cosine TF-IDF		Cosine TF	
		Mean NPMI	Median NPMI	Mean NPMI	Median NPMI	Mean NPMI	Median NPMI
GTNN	0	0.10	0.08	0.10	0.08	0.10	0.08
	0.51	0.10	0.09	0.09	0.08	0.10	0.08
	0.52	**0.21**	**0.13**	0.08	0.06	0.09	0.08
	0.53	0.15	0.13	**0.22**	**0.15**	0.09	0.08
	0.54	0.11	0.08	0.14	0.11	0.21	0.17
	0.55	0.10	0.08	0.11	0.10	**0.20**	**0.22**
	AVG	*0.13*	*0.10*	*0.12*	*0.10*	*0.13*	*0.12*
GATNN	0	0.09	0.04	0.09	0.04	0.09	0.04
	0.51	0.09	0.03	0.09	0.04	0.09	0.04
	0.52	**0.18**	**0.15**	0.08	0.04	0.09	0.03
	0.53	0.12	0.12	**0.17**	**0.13**	0.10	0.04
	0.54	0.11	0.10	0.10	0.10	0.19	0.15
	0.55	0.10	0.07	0.09	0.07	**0.19**	**0.16**
	AVG	*0.12*	*0.09*	*0.10*	*0.07*	*0.13*	*0.08*

Table 2: Comparison of Topic Coherence between GTNN and GATNN using different content similarity metrics and cut-off thresholds derived adjacency matrices.

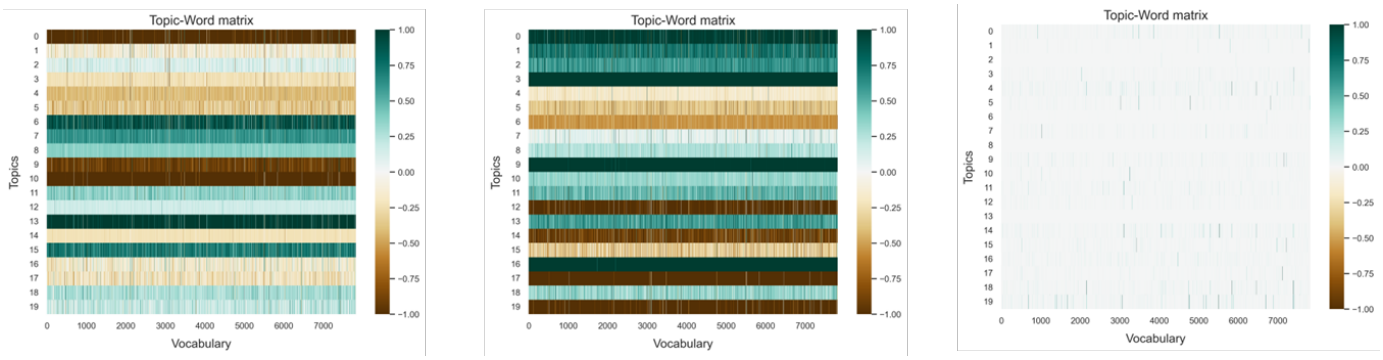


Figure 6: Heatmaps of learnt word-topic distributions by model (Left: GTNN, Centre: GATNN, Right: NaNMF

Cut-off Threshold	Jaccard	Cosine TF-IDF	Cosine TF
0	89.6	89.5	89.5
0.51	62.1	9.3	89.1
0.52	15.2	3.1	71.9
0.53	5.4	1.2	42.4
0.54	2.0	0.5	22.9
0.55	0.7	0.3	13.0

Table 3: Comparison of Density % of Adjacency Matrix A between content similarity measures at different cut-off thresholds.

Model	Mean NMPI	Median NMPI
GTNN	0.20	0.22
GATNN	0.19	0.16
NaNMF	0.08	0.08

Table 4: Mean and Median Topic Coherence NMPI per model

of frequently repeated common words like *green*, *water*, *power*, *energy*. This difference can be explained by more distinct words within topics 12 and 17, where certain words exhibit highly distinctive probabilities. In contrast, for topics 3 and 9, the models have learned that all words contribute equally, resulting in picking ten random words to represent the topic. It is also interestingly noted that the noisier topics exhibit higher coherence scores than the interpretable ones. This can be because common words can appear frequently in many documents and can co-occur with a variety of other words and hence are associated with diluted PMI values. Consequently, these common words might spread across various topics and lack strong association with a specific topic.

4.2.3 *Document-Topic Representations*

Figure 8 shows that all models could learn distinct document relationships and group related documents together. Comparing GTNN and GATNN, GATNN appeared to learn tighter embeddings, indicating that the additional fine-tuning through attention mechanisms allowed the model to identify deeper levels of document relations, resulting in more well-defined clusters of documents. Conversely, GTNN displays a bit more variance in the observed data points. In contrast, NaNMF learnt many distinct groups of documents at a finer granularity than the other two. Upon examining the topic-document heatmaps, it becomes evident that NaNMF shows a stronger ability to distinctly group documents by topics, except for topics 1, 2, 6, 8, 13 which show uniform patterns than the others. A similar conclusion for word-topic distribution for GTNN and GATNN can be made for document-topic, indicating a lower degree of distinctiveness in document-topic associations, despite the ability to cluster related documents.

4.2.4 *Topic Interpretability*

The table 5 presents the top 3 topics, each accompanied by the top 10 probable keywords based on the word-topic distribution, along with the topic coherence score for each topic. Although the table demonstrates that GTNN and GATNN achieve higher topic coherence scores compared to NaNMF, it is noteworthy that NaNMF yields topics that are more coherent and interpretable than those extracted by GTNN and GATNN, with the latter methods appearing to capture

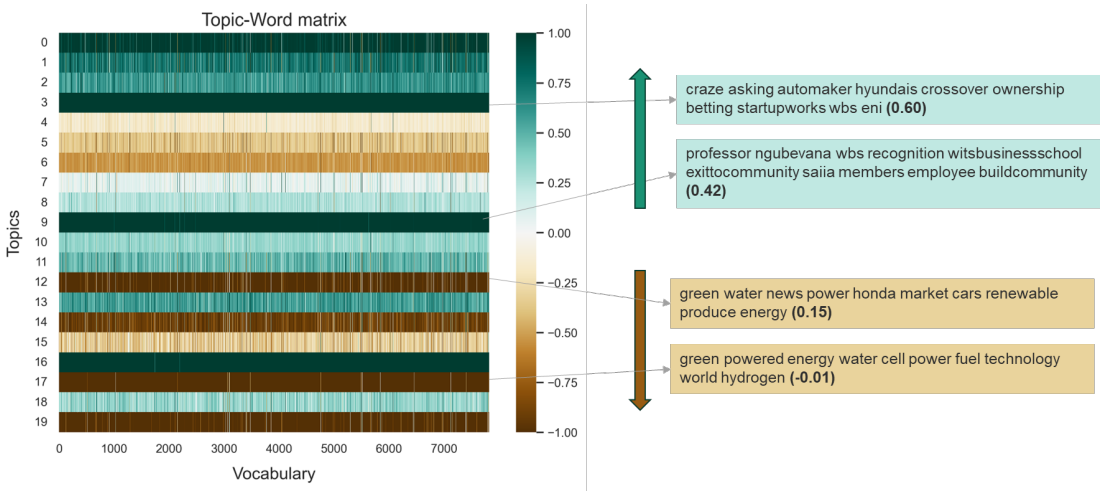


Figure 7: Heatmaps of learnt word-topic distributions and extracted topics of GATNN.

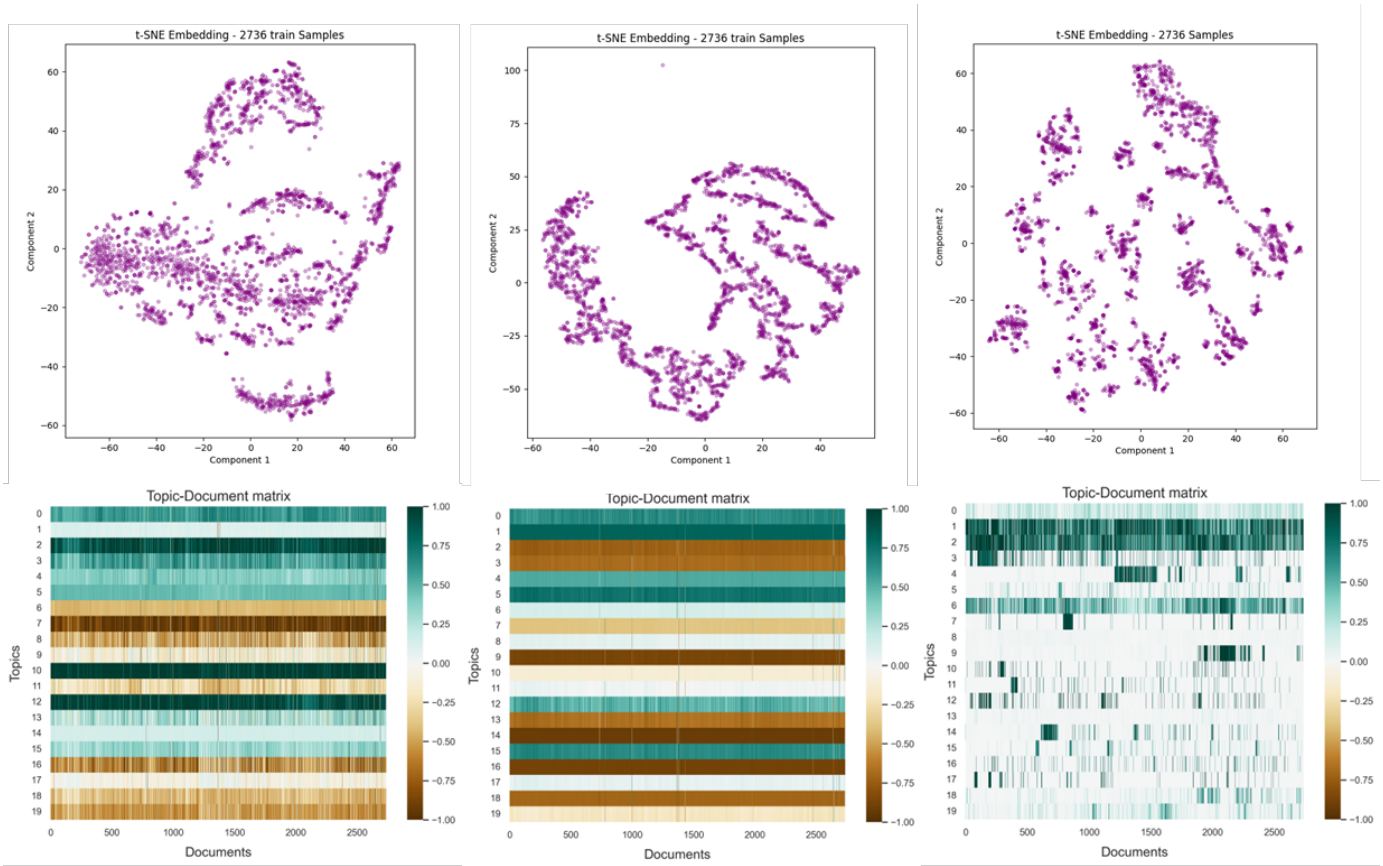


Figure 8: TSNE (top row) and heatmaps (bottom row) of learnt doc-topic distributions by model (Left: GTNN, Centre: GATNN, Right: NaNMF

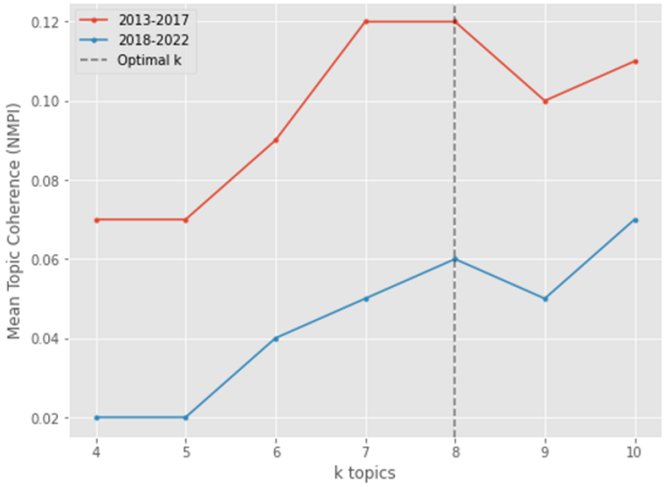


Figure 9: Graph showing the topic coherence score at different numbers of k for 2013-2017 (red) and 2018-2022 (blue) datasets. The final chosen k is 8, indicated by the dashed line.

noisy words with occurrences of hashtags and topics unrelated to Hydrogen Energy. This coincides with the findings above that interpretable topics tend to be associated with lower coherence scores, highlighting the intricate relationship between quantitative topic coherence and the human qualitative assessment of topic quality. To facilitate a more human-understandable approach to topic analysis, we have opted for NaNMF as our method of choice, rooted in its capacity for generating interpretable, meaningful and comparatively distinctly defined topics.

4.3 Topic Analysis

8 topics were chosen as the optimal k for both datasets derived from Fig. 9, using which we analysed the key themes and topics in the following two subsections. Generally, we observed common discussions on Twitter over the two periods around hydrogen technology, fuel cells, fueling infrastructure and transportation applications, indicating the enduring importance of these most-conversed aspects of hydrogen energy. It also indicates a competitive landscape of energy alternatives like electric, solar and wind power. Comparing the two periods, the later topics appear to demonstrate more global perspectives, with mentions of Germany, India, China, compared to a more industry-specific focus specifically on car manufacturers. Moreover, the earlier topics largely centered on hydrogen-powered vehicles and technology, while the more recent topics encompass a wider range of applications, including energy storage, renewables, and industrial use. The later period also involved a more prominent discussion about the transition to clean and renewable energy sources, energy storage and reducing carbon emissions. The shifts highlight the growing worldwide interest in and recognition of hydrogen's potential as a key player in transitioning to cleaner and more sustainable energy solutions. This also suggests the sustained research and development efforts in this field.

4.3.1 2013 - 2017

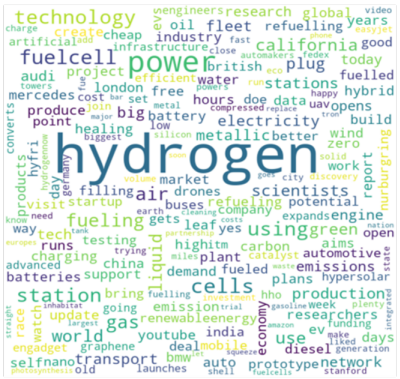
1) **Fuel Cell Technology**: centered around hydrogen power cells, fuel cell technology, and fueling stations, focusing on the infrastructure and technology for hydrogen-powered vehicles.
2) **Electric Vehicles Competition**: highlights the competition among electric car manufacturers, focusing on Tesla, Daimler, and advancements in electric vehicle technology.
3) **Toyota's Hydrogen Vehicles**: revolves around Toyota's Mirai and other hydrogen-fueled vehicles, covering aspects like patents, CES presentations, and sales of hydrogen-powered sedans.
4) **Fuel Cell Vehicle Costs and Technology**: delves into the technology and costs associated with fuel cell vehicles, featuring Hyundai and discussions about lowering the costs.
5) **Future Honda and Hyundai Vehicles**: suggests a look into the future of Honda and Hyundai in terms of hydrogen-powered vehicle production, including SUVs.
6) **Hydrogen Cars and Concepts**: centered around hydrogen-powered vehicles, including concepts from BMW, Tesla, and Riversimple, focusing on the American market.
7) **Diverse Hydrogen Transportation Applications**: explores various hydrogen-powered modes of transportation, such as drones, trams, and bikes, as well as hydrogen's use in achieving zero-emission flights.
8) **Clean and Renewable Hydrogen Energy**: highlights the clean and renewable aspects of hydrogen energy, especially in the context of solar power, water storage, and scientific advancements.

4.3.2 2018 - 2022

1) **Hydrogen in Clean Energy Economy**: highlights hydrogen's role in the clean energy economy, focusing on news, renewables, and projects that aim to produce clean energy using hydrogen.
2) **Hyundai's Contribution to Hydrogen-Powered Vehicles**: centers on Hyundai's involvement in developing fuel cell technology for electric vehicles, particularly hydrogen fuel cell buses.
3) **Hydrogen-Powered Transportation**: focuses on Germany and various modes of transportation with the mentions of buses, trains, cars, trucks and ships, highlighting the global shift towards more diverse transportation applications.
4) **Green Projects in India and Renewable Hydrogen**: delves into green projects, emphasizing India's role, and the production of renewable hydrogen through solar power and ammonia projects, showing a commitment to sustainable energy solutions.
5) **New Technology and Catalysts for Electric Cars**: discusses about new electric technology, production methods, and catalysts for cars, with a specific reference to Toyota and China.
6) **Renewable Energy Transition and Storage**: centered around the transition to renewable energy sources and the vital role of hydrogen in energy storage, supporting the broader goal of clean and sustainable energy.

Model	Top 3 Topics	NMPI
GTNN	counting creacleanair travelgreentraxx decarbonising steve invaded infoforeign grobbelaar speak lwazi	0.50
	automaker hyundais craze crossover betting americas cohort employee formation fulcrum	0.42
	members lwazi nygovcuomo infoforeign neuma ngubevana industries advisor intensive policy	0.41
GATNN	cox convention carbonzero getontraxx chad ceron sadowy jupiter speak matthew	0.60
	hon ward clip steelworks keanmp matt kembla welcomed press garethjward	0.54
	professor ngubevana wbs recognition witsbusinessschool exittocommunity saiia members employee buildcommunity	0.42
NaNMF	toyota mirai vehicle fcv fueled patents tesla sedan sell away	0.16
	powered world bus trains germany worlds train drone year diesel	0.14
	new catalyst water way seawater year produces method efficiently post	0.14

Table 5: Top three topics represented by top 10 words per model.



Fuel Cell Technology | Electric Vehicles Competition | Toyota's Hydrogen Vehicles

Fuel Cell Vehicle Costs and Technology | Future Honda and Hyundai Vehicles | Hydrogen Cars and Concepts

Diverse Hydrogen Transportation Applications | Clean and Renewable Hydrogen Energy

Figure 10: 8 Topics Extracted for 2013 - 2017 using NaNMF

7) **Reducing Carbon Emissions and Low-Carbon Energy Sources**: highlights the pursuit of low carbon emissions and the utilization of low-carbon energy alternatives, particularly wind, solar, and hydrogen production, reflecting a commitment to addressing climate change.
8) **Hydrogen as a Replacement for Fossil Fuels**: centers on the potential for hydrogen to replace fossil fuels, including natural gas and oil, in various industries, emphasizing its role in achieving sustainability goals.

4.4 Future Recommendations

Although GTNN and GATNN outperformed NaNMF in terms of topic coherence, the interpretability of the topics was hindered by the presence of noisy and irrelevant content. The extracted topics exhibited less distinctiveness across documents and words, suggesting potential oversmoothing of neighborhood information in models trained with two layers of BGMP. As pointed out in [18], using just one layer of BGMP produced the best results because the 2-hop low-pass graph filtering already captured order-2 document relation information. Additionally, the inclusion of noisy words might be attributed to employing PPMI values in word-word connections, with higher weights assigned to unique but infrequent word pairs and diluted weights for common and relevant words that frequently occurred in the corpus. High document frequency for common words might lead to an overall diluted PPMI, potentially causing the models to struggle to distinguish such words. Future research should explore the impact of PMI on topic quality and assess the suitability of different word-association and topic coherence measures for this dataset.

Furthermore, the effectiveness of GCN-based models could be compromised by the lack of distinct connections established initially, given the abundance of self-contained tweets. The downsampling approach resulted in no observed interactions between any two documents, and the density of the people connection matrix was only 15.9%, indicating a bias toward documents sharing mentions of common people, with insufficient consideration of meaningful replies and retweet interactions due to dataset limitations.

Another limitation of our study is the absence of labels, which restricted comprehensive clustering and classification analysis commonly found in related literature. Future work could assess model performance on different labeled tweet datasets, potentially shedding light on the reasons for the suboptimal performance. The high sparsity of document-word representations (2736 documents vs. 7821 words) is worth investigating further, and potential mitigation strategies such as incorporating spelling corrections in the preprocessing phase may be explored.

Finally, it's important to note that the topic analysis provided a static snapshot of two distinct periods without considering the sentiments expressed in these discussions. A longitudinal study of topics combined with sentiment analysis over the years could yield more insightful perspectives on evolving conversations. Moreover, the topics extracted from the final sampled dataset may not truly represent all topics.

5 Conclusion

In conclusion, this paper proposed a novel topic model for short-text modelling in Twitter context to deal with the extreme sparseness in the short-text document representations. We identified that no study has explored integrating document neighbourhood information combined with the rich semantic features available in Twitter into graph convolutional neural network for high-order relational topic modelling. Due to the short length of the tweet body, we hypothesised that using neighbourhood information propagation could potentially promote complementary topic learning between similar documents. As a solution, we introduced a novel model GATNN, established upon the GTNN backbone, with the extension of attention mechanism that fuses the document affinity derived from the combined interaction and content relations.

In summary, this study endeavoured to show how document similarity can be integrated into GTNN. Our findings show that, in general, GCN methods tend to yield interpretable, relevant and unique topics as compared to NaNMF, despite a higher NPMI topic coherence observed. Comparing GTNN and GATNN, our proposed method yields more distinct TSNE clusters, indicating its ability to capture deeper similarity relations via the attention network. However, the degraded performance can be attributed to the inherent sparsity in the dataset as a result of downsampling to 2736 documents, highlighting the importance of leveraging larger dataset derived from the initial data collection. Moreover, the effect of PMI in the topic learning and topic coherence evaluation warrants further investigation, as it was observed that relevant words tend to exhibit diluted PMI values due to high document frequency within the domain-specific dataset pertained to Hydrogen Energy as opposed to other diverse datasets. Future work involves applying the proposed model and experimental framework to other labelled datasets for comprehensive topic evaluation using classification and colour-encoding clustering, and ones that accommodate more document interactions.

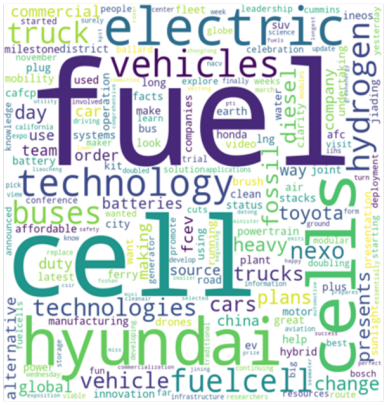
Moreover, the research has effectively closed the existing knowledge gaps of topic exploratory on Hydrogen Dataset by shedding light on the common themes and shifts across two periods, highlighting the transition to a global application. Future directions can focus on trend detection across smaller timeframes to detect finer patterns and shifts over the years, given that a larger dataset is used to derive representable topics.

Nevertheless, this study establishes a fundamental foundation for future advancements in the field of topic modeling, highlighting its potential for further improvements and innovations.

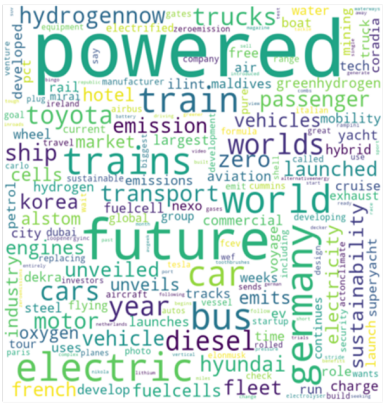
Acknowledgments

I want to extend my gratitude towards my supervisor, Deepak Uniyal, and my cluster tutor, Wathsala Mohotti, for their patience, guidance and expertise throughout these few months. I would also like to thank Qianqian Xie, one of the co-authors of [18] and Shalani Athukorala, co-author of [14] for sharing the source codes with me.

Hydrogen in Clean Energy Economy



Hyundai's Contribution to Hydrogen-Powered Vehicles

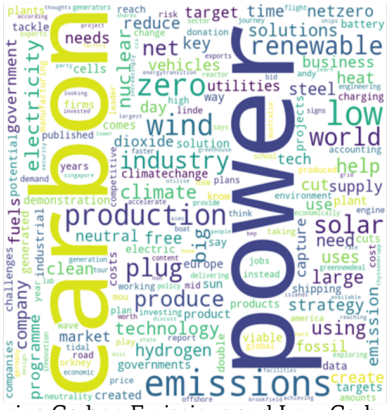


Hydrogen-Powered Transportation

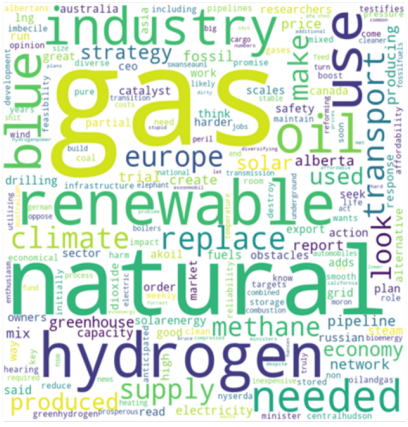
Green Projects in India and Renewable Hydrogen

New Technology and Catalysts for Electric Cars

Renewable Energy Transition and Storage



Reducing Carbon Emissions and Low-Carbon Energy Sources



Hydrogen as a Replacement for Fossil Fuels

Figure 11: 8 Topics Extracted for 2018 - 2022 using NaNMF

References

[1] "Deep learning based topic and sentiment analysis: Covid19 information seeking on social media," *Social network analysis and mining*, vol. 12, no. 1, pp. 90–90, 2022.

[2] "A survey of recent methods on deriving topics from twitter: algorithm to evaluation," *Knowledge and information systems*, vol. 62, no. 7, pp. 2485–2519, 2020.

[3] F. Qiao and K. Jiang, "Attitudes towards global warming on twitter: A hedonometer-appraisal analysis," *Journal of global information management*, vol. 30, no. 7, pp. 1–20, 2022.

[4] J. Schreiber, A. Scherrer, and H. L. Breetz, "Driving discussion: Media framing of electric, hydrogen, and conventional vehicles in german newspapers and twitter," *Energy research social science*, vol. 103, pp. 103193–, 2023.

[5] D. Fischer-Preßler, C. Schwemmer, and K. Fischbach, "Collective sense-making in times of crisis: Connecting terror management theory with twitter user reactions to the berlin terrorist attack," *Computers in human behavior*, vol. 100, pp. 138–151, 2019.

[6] G. D. Sharma, M. Verma, B. Taheri, R. Chopra, and J. S. Parihar, "Socio-economic aspects of hydrogen energy: An integrative review," *Technological forecasting social change*, vol. 192, pp. 122574–, 2023.

[7] R. Sharma, H. Sharda, A. Dutta, A. Dahiya, R. Chaudhary, A. Singh, K. Rathi, S. Kumar, A. Sharma, S. Maken, and S. Nehra, "Optimizing green hydrogen production: Leveraging load profile simulation and renewable energy integration," *International journal of hydrogen energy*, 2023.

[8] Australian Renewable Energy Agency, "Hydrogen energy," 2023.

[9] T. Saheb, M. Dehghani, and T. Saheb, "Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis," *Sustainable computing informatics and systems*, vol. 35, pp. 100699–, 2022.

[10] E. M. Rogers, *Diffusion of innovations*. New York: Free Press, 5th ed. ed., 2003.

[11] C. D. P. Laureate, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *The Artificial intelligence review*, vol. 56, no. 12, pp. 14223–14255, 2023.

[12] F. Albanese and E. Feuerstein, "Improved topic modeling in twitter through community pooling," *arXiv.org*, 2021.

[13] A. Ozgur, A. Yildirim, and S. Uskudarli, "Identifying topics in microblogs using wikipedia," *PloS one*, vol. 11, no. 3, pp. e0151885–e0151885, 2016.

[14] S. Athukorala and W. Mohotti, "An effective short-text topic modelling with neighbourhood assistance-driven nmf in twitter," *Social network analysis and mining*, vol. 12, no. 1, pp. 89–89, 2022.

[15] Q. Zhu, Z. Feng, and X. Li, "GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 4663–4672, Association for Computational Linguistics, Oct.-Nov. 2018.

[16] D. Zhou, X. Hu, and R. Wang, "Neural topic modeling by incorporating document relationship graph," *arXiv.org*, 2020.

[17] H. Zhao, D. Phung, V. Huynh, J. Yuan, L. Du, and W. Buntine, "Topic modelling meets deep neural networks: A survey," *arXiv.org*, 2021.

[18] Q. Xie, J. Huang, P. Du, M. Peng, and J.-Y. Nie, "Graph topic neural network for document representation," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 3055–3065, Association for Computing Machinery, 2021.

[19] C. Zhang and H. W. Lauw, "Topic modeling on document networks with adjacent-encoder," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 6737–6745, Apr. 2020.

[20] L. Yang, F. Wu, J. Gu, C. Wang, X. Cao, D. Jin, and Y. Guo, "Graph attention topic modeling network," in *Proceedings of The Web Conference 2020*, WWW '20, (New York, NY, USA), p. 144–154, Association for Computing Machinery, 2020.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine learning research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[22] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," p. 535–541, 2000.

[23] "Using time-sensitive interactions to improve topic derivation in twitter," *World wide web (Bussum)*, vol. 20, no. 1, pp. 61–87, 2017.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv.org*, 2017.

[25] Q. Xie, J. Huang, P. Du, and M. Peng, "Graph relational topic model with higher-order graph attention auto-encoders," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2604–2613, Association for Computational Linguistics.

[26] Y. Zhang, T. Senjyu, C. So-In, and A. Joshi, *Smart Trends in Computing and Communications: Proceedings of SmartCom 2022*. Lecture Notes in Networks and Systems, Springer Nature Singapore.

[27] A. Schofield and D. Mimno, "Comparing apples to apple: The effects of stemmers on topic models," vol. 4, pp. 287–300.

[28] R. W. N. Nugroho, J. Yang, J. Yang, Jian Yang, Jian Yang, Jian Yang, Y. Zhong, C. Paris, and S. Nepal, "Deriving topics in twitter by exploiting tweet interactions," pp. 87–94. MAG ID: 1603381115.

[29] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (S. Wintner, S. Goldwater, and S. Riezler, eds.), (Gothenburg, Sweden), pp. 530–539, Association for Computational Linguistics, Apr. 2014.

[30] Y. Chen and M. Zaki, "Kate: K-competitive autoencoder for text," in *Proceedings of the 23rd ACM SIGKDD International Conference on knowledge discovery and data mining*, vol. 129685, pp. 85–94, ACM, 2017.