IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 January 2023, accepted 20 March 2023, date of publication 28 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262653

RESEARCH ARTICLE

DATM: A Novel Data Agnostic Topic Modeling Technique With Improved Effectiveness for Both Short and Long Text

MICHAEL BEWONG[1], (Member, IEEE), JOHN WONDOH[2], SELASI KWASHIE[3], JIXUE LIU[2], LIN LIU[2], JIUYONG LI[2], (Member, IEEE), MD. ZAHIDUL ISLAM[1], AND DAVID KERNOT[4]

[1]School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga Wagga, NSW 2650, Australia
[2]School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA 5095, Australia
[3]Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Bathurst, NSW 2795, Australia
[4]Defence Science Technology Group, Department of Defence, Edinburgh, SA 5111, Australia

Corresponding author: Michael Bewong (mbewong@csu.edu.au)

This work was supported in part by Charles Sturt University publication grant, and the Charles Sturt University Early Career Researcher Award 2021 (Michael Bewong).

ABSTRACT Topic modelling is important for tackling several data mining tasks in information retrieval. While seminal topic modelling techniques such as Latent Dirichlet Allocation (LDA) have been proposed, the ubiquity of social media and the brevity of its texts pose unique challenges for such traditional topic modelling techniques. Several extensions including *auxiliary aggregation*, *self aggregation* and *direct learning* have been proposed to mitigate these challenges, however some still remain. These include a lack of consistency in the topics generated and the decline in model performance in applications involving disparate document lengths. There is a recent paradigm shift towards neural topic models, which are not suited for resource-constrained environments. This paper revisits LDA-style techniques, taking a theoretical approach to analyse the relationship between word co-occurrence and topic models. Our analysis shows that by altering the word co-occurrences within the corpus, topic discovery can be enhanced. Thus we propose a novel data transformation approach dubbed *DATM* to improve the topic discovery within a corpus. A rigorous empirical evaluation shows that DATM is not only powerful, but it can also be used in conjunction with existing benchmark techniques to significantly improve their effectiveness and their consistency by up to 2 fold.

INDEX TERMS Document transformation, greedy algorithm, information retrieval, latent dirichlet allocation, multi-set multi-cover problem, probabilistic generative topic modelling.

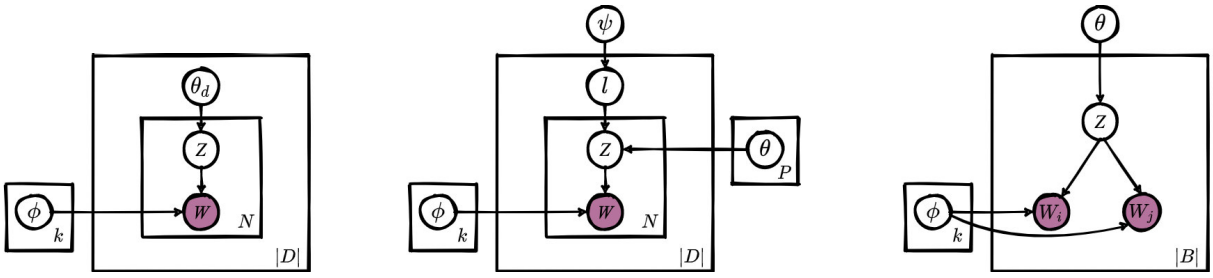
I. INTRODUCTION

Topic modelling is the task of discovering latent thematic topics in any given corpus. An example of such a task is the autonomous detection of topical discussions on social media.[1] This has use in many different disciplines such as in national security operations for riot prediction and misinformation detection where the manual monitoring of social media feeds for anti-social and anti-factual behaviour is seriously limiting [1]. In technology, topic modelling is used in everyday recommender systems, document tagging and efficient search [2], [3]. Even in healthcare data analysis, topic models are used to identify molecular subtypes of cancer, which significantly helps improve treatment regimes and health outcomes [4]. These and several other applications make the task of topic modelling crucial, requiring well thought-out solutions.

Conventional topic modelling techniques rely on probabilistic generative models for mimicking the human document generation process [5], [6]. In [7], the seminal LDA model assumes each document d in the corpus D is associated

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

[1]Topical discussions here are contextual and may refer to precursory information relevant to real life events such as riots (e.g., U.S, capitol riots-2021), scale of disasters (e.g., Lebanon explosion-2020), or extremist or antisocial cyber behaviour.



(a) Conventional Topic Model: LDA (b) Self aggregation: PTM (c) Direct Learning: BTM

FIGURE 1. Plate notation representation of various topic models.

with a distribution θ_d over the set of topics Z, and each topic $z \in Z$ is associated with a distribution ϕ_z over words. A document $d \in D$ is generated by iteratively selecting a topic from the distribution θ_d and then a word w from the distribution ϕ_z. Figure 1a illustrates this process. In the figure, k is the number of topics while $|D|$ is the size of the corpus. However, LDA and other conventional techniques are ineffective with short texts [6], [8], [9].

Several approaches have been proposed to improve the effectiveness of topic models for short text,[2] namely (1) *auxiliary aggregation* [9], [10], [11], [12], [13], [14], [15]; (2) *self aggregation* [16], [17], [18]; and (3) *direct learning* [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. In auxiliary aggregation, related texts are associated using some auxiliary information such as authors or publication venues to increase their length [13] or generate more context [10]. In self aggregation, it is assumed that short texts are generated from some longer latent document represented as a variable. Figure 1b illustrates PTM [17], which is a self aggregation technique. In PTM, each short document $d_s \in D$ is generated from some longer pseudo-document l. Each pseudo-document l is sampled from a distribution ψ. Each pseudo-document is associated with a topic distribution θ_l from which a topic z is selected, and there are P pseudo-documents. The word w is then sampled from the corresponding word distribution ϕ_z. In direct learning techniques, the global topics are learnt directly from word co-occurrences within the corpus. Figure 1c illustrates BTM [19], where it is assumed that documents can be represented as biterms B denoted (w_i, w_j). For each biterm $(w_i, w_j) \in B$, a topic z is sampled from the distribution θ and the biterms (w_i, w_j) sampled from the corresponding word distribution ϕ_z.

However, these approaches face 3 main challenges. Firstly, due to their probabilistic generative assumptions, they lack reliability. That is, each run of the algorithm on the same set of documents can generate a different set of topics. This makes it challenging to apply these techniques to scenarios where the consistency of topics derived from the same set of documents in different iterations is important. Such applications include tracking changes in social media content [29], [30]; cross-site user generated content modelling [31]; and opinion summarisation [32].

Secondly, in corpora involving a mixture of both long and short text, these techniques do not fully apply. For example, customer reviews are often a mixture of short and long text.[3] Analysing such data require techniques that can simultaneously derive topics for both short and long text. Unfortunately, as demonstrated in the empirical section, these techniques do not cope well for such cases.

Thirdly, these approaches are not robust under realistic assumptions. In PTM, the assumption that each piece of short text is associated with a single topic may not hold. For example, soundbites are short text but often cover multiple topics. Further, the availability of large-sized corpora is not always guaranteed. When these assumptions fail, the effectiveness of the existing techniques diminish. Also, their parameters are challenging to set, *i.e.,* there is no intuitive way to specify the number of pseudo-documents in PTM, the misspecification of which leads to ineffective learning [25]. For example in SATM [16] the parameters increases as the number of documents increase making it prone to over fitting.

Neural topic models (NTMs) [33], [34], [35], [36], [37], [38], [39], [40], [41] have shown promising results, however they are not absolute. A recent extensive study in [42] on NTMs shows that in some topic modelling contexts (*eg.* document modelling), LDA-style techniques are better suited. Secondly, pre-trained word embeddings often used in NTMs may not be relevant (*eg.* out-of-vocabulary words often arises in social media lingo) or trustworthy (*eg.* due to adversarial ML). At the same time, learning new embeddings can be impractical in resource-constrained environments [43]. Thus, in this work, we focus on LDA-style techniques which plays a key role in topic modelling.

Motivated by these challenges, our work focusses on investigating the relationship between the underlying topic concentrations within documents and the effectiveness of LDA-style topic models. We make the observation that, the topic concentration within a document positively affects the learning of topics [8]. Specifically, by improving the topic concentration within the input documents, we can improve

[2]Conventionally, tweet length of 280 characters is considered as short text.

[3]The Maccas dataset used in this paper has a document length range of 5-2910 words.

the reliability and effectiveness of the topics discovered. We propose a novel data transformation technique called *data agnostic topic modelling* (DATM) which improves the reliability and effectiveness of topic models for both short and long text. Our work seeks to make the following specific contributions:

- First, we formalise the relationship between word co-occurrences and topic concentration. We then cast the problem of enhancing topic concentration as a multi-set multi-cover problem which when optimised, can optimise the topic concentrations in documents.
- Second, we propose a greedy solution and theoretically analyse the bounds of optimality using primal-dual analysis. Further, we study the complexity of our greedy solution and develop a new data structure enabling the development of a second, more efficient solution with an improved quadratic time complexity.
- Third, we develop a suite of experiments to evaluate the reliability via consistency and purity; effectiveness via topic correctness and coherence; and quality via a qualitative analysis. To the best of our knowledge, this is the first work of its kind to evaluate the reliability of topic models comprehensively.
- Fourth, we present an empirical analysis using 3 diverse datasets and 4 benchmark techniques as baselines to demonstrate the efficacy of DATM. Our analysis shows that DATM significantly improves the reliability and effectiveness of topic models. Further, we show that even for challenging datasets that are not amenable to topic modelling, DATM can lead to considerable improvements. In particular, our experiments show that purity of topics and consistency of topics generated over time is improved on average by 2 fold. In addition, topic correctness, coherence and quality are also significantly improved.

The rest of this paper is organised as follows. In Section II, we present the related work. In Section III, we present the preliminaries, and formalise the data transformation approach in Section IV. Section V presents the framework of our solution: a basic greedy solution and an improved greedy solution called DATM. Section VI presents a rigorous evaluation of our approach in comparison with benchmark techniques. Finally, a conclusion is presented in Section VII.

II. RELATED WORK

Conventional LDA-style topic modelling techniques are often ineffective in short text [6], [9]. This ineffectiveness is due to the sparsity of word co-occurrences in the short texts compared to longer text. Several approaches have been proposed to improve the effectiveness of topic models for short text including (1) auxiliary aggregation [9], [10], [11], [12], [13], [14], [15]; (2) self aggregation [16], [17], [18]; and (3) direct learning [19], [20], [21], [22], [23], [24], [25], [26], [27], [28].

A. AUXILIARY AGGREGATION

In Auxiliary aggregation, the short text is made longer or further contextualised by associating the shorter text to form longer contextual documents by some auxiliary information. These techniques include [9], [10], [11], [12], [13], [14], [15]. Recently in [10], the authors propose a novel variational graph auto-encoder approach to align authorship, publication venue and semantic interpretability. In [11], the authors propose a model that combines both clustering and topic modelling. The aim is to discover the topics specific to a cluster as well as the global topics that relate to the corpus. Reference [12] proposes a pooling technique for topic modelling that groups short texts that occur in the same user-to-user conversation. Reference [13] proposed combining tweets from a single user to form longer documents. Reference [9] also proposed combining tweets that contained the same words. Other techniques in this category include [14], [15]. However, these approaches rely on auxiliary information that may not always be available and in some cases such aggregation results in ineffective groupings.

B. SELF AGGREGATION

In self aggregation, the short text is often assumed to be derived from some longer latent document, which is taken into account in the inference procedure by modelling it as a variable. They include [16], [17], [18]. In [16], the authors propose **SATM** which is an aggregation of short text based on the latent longer documents to which they are related. Its model assumes that regular sized documents can be generated by using typical topic models like LDA, and then subsequently a sampling of words from the longer documents to form the shorter texts. In [17], the authors propose **PTM**, the primary assumption is that high volume of short texts are generated from much less regular sized pseudo-documents. In the model, this distribution is specified by the user. A key aspect of PTM is that, by modelling the pseudo-document, higher co-occurence in the pseudo-document can be leveraged. Reference [18] proposes a technique to improve SATM by an adaptive aggregation process. Further, [44] considers co-occurring short and long text by likening blogs to normal text and their comments as short text. In so doing, they leverage the normal text to enrich the topic learning from the short texts and vice versa.

The techniques above, while formidable do not offer a principled approach to specifying the number of pseudo-documents. For example when the number of pseudo-documents specified is either higher or lower, the learning process suffers. In [17], the authors propose SPTM, which uses spike and slab priors to mitigate this issue, however, the choice between PTM or SPTM depends on the number of pseudo-documents and its constitution, both of which are impossible to determine a priori. Further [16] makes a strong assumption that short text will always have a single topic while [17] assumes each short text comes from at most one document. Both of these assumptions can be

impractical. For example soundbites which are short texts retrieved from longer, and often different documents and can talk about different topics.

Unlike [16], [17], our data transformation does not make such assumptions, nor does it require a user to specify the nature of the distribution of short text over longer pseudo-documents. In our case we improve the co-occurrence by using the underlying co-occurrence structure in the whole corpus. Our method learns the word co-occurrences in the original corpus and aggregates the words from the corpus based on the commonly co-occurring words to form a new corpus.

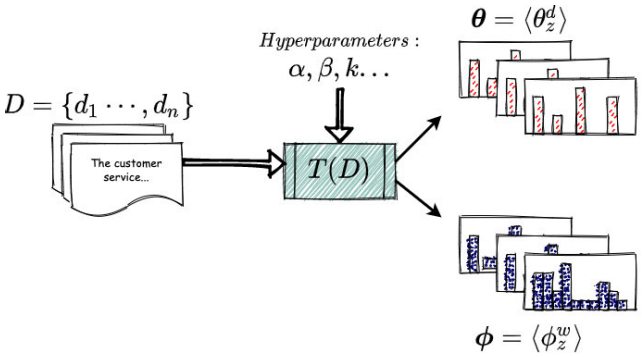


FIGURE 2. Topic modelling process.

C. DIRECT LEARNING FROM WORD STRUCTURES

A key relationship between documents and the topic models is that, for better topic models, documents must exhibit a skewed distribution of topics in a document and each topic must exhibit a skewed distribution of words [8]. Techniques such as [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28] which adopt this approach aim to enhance these properties. In [19], **BTM** is proposed to learn topics directly from the word co-occurrences rather than the documents. In this way the model learns the global topics based on how words co-occur within the corpus. Reference [20] further improves the efficiency of BTM by reducing the sampling complexity using *Metropolis Hastings and alias method*. Reference [45] proposes the relational biterm topic modelling to improve BTM which may not capture some related words which never co-occur. The relational biterm topic model captures these by linking short texts using a similarity list of words computed employing word embeddings. Reference [21] proposes an online biterm topic model for short texts which extends short texts with an external resource to make up for the data sparsity, and using online BTM to select representative topics instead of the word vector to represent the feature of short texts. Reference [22] improves the biterm topic model by introducing an embeddings based topic model for short texts. The method is purported to be able to distinguish noise topics from latent topics. Reference [23] proposes a technique called GraphBTM which aims to represent biterms with a graph convolutional network with residual connections to extract transitive features from biterms. Other similar works include [24], [25], [26], [27], [28].

Further works include [24], which proposes a non-negative factorization technique to solve the problem of learning topics from short texts. The work first generates a term correlation matrix which can be decomposed into two factors, a term topic matrix and its transpose. The term-topic matrix together with the term-document matrix can then be used to determine the topic-document matrix. Reference [25] proposes a novel technique based on semantics-assisted non-negative matrix factorization model (SeaNMF) to discover topics for short texts. The work of [26] proposes a word co-occurrence network based model called word network topic model (WNTM), to mitigate sparsity in short text. Reference [27] extends this work by introducing RWNTM which filters out irrelevant information during sampling. Further, [28] proposes the use of word embeddings to improve upon the works in [26] and [27].

In contrast to these approaches, our work does not model the topic directly from the biterms, but rather a collection of biterms which apparently belong to the same topic. As seen in the experimental section, this enhances the reliability and effectiveness of the learning process. At the same time [25], [28] requires word embeddings which relies on pre-trained word embeddings which may not be relevant nor readily available. We refer the reader to [33], [34], [35], [36], [37], [38], [39], and [46] for other word embeddings based approaches.

It can be seen from the related the works that none provide an approach to addressing the topic consistency problem. Further, some of the works make strong assumptions which may be implausible in certain scenarios. Finally, some of the works require extraneous techniques such as word embeddings or auxiliary information about the documents which may be impractical. We refer the reader to [47] for a recent comprehensive survey on topic modelling techniques.

III. PRELIMINARIES & BACKGROUND

Let D be a set of documents $\{d_1, \cdots, d_n\}$, also called a corpus such that each document d is a multiset of words $\{w_1, \cdots, w_{v_d}\}$. Each word w is drawn from the universal set of words W, also called vocabulary. Let $Z = \{z_1, \cdots, z_k\}$ be a set of k topics in the corpus D. Each topic z is associated with the distribution $\boldsymbol{\phi}_z = \langle \phi_z^1, \cdots, \phi_z^{|W|} \rangle$ over the set W of words, and ϕ_z^w is the probability of the word w in topic z. Each document d is associated with the distribution $\boldsymbol{\theta}^d = \langle \theta_1^d, \cdots, \theta_k^d \rangle$ over k topics, and θ_z^d is the probability of topic z in document d.

Given a corpus D, a generative topic model denoted by T aims to estimate the distributions $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for all topics Z and documents D respectively as depicted in Figure 2. The works in [8], [48], [49], [50] have demonstrated that the effectiveness of a topic model depends on the input dataset and correct hyperparameter specification. In [8], two categories of guidelines are provided for generative topic models. Firstly, in parameter specification, smaller values for the dirichlet

hyperparameters α, where documents are associated with a smaller subset of topics, and β, where topics are well separated, yield demonstrably better results. In addition, a larger k parameter than the true number of topics lead to poorer topic models. Secondly, for the input data, a larger corpus made up of longer documents each of which relates to a smaller subset of topics yields better topic models. These guidelines are re-echoed in [17], [19], and [16].

However, the interpretation of these guidelines is neither obvious nor trivial. For example, in most real scenarios such as online articles and comments, the corpus is often made up of a mixture of both long and short text. Intuitively, longer texts may cover more topics than the shorter texts. For instance, a longer review of a restaurant may cover more topics such as cleanliness, location and customer service than a shorter review. In this scenario, it is ambiguous how the hyperparameters should be specified. Some proposals on treating the hyperparameters as random variables to be learnt, somewhat optimally, leads to complex models that are slow to converge, and not practical in real scenarios [50] and [8].

In this paper, we seek to improve the robustness of a topic model T in estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by leveraging a data transformation approach. In particular we seek to make topic models data agnostic leading to less guesswork in hyperparameter specification and greater reliability of topic models in dynamic environments.

It is worth noting that, while the works in [17], [19], and [16] focus on improving topic models for short text, they do not work well in scenarios involving a mixture of short and long text. In such scenarios, their underlying assumption on how the documents were generated fails, resulting in poorer topic models. This is demonstrated and further discussed in the empirical section (*c.f.* Section VI).

IV. DATA TRANSFORMATION FOR TOPIC MODELLING

Our aim is to improve the effectiveness of generative topic models in estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. In particular, we develop a data transformation F that transforms any given set of documents D into an "ideal" set of documents D', *i.e.* $F(D) = D'$, such that the resulting topics from D' are better than those from D. Our data transformation improves the topic concentration within documents by ensuring that each document in our transformed corpus D' is associated with a smaller subset of topics as topic concentration is a key factor in achieving effective topic models [8], [17]. This is further illustrated by the following example.

Example 1: In this example, we sample two equally sized sets of corpora ($\approx$ 3000) from the real Airline dataset (c.f. *Table 1). There are 3 topics in each corpus. The first corpus, denoted Airline_Single, is a set of documents each of which has been assigned only one topic in the ground truth. The second, denoted Airline_Multiple, is a set of documents each of which has been assigned more than one topic in the ground truth. The topic models PTM and LDA were then applied and their purity*[4] *scores calculated. Figure 3 shows the results of this experiment. In the figure, it is clear that Airline_Single whose documents each have only one topic* i.e. *"higher" concentration of topics performs much better on both PTM and LDA than the Airline_Multiple corpus.*

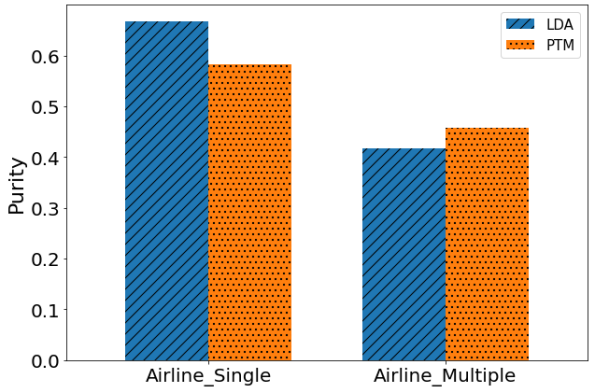
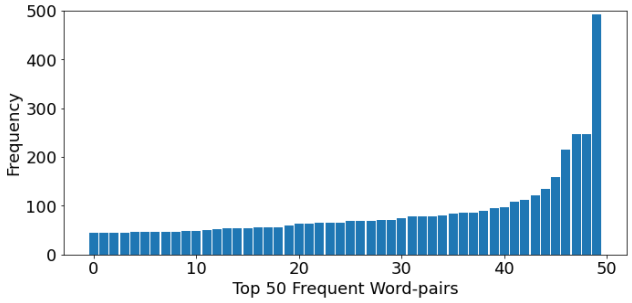
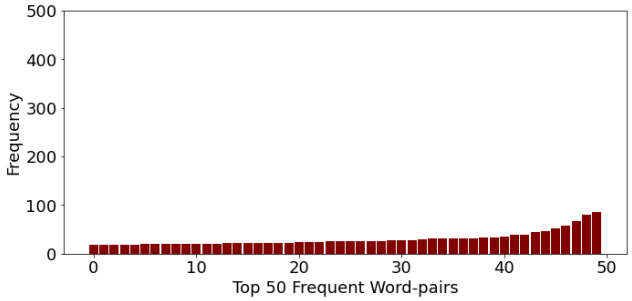


FIGURE 3. High topic concentration *vs.* low topic concentration.



(a) Airline_Single



(b) Airline_Multiple

FIGURE 4. Word-pair distribution.

We further analyse the word-pair distribution of each corpus in Example 1 and make an important observation from the results in Figure 4 that, when documents are concentrated on fewer topics (*i.e.* Airline_Single), their word co-occurrence is also concentrated on fewer word-pairs (Figure 4a). Subsequently, by controlling the word co-occurrence structure we can control the topic concentration within documents.

[4]Purity is a measure of the clarity of the topics found, further described in Section VI.

We now present some definitions and proofs to support this observation.

Given a corpus D and a topic model T such as LDA,[5] let $\mathrm{T}(D) = M$ be the resulting topics discovered from the corpus, such that $M = \{m_{wd}\}$ is a word-document matrix and each element m_{wd} is the topic allocated to the word w in document d. Following [48], we formally define the topic distribution $\boldsymbol{\phi}$ as follows.

Definition 1 (Topic Distribution $\boldsymbol{\phi}$): Given a corpus D, a topic model T *and the matrix* $\mathrm{T}(D) = M$*, the topic distribution for any topic $z \in Z$ is given by $\boldsymbol{\phi}_z = \langle \phi_z^1, \cdots, \phi_z^{|W|} \rangle$, where $\phi_z^w = \frac{N_z^{(w)} + \beta}{N_z^{(\cdot)} + |W|\beta}$ is the probability that the word w belongs to topic z, $N_z^{(w)}$ is the number of copies of word w assigned to topic z, $N_z^{(\cdot)}$ is the number of words assigned to topic z, $|W|$ is the vocabulary size, and β is the LDA dirichlet parameter.*

The number of copies of word w assigned to topic z, $N_z^{(w)}$ is calculated from the matrix M. The distribution $\boldsymbol{\phi} = \langle \boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_k \rangle$ is the topic distribution for all topics.

Definition 2 (Document Distribution $\boldsymbol{\theta}$): Given a corpus D, a topic model T *and the matrix* $\mathrm{T}(D) = M$*, the document distribution for any document $d \in D$ is given by $\boldsymbol{\theta}^d = \langle \theta_1^d, \cdots, \theta_k^d \rangle$, where $\theta_z^d = \frac{N_z^{(d)} + \alpha}{N_{(\cdot)}^{(d)} + k\alpha}$ is the probability for any word in document d to belong to topic z, $N_z^{(d)}$ is the number of words in document d assigned to topic z, $N_{(\cdot)}^{(d)}$ is the number of words in document d, k is the number of topics, and α is the LDA dirichlet parameter.*

The number of copies of words in document d assigned to topic z, $N_z^{(d)}$ is calculated from the matrix M. The distribution $\boldsymbol{\theta} = \langle \boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^n \rangle$ is the document distribution for all documents.

Following the definition of the distribution of a document over topics $\boldsymbol{\theta}$, the topic concentration within a document may be described geometrically as the location of the document within a topic polytope. We quantify this degree of concentration by defining a *topic spread* measure based on *gini-simpson index*.[6]

Definition 3 (Topic Spread, $\mathcal{T}_S$): Given the document d and its topic distribution $\boldsymbol{\theta}^d = \langle \theta_1^d, \cdots, \theta_k^d \rangle$, the topic spread in d is defined via its gini-simpson index as:

$$\mathcal{T}_S(d) := 1 - \sum_z (\theta_z^d)^2.$$

We note that $\mathcal{T}_S(d) \in [0, 1]$, and the use of *gini-simpson index* to define topic spread ensures that a low value reflects a low topic spread and thus a high topic concentration, and vice versa. A low topic spread in document d means that it has a higher degree of association with a smaller subset of topics in the distribution $\boldsymbol{\theta}^d$. Consequently, documents with lower topic spread are closer to the boundaries of the topic polytope.

[5]For simplicity we assume LDA topic model, however by treating the hyperparameters α and β as random variables, the analysis easily extends to other topic models.

[6]A well-known variance based measure of diversity [51], [52], [53].

The topic spread within a corpus D is the average topic spread of all the documents in D, *i.e.* $\mathcal{T}_S(D) := AVG_{\forall d \in D}(\mathcal{T}_S(d))$.

As observed in Example 1, when there is a high topic concentration, the word co-occurrence is also high. In the following, we present two new concepts, *word-pair ratio* and *word-pair spread*, for estimating the word co-occurrence in a document.

Given a document d, let $Pairs(d) = \{\boldsymbol{w}\}$ be the set of all word-pairs in d. $\boldsymbol{w}$ is an unordered pair of words (w_i, w_j) for all $w_i, w_j \in d \wedge i \neq j$. Let $\mathcal{M}(\boldsymbol{w}, d)$ be the multiplicity of each word-pair $\boldsymbol{w}$ in d, *i.e.* $\mathcal{M}(\boldsymbol{w}, d)$ is the number of occurrences of $\boldsymbol{w}$ in d.

Definition 4 (Word-Pair Ratio $P(\boldsymbol{w}_i, d)$): Given the corpus $D = \{d_1, \cdots, d_n\}$ and the word-pairs $Pairs(d) = \{\boldsymbol{w}\}$ for every document $d \in D$, the word-pair ratio $P(\boldsymbol{w}_i, d)$ of any word-pair $\boldsymbol{w}_i$ in a document d is given by:

$$P(\boldsymbol{w}_i, d) = \frac{\mathcal{M}(\boldsymbol{w}_i, d)}{\max_{d_j} \sum_{\boldsymbol{w}} \mathcal{M}(\boldsymbol{w}, d_j)}.$$

$P(\boldsymbol{w}_i, d)$ is the ratio of the number of copies of $\boldsymbol{w}_i$ in document d w.r.t. the total number of copies of all word-pairs in any document d_j with the most number of total word-pairs.

Definition 5 (Word-Pair Spread $\mathcal{W}_S$): Let d, $Pairs(d)$ and $P(\boldsymbol{w}, d)$ be a document, its word-pairs and the word-pair ratio of $\boldsymbol{w}$ in d respectively. The word-pair spread $\mathcal{W}_S(d)$ in a document d is given by:

$$\mathcal{W}_S(d) := 1 - \sum_{\forall \boldsymbol{w}_i \in Pairs(d)} P(\boldsymbol{w}_i, d)^2.$$

Word-pair spread $\mathcal{W}_S \in [0, 1]$ characterises the topic spread in a document. Our definition above uses the word-pair ratio $P(\boldsymbol{w}_i, d)$, which differs from conventional measures in [51], [52], and [53] that calculate $P(\boldsymbol{w}_i, d)$ as a ratio of the frequency of $\boldsymbol{w}_i \in d$ to its own document size $\sum_{\boldsymbol{w}} \mathcal{M}(\boldsymbol{w}, d)$. Our approach *i.e.* $P(\boldsymbol{w}_i, d) = \frac{\mathcal{M}(\boldsymbol{w}_i, d)}{\max_{d_j} \sum_{\boldsymbol{w}} \mathcal{M}(\boldsymbol{w}, d_j)}$ not only allows us to compare the relative prominence of word-pairs in a document, as does the conventional approach, but it also allows us to compare the prominence of any word pair in one document to another. The following example illustrates this point.

Example 2: ***Scenario 1****: $d_1 = (\boldsymbol{w}_1 : 2, \boldsymbol{w}_2 : 2)$ and $d_2 = (\boldsymbol{w}_1 : 5, \boldsymbol{w}_2 : 5)$ and let $\max_d \sum_{\boldsymbol{w}} m(\boldsymbol{w}, d) = 10$. By our definition, word-pair spread of d_1 is* 0.92 *while that of d_2 is* 0.5*. This implies that d_2 has lower word-pair spread than d_1. Intuitively, d_2 contributes more to topic concentration due to having more copies of the word-pairs than d_1. In contrast, conventional measures such as entropy will assign the same value of* 1 *to both documents.*

Scenario 2*: $d_1 = (\boldsymbol{w}_1 : 5)$, $d_2 = (\boldsymbol{w}_1 : 5, \boldsymbol{w}_2 : 1)$ and $d_3 = (\boldsymbol{w}_1 : 1, \boldsymbol{w}_2 : 1, \boldsymbol{w}_3 : 4)$ and let $\max_d \sum_{\boldsymbol{w}} m(\boldsymbol{w}, d) = 10$. By our definition, word-pair spread of d_1 is* 0.75*, d_2 is* 0.74 *while d_3 is* 0.80*. This implies that d_2 has lower word-pair spread than d_1, and d_1 is lower than d_3. Intuitively, d_2 contributes more to topic concentration due to having the same number of copies of the word pairs in d_1 plus extra information in the additional word pair. On the other hand, d_3 has the*

highest word-pair spread even though it has the same total number of word pairs as d_2, but because the distribution is more spread out than either of d_1 and d_2, it has the highest. In contrast, the traditional measures such as entropy will assign the value of 0 to d_1, 0.65 to d_2 and 1.25 to d_3. This subtle difference is important since our measure ensures that additional information is rewarded.

In the following Lemma 1, we formally relate the word-pair spread $\mathcal{W}_\mathcal{S}$ calculated from the observed word-pair distribution of documents, and the topic spread $\mathcal{T}_\mathcal{S}$ calculated from the topic distributions within documents. First, we make the following simplifying assumption about the corpus.

Assumption 1: Given a corpus $D = \{d_1, \cdots, d_n\}$, such that each document $d = \{w_1, \cdots, w_{v_d}\}$ is a multiset of words and Pairs(d) is the set of word-pairs in d. Also let $\{z_1, \cdots, z_k\}$ be the set of topics associated with the corpus D denoted $\boldsymbol{\theta}^d = \langle \theta^d_1, \cdots, \theta^d_k \rangle$ where θ^d_z is the proportion of topic z in document d. We make the following simplifying assumptions about the corpus D.

a. *All documents in the corpus D are of equal length,* i.e. $|d_i| = |d_j|$ *for all* $i, j \in \{1, \cdots, n\}$.
b. *There exists a non-injective function $f : w \mapsto z$ mapping every word in a document to at most one topic* i.e. *each word w belongs to only one topic z.*

Assumption 1.a. simplifies the analysis of the word counts required in Definition 2. Assumption 1.b. relates to the complete topic separation of the underlying true topics. Thus, under the condition that all documents are of equal length *e.g.* tweets,[7] and the underlying latent topics are mutually exclusive, the following lemma holds.

Lemma 1: Given any two documents d_i and d_j, and their respective topic proportions $\boldsymbol{\theta}^{d_i}$ and $\boldsymbol{\theta}^{d_j}$, the following monotonicity holds $\mathcal{W}_\mathcal{S}(d_i) \geq \mathcal{W}_\mathcal{S}(d_j) \implies \mathcal{T}_\mathcal{S}(d_i) \geq \mathcal{T}_\mathcal{S}(d_j)$.

Proof: The proof follows from analysing the geometric locations of a document in its topic and word-pair polytopes. Let $\overline{Pairs}(D) = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_l\}$ be the set of word-pairs derived from D such that each word-pair $\boldsymbol{w}$ belongs to a single topic. Let $G = conv(\overline{Pairs}(D))$ be the convex hull of the word-pairs. G is also called a word-pair polytope. Every document d is a point in the word-pair polytope space *i.e.* $d \in G$, such that $\boldsymbol{d} = \langle P(\boldsymbol{w}_1, d), \cdots, P(\boldsymbol{w}_l, d) \rangle$ is its vector representation. Similarly let $H = conv(Z)$ be the topic polytope, such that each document is a point within the topic polytope. Let $f(d) := \{\boldsymbol{w}\}_d \mapsto \{z\}$ be the resultant topic mapping of the word-pairs in d. From this mapping, the topic distribution $\boldsymbol{\theta}^d = \langle \theta^d_1, \cdots, \theta^d_k \rangle$ can be calculated via $\theta^d_z = \frac{N^{(d)}_z + \alpha}{N^{(d)}_{(\cdot)} + k\alpha}$ (*cf.* Definition 2). Also let $dist(\boldsymbol{d}, G) = \min_{\boldsymbol{g} \in extr(G)} ||\boldsymbol{d} - \boldsymbol{g}||$ be the distance between the document point $\boldsymbol{d}$ and the closest extreme point in G. We need to show that when $dist(\boldsymbol{d_i}, G) \geq dist(\boldsymbol{d_j}, G) \implies dist(\boldsymbol{\theta}^{d_i}, H) \geq dist(\boldsymbol{\theta}^{d_j}, H)$. From Definition 2 and for any given topic z, $\theta^d_z \propto N^{(d)}_z$ as all other variables are constant. When $\mathcal{W}_\mathcal{S}(d_i) \geq \mathcal{W}_\mathcal{S}(d_j)$ the distribution of word-pairs in d_j is more skewed than those in d_i. Consequently for document d_j some word-pairs have higher probability and appear more in the document. This increases the values of $N^{(d)}_z$ for the topics to which the word-pairs are related than any other topics. Subsequently the distribution $\langle \theta^d_1, \cdots, \theta^d_k \rangle$ of the document follows the same distribution as $\langle P(\boldsymbol{w}_1, d), \cdots, P(\boldsymbol{w}_l, d) \rangle$ due to our non-injection function f. Thus $dist(\boldsymbol{\theta}^{d_i}, H) \geq dist(\boldsymbol{\theta}^{d_j}, H)$ holds and $\mathcal{T}_\mathcal{S}(d_i) \geq \mathcal{T}_\mathcal{S}(d_j)$ is true when $\mathcal{W}_\mathcal{S}(d_i) \geq \mathcal{W}_\mathcal{S}(d_j)$. □

The lemma is correct even when the assumptions are removed. We look at the implications of removing the assumptions.

Assumption 1.a. ensures that the value $N^{(d)}_{(\cdot)}$ (*cf.* Definition 2) remains constant. When this does not hold, $N^{(d)}_{(\cdot)}$ may vary thus $\theta^{(d)}_z \propto \frac{N^{(d)}_z}{N^{(d)}_{(\cdot)}}$ generally as $N^{(d)}_z \gg \alpha \wedge N^{(d)}_{(\cdot)} \gg k\alpha$ in practice.

Assumption 1.b. relates to the degree of separation of the underlying topics. Without this assumption, a word-pair may relate to multiple topics. When the underlying topics are completely separated, the distance between the word-pairs and the polytope is 0 since each word-pair coincides with at least one extreme point in the polytope. When the distance is > 0 a shift in the word-pairs from the extreme points have occurred. Consequently, regardless of the word-pair spread in a document, the best possible topics that can be derived will still share overlapping word-pairs. The proof holds but there is a lower bound for the topic spread determined by the location of the word-pairs in the topic polytope. This is a limiting property of the corpus rather than an artefact of our proof.

□

From the analysis of the proof above, the lemma holds true even when the simplifying assumptions are removed.

Note that, this lemma formalises the observation made in Example 1. Thus, by improving the word-pair co-occurrence structure within a document we can improve the topic concentration within documents. In the following section we formally present this objective as an optimisation problem.

A. PROBLEM FORMULATION

In this section we model the problem of improving the topic concentrations as an optimisation problem. Particularly we observe that this optimisation problem is a multi-set multi-cover problem.

Given a corpus $D = \{d_1, \cdots, d_n\}$ of documents and a set of word-pairs $Pairs(D) = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m\}$ denoted by U, let $\mathcal{S} = \{S : Set(S) \subseteq U\}$ be a collection of subsets of U. Each S is a multiset associated with a cost $c_S := \mathcal{W}_\mathcal{S}(S)$; and each word-pair $\boldsymbol{w}$ may appear in the multiset S with a multiplicity of $\mathcal{M}(\boldsymbol{w}, S) \leq \mathcal{M}(\boldsymbol{w}, D)$. The goal of the optimisation problem is to obtain the subset C of $\mathcal{S}$ with the cheapest cost such that each word-pair $\boldsymbol{w}$ appears at least $r_{\boldsymbol{w}}$ times in C. $r_{\boldsymbol{w}} = \mathcal{M}(\boldsymbol{w}, D)$ is the number of occurrences of $\boldsymbol{w}$ in the original

[7]Strictly speaking, tweets are not of equal length however each tweet is bound to a maximum of 280 characters.

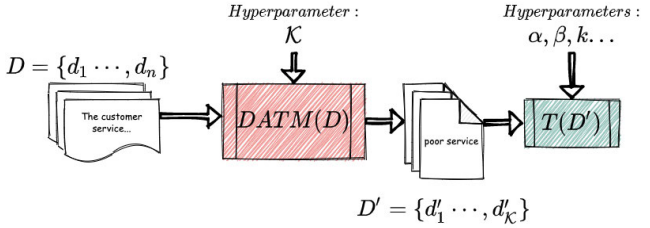


FIGURE 5. DATM framework.

corpus D. C, also called the cover, becomes the transformed corpus D', trivially $\bigcup_{S \in C} S = U$. This is formally defined as follows.

Definition 6 (Problem Definition): Given a set of word-pairs $U := Pairs(D) = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m\}$ and a collection of subsets $\mathcal{S} = \{S : Set(S) \subseteq U\}$. The objective is to select a subset C of $\mathcal{S}$ that minimises the total word-pair spread $\mathcal{W}_{\mathcal{S}}(C)$ such that each word-pair $\boldsymbol{w}$ is covered at least $r_{\boldsymbol{w}}$ times:

$$\begin{aligned} &minimise \quad \sum_{\forall S \in \mathcal{S}} \mathcal{W}_{\mathcal{S}}(S) \cdot x_S \\ &s.t. \quad \sum_{\forall S \in \mathcal{S}} \mathcal{M}(\boldsymbol{w}, S) \cdot x_S \geq r_{\boldsymbol{w}} \\ &\quad -x_S \geq -1 \\ &\quad x_S \geq 0 \end{aligned} \tag{1}$$

x_S is a selection variable.

$\mathcal{S}$ does not contain any S with spurious word-pairs. A spurious word-pair is one that does not exist in the original corpus. Also $\mathcal{W}_{\mathcal{S}}(S) : S \rightarrow \mathbb{R}^+$. This implies every element S in $\mathcal{S}$ is a non-negative cost element. We note that the optimisation problem is a generalisation of the set cover problem which is NP-Hard, by induction this problem is also NP-Hard [54], [55].

The following presents our *data agnostic topic modelling*, *DATM*, which solves this optimisation problem for improved topic modelling.

V. SOLUTION FRAMEWORK

DATM framework has two parts: (1) a transformation of any input set of documents D into an ideal set of documents D' (Definition 6); and (2) the discovery of topics from the transformed set of documents D'. To address the first part, we present a basic greedy algorithm and show it is an n'th harmonic factor approximation of the optimal solution. By developing a new data structure, we introduce a more efficient candidate set generation and foldback technique which improves the computational complexity without sacrificing optimality. In the second part, we adopt existing generative topic models such as LDA, with reasonably specified parameters α, β and k. Figure 5 illustrates our approach.

Algorithm 1 Basic Greedy Algorithm

Input: A collection $\mathcal{S}$ of all possible multisets S of word-pairs in D
Output: A collection of word-pair multisets C

1 Initialization $C \leftarrow \emptyset$;
2 **while** *there exists* $w : r(\boldsymbol{w}) < r_{\boldsymbol{w}}$ **do**
3 find $S \in \mathcal{S}$ with most cost-effectiveness $\frac{\mathcal{W}_{\mathcal{S}}(S)}{|A(S)|}$;
4 update $C \leftarrow \{C + S\}$;
5 update the residuals $r(\boldsymbol{w})$;
6 **return** C.

A. BASIC GREEDY ALGORITHM

The algorithm is summarised in Algorithm 1. Let $r(\boldsymbol{w})$ be the residual requirement of $\boldsymbol{w}$ at any iteration in the algorithm (line 2). $r(\boldsymbol{w})$, which is initially $r_{\boldsymbol{w}}$, is decreased by $\mathcal{M}(\boldsymbol{w}, S)$ each time a multiset S containing $\boldsymbol{w}$ is added to the cover C (lines 3-5).

Let $a(\boldsymbol{w}, S) = min\{\mathcal{M}(\boldsymbol{w}, S), r(\boldsymbol{w})\}$. The set of alive word-pairs in S denoted $A(S)$ will be the multiset (a subset of S) containing exactly $a(\boldsymbol{w}, S)$ copies of any word-pair $\boldsymbol{w}$ if S is not already in the cover C; and it is an empty set if S is already in the cover. We denote the total number of alive elements in S as $|A(S)| = \sum_{\boldsymbol{w}} a(\boldsymbol{w}, S)$.

The greedy algorithm repeatedly adds a set S minimizing $\frac{\mathcal{W}_{\mathcal{S}}(S)}{|A(S)|}$ to the cover (line 3). When a set S is picked, the cost $\mathcal{W}_{\mathcal{S}}(S)$ is shared equally to the alive elements; let the tuple $(\boldsymbol{w}, i)$ denote the word-pair $\boldsymbol{w}$ covered the i'th time by S ($i \in [1, r_{\boldsymbol{w}}]$). The cost for every element $\boldsymbol{w}$ covered by S the i'th time is $cost(\boldsymbol{w}, i) = \frac{\mathcal{W}_{\mathcal{S}}(S)}{|A(S)|}$.

Note that the cost of covering $\boldsymbol{w}$ is independent of the $i'th$ time it was covered. Consider the scenario where two sets both cover word-pair $\boldsymbol{w}$. Assume S_1 covers $\boldsymbol{w}$ the first 3 times and S_2 covers $\boldsymbol{w}$ the last 2 times *i.e.* $r_{\boldsymbol{w}} = 5$; then S_1 will be associated with the tuples $(\boldsymbol{w}, 1)$, $(\boldsymbol{w}, 2)$, $(\boldsymbol{w}, 3)$ such that $cost(\boldsymbol{w}, 1) = cost(\boldsymbol{w}, 2) = cost(\boldsymbol{w}, 3)$ and S_2 will be associated to the tuples $(\boldsymbol{w}, 4)$ and $(\boldsymbol{w}, r_{\boldsymbol{w}})$ such that $cost(\boldsymbol{w}, 4) = cost(\boldsymbol{w}, r_{\boldsymbol{w}})$ and $cost(\boldsymbol{w}, 4) \geq cost(\boldsymbol{w}, 1)$. This observation is essential to the following n'th harmonic number H_n approximation of optimality of our solution in Lemma 2 and 3.

While the optimisation problem (Definition 6) is an integer program, we relax it to a linear program to allow x_S to be fractional, making the duality analysis possible to deal with fractional costs.

Lemma 2: There exist a dual to the primal function in the optimisation problem (c.f. *Definition 6*) *given by*:

$$\begin{aligned} &maximise \quad \sum_{\forall \boldsymbol{w} \in U} r_{\boldsymbol{w}} \cdot y_{\boldsymbol{w}} - \sum_{\forall S \in \mathcal{S}} z_S \\ &s.t. \quad \sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(S, \boldsymbol{w}) \cdot y_{\boldsymbol{w}} - z_S \leq \mathcal{W}_{\mathcal{S}}(S) \\ &\quad z_S \geq 0, y_{\boldsymbol{w}} \geq 0 \end{aligned} \tag{2}$$

Proof: The proof of Lemma 2 via duality analysis shows the primal and dual have the common objective function $\sum_{\forall \boldsymbol{w} \in U} \sum_{i=1}^{r_{\boldsymbol{w}}} cost(\boldsymbol{w}, i)$. We make use of duality analysis to show the H_n approximation of our algorithm. From the corresponding dual form (*c.f.* Formula 2), we want to find the values for $y_{\boldsymbol{w}}$ and z_S that maximises the objective function. We have the following:

$$y_{\boldsymbol{w}} := \frac{max_i\{cost(\boldsymbol{w}, i)\}}{H_n} = \frac{cost(\boldsymbol{w}, r_{\boldsymbol{w}})}{H_n} \tag{3}$$

$$z_S := \begin{cases} 0, & \text{if } S \text{ is not picked} \\ \frac{\sum_{(\boldsymbol{w},i) coveredbyS} (cost(\boldsymbol{w}, r_{\boldsymbol{w}}) - cost(\boldsymbol{w}, i))}{H_n}, & \text{otherwise} \end{cases} \tag{4}$$

Following Algorithm 1, we show the relationship between the primal and dual solutions. The primal function (*c.f.* Definition 6) can be defined in terms of the cost of covering each word-pair *i.e.*

$$\sum_{\forall S \in \mathcal{S}} \mathcal{W}_{\mathcal{S}}(S) \cdot x_S := \sum_{\forall \boldsymbol{w} \in U} \sum_{i=1}^{r_{\boldsymbol{w}}} cost(\boldsymbol{w}, i) \tag{5}$$

Assuming $H_n = 1$ and substituting the values of $y_{\boldsymbol{w}}$ and z_S, our dual function becomes:

$$\begin{aligned} &\sum_{\forall \boldsymbol{w} \in U} r_{\boldsymbol{w}} \cdot y_{\boldsymbol{w}} - \sum_{\forall S \in \mathcal{S}} z_S \\ &:= \sum_{\forall \boldsymbol{w} \in U} r_{\boldsymbol{w}} \cdot cost(\boldsymbol{w}, i) - \sum_{(\boldsymbol{w},i) coveredbyS} (cost(\boldsymbol{w}, r_{\boldsymbol{w}}) - cost(\boldsymbol{w}, i)) \end{aligned} \tag{6}$$

Considering each word-pair at a time, and for all the sets that cover w_1 the *r.h.s* of Formula 6 becomes:

$$r_{\boldsymbol{w}} \cdot cost(\boldsymbol{w}, i) - r_{\boldsymbol{w}} \cdot (cost(\boldsymbol{w}_1, r_{\boldsymbol{w}})) + \sum_{i=1}^{r_{\boldsymbol{w}}} cost(\boldsymbol{w}_1, i) \tag{7}$$

For all word pairs, the first two terms of the above equation cancel out, thus Formula 6 simplifies to:

$$\sum_{\forall \boldsymbol{w} \in U} r_{\boldsymbol{w}} \cdot y_{\boldsymbol{w}} - \sum_{\forall S \in \mathcal{S}} z_S := \sum_{\forall \boldsymbol{w} \in U} \sum_{i=1}^{r_{\boldsymbol{w}}} cost(\boldsymbol{w}, i) \tag{8}$$

This value in the dual function also happens to be the value in the primal function (*c.f* Formula 5).

□

With the common objective function, we need to show that a solution to the dual function is feasible; and subsequently, that the dual solution is related to the primal solution by the factor H_n.

*Lemma 3: Given the dual objective function (*c.f. *Formula 1); y, z is dual feasible.*

Proof: The proof shows feasibility and an H_n bound on the greedy solution. y and z are non-negative, so the non-negativity constraints are trivially satisfied. We need to show that for any S, the remaining dual constraint $\sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(\boldsymbol{w}, S) \cdot y_{\boldsymbol{w}} - z_S \leq \mathcal{W}_{\mathcal{S}}(S)$ in Formula 2 is satisfied.

We focus on the *l.h.s* by substituting values of $y_{\boldsymbol{w}}$ and z_S from formula 3 and 4:

$$\frac{1}{H_n} \Big(\sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(\boldsymbol{w}, S) cost(\boldsymbol{w}, r_{\boldsymbol{w}}) \sum_{(\boldsymbol{w},i) coveredbyS} (cost(\boldsymbol{w}, r_{\boldsymbol{w}}) - cost(\boldsymbol{w}, i)) \Big) \tag{9}$$

The first term $\sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(\boldsymbol{w}, S) cost(\boldsymbol{w}, r_{\boldsymbol{w}})$ can be re-written as

$$\sum_{(\boldsymbol{w},i) coveredbyS} cost(\boldsymbol{w}, r_{\boldsymbol{w}}) + \sum_{(\boldsymbol{w},i) notcoveredbyS} cost(\boldsymbol{w}, r_{\boldsymbol{w}})$$

By substituting into Formula 9, with algebraic manipulation we have:

$$\frac{1}{H_n} \Big(\sum_{(\boldsymbol{w},i) notcoveredbyS} cost(\boldsymbol{w}, r_{\boldsymbol{w}}) + \sum_{(\boldsymbol{w},i) coveredbyS} cost(\boldsymbol{w}, i) \Big) \tag{10}$$

Returning to the *r.h.s* of the constraint $\sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(\boldsymbol{w}, S) \cdot y_{\boldsymbol{w}} - z_S \leq \mathcal{W}_{\mathcal{S}}(S)$, we need to find some value $V : V \leq \mathcal{W}_{\mathcal{S}}(S)$, such that $\sum_{\forall \boldsymbol{w} \in U} \mathcal{M}(\boldsymbol{w}, S) \cdot y_{\boldsymbol{w}} - z_S \leq V$, then the relation is proved.

Assuming S covered the elements $(\boldsymbol{w}, n)$ to $(\boldsymbol{w}, n+2)$ then the inner term of Formula 9 becomes

$$\begin{aligned} &[cost(\hat{\boldsymbol{w}}, r_{\boldsymbol{w}}) + \cdots + cost(\hat{\boldsymbol{w}}, n-1)] + [cost(\boldsymbol{w}, n) \\ &+ cost(\boldsymbol{w}, n+1) + cost(\boldsymbol{w}, n+2)] \end{aligned} \tag{11}$$

where $\hat{cost}(\cdot)$ is the cost not covered by S. Note that $cost(\boldsymbol{w}, n) = cost(\boldsymbol{w}, n+1) = cost(\boldsymbol{w}, n+2)$ and $[cost(\boldsymbol{w}, n) + cost(\boldsymbol{w}, n+1) + cost(\boldsymbol{w}, n+2)] = \mathcal{W}_{\mathcal{S}}(S)$. Thus Formula 11 simplifies to:

$$[cost(\hat{\boldsymbol{w}}, r_{\boldsymbol{w}}) + \cdots + cost(\hat{\boldsymbol{w}}, n-1)] + \mathcal{W}_{\mathcal{S}}(S) \tag{12}$$

The worst case scenario is that S is the longest subset with each element in S being unique, and covers only one element of $\boldsymbol{w}$:

$$[cost(\hat{\boldsymbol{w}}, r_{\boldsymbol{w}}) + \cdots + cost(\hat{\boldsymbol{w}}, n-1)] + \frac{\mathcal{W}_{\mathcal{S}}(S)}{1} \tag{13}$$

Assume the unique elements in S are ordered from those that were covered first until those that were covered last before the current set S was chosen. For the first element that was covered, the cost of the final copy of the element to be covered will still be less than the cost of the elements in the set S ie $cost(\boldsymbol{w}, r_{\boldsymbol{w}})_1 \leq \mathcal{W}_{\mathcal{S}}(S)/n$, $cost(\boldsymbol{w}, r_{\boldsymbol{w}})_2 \leq \mathcal{W}_{\mathcal{S}}(S)/n-1$ and so on until $cost(\boldsymbol{w}, r_{\boldsymbol{w}})_{n-1} \leq \mathcal{W}_{\mathcal{S}}(S)/2$. By this, we can arrive at a compact representation of

$$\sum_{i=1}^{n} \frac{1}{i} \mathcal{W}_{\mathcal{S}}(S).$$

Every term in this compact statement will be $\geq$ its corresponding term in:

$$\sum_{(w,i)} cost(\hat{\boldsymbol{w}}, r_{\boldsymbol{w}}) + \sum_{(w,i)} cost(\boldsymbol{w}, i)$$

Our V is $\frac{1}{H_n} \sum_{i=1}^{n} \frac{1}{i} \mathcal{W}_{\mathcal{S}}(S)$ which is also $\leq \mathcal{W}_{\mathcal{S}}(S)$ *i.e.* $\frac{1}{H_n} \sum_{i=1}^{n} \frac{1}{i} \mathcal{W}_{\mathcal{S}}(S) \leq \mathcal{W}_{\mathcal{S}}(S)$. Thus it can be concluded that

$\sum_{\boldsymbol{w}} \mathcal{M}(\boldsymbol{w}, S) \cdot y_{\boldsymbol{w}} - z_S \leq \mathcal{W}_{\mathcal{S}}(S)$ holds, and y, z is feasible. Note that H_n is the n'th harmonic number *i.e.* $H_n = \sum_{i=1}^{n} \frac{1}{i}$. □

It is worth mentioning that our greedy solution (Algorithm 1) does not limit the size of the word-pair multisets C which later becomes D', consequently, it is possible to encounter the situation where $|D'| \gg |D|$. This is not desirable as it can lead to a very slow inference of the topics from the transformed corpus D'. We further impose a constraint on the size of C without affecting our optimality analysis. Namely, we propose a user defined number of transformed documents $\mathcal{K}$ which limits the size of C *i.e.* $|C| \leq \mathcal{K}$. A *foldback* procedure (*c.f.* Procedure Foldback) can be applied to the output of Algorithm 1 to satisfy this constraint as follows.

Procedure Foldback($C, \mathcal{K}$)

while $|C| > \mathcal{K}$ **do**
  find $S_{lst}, S_{mst} \in C$; ▷ where S_{lst}, S_{mst} have the least and most cost $\mathcal{W}_{\mathcal{S}}(S)$ respectively
  update $C \leftarrow \{C \setminus \{S_{lst}, S_{mst}\}\}$;
  update $C \leftarrow \{C + \{\biguplus(S_{lst}, S_{mst})\}\}$;

The following lemma guarantees the H_n approximation even when the *foldback* technique is applied.

Lemma 4: Foldback procedure finds the multi-set multi-cover within H_n factor of the primal optimal solution.

Proof: Lemma 2 & 3 are correct and the foldback technique does not invalidate Lemma 2 & 3. We need to show that the foldback technique satisfies the following inequality:

$$Cost(S_1) + Cost(S_2) \geq Cost(\biguplus(S_1, S_2)) \tag{14}$$

From Definition 5 $Cost(S) = 1 - \sum_{i=1}^{n_S} p(\boldsymbol{w}_i, S)^2$; simplifying to $Cost(S) = 1 - \sum_{i=1}^{n_S} (\frac{\mathcal{M}(\boldsymbol{w}_i S)}{Const})^2$ where $Const$ is the maximum number of words in a document for all documents. Formula 14 becomes:

$$1 - \sum_{i=1}^{n_{S_1}} \left(\frac{\mathcal{M}(\boldsymbol{w}_i S_1)}{Const}\right)^2 + 1 - \sum_{i=1}^{n_{S_2}} \left(\frac{\mathcal{M}(\boldsymbol{w}_i S_2)}{Const}\right)^2 \geq 1 - \sum_{i=1}^{n_{S_{1,2}}} \left(\frac{\mathcal{M}(\boldsymbol{w}_i S_1) + \mathcal{M}(\boldsymbol{w}_i S_2)}{Const}\right)^2$$

By algebraic manipulation we get:

$$\sum_{i=1}^{n_{S_1}} \mathcal{M}(\boldsymbol{w}_i S_1)^2 + \sum_{i=1}^{n_{S_2}} \mathcal{M}(\boldsymbol{w}_i S_2)^2 \leq Const^2 + \sum_{i=1}^{n_{S_{1,2}}} (\mathcal{M}(\boldsymbol{w}_i S_1) + \mathcal{M}(\boldsymbol{w}_i S_2))^2$$

By comparing like terms we have:

$$\mathcal{M}(\boldsymbol{w}_i S_1)^2 + \mathcal{M}(\boldsymbol{w}_i S_2)^2 \leq (\mathcal{M}(\boldsymbol{w}_i S_1) + \mathcal{M}(\boldsymbol{w}_i S_2))^2$$

Which holds for any $\boldsymbol{w}_i$ due to the following algebraic identity which applies to the *r.h.s.* of the equation

$$(a+b)^2 = a^2 + b^2 + 2ab$$

Thus the *r.h.s.* of the equation can be expressed as:

$$\mathcal{M}(\boldsymbol{w}_i S_1)^2 + \mathcal{M}(\boldsymbol{w}_i S_2)^2 \leq \mathcal{M}(\boldsymbol{w}_i S_1)^2 + \mathcal{M}(\boldsymbol{w}_i S_2)^2 + 2\mathcal{M}(\boldsymbol{w}_i S_1)\mathcal{M}(\boldsymbol{w}_i S_2)$$

For any $\mathcal{M}(\boldsymbol{w}_i S_j) \in \mathbb{R}^+$ the Lemma holds. □

It follows from the above that $foldback(C, \mathcal{K})$ uses S_{lst} and S_{mst} in each iteration because the resultant cost $\mathcal{W}_{\mathcal{S}}(\biguplus(S_{lst}, S_{mst}))$ is bound to be cheaper than $\mathcal{W}_{\mathcal{S}}(S_{lst})$. This is further illustrated in Example 3. Approximation algorithms for multi-set multi-cover problems are well studied in [56], [57], [58], and [59]. Similar to these works, our basic greedy algorithm (Algorithm 1) has $O(log\ m)$ wrt to m word-pairs. However, we see that the main computational challenge is the generation of the candidate subsets in $\mathcal{S}$ which is exponential w.r.t. corpus size n *i.e.* $O(n \cdot 2^n)$. The derivation is as follows:

1) WORD-PAIR GENERATION

Given a corpus $D = \{d_1, \cdots, d_n\}$. Word-pair generation for each document d is $\binom{d}{2}$. Total worst case scenario is $n \cdot \binom{d_{max}}{2}$ where d_{max} is the longest document. By combinatorics and algebraic manipulation this becomes $n \cdot \frac{d_{max}^2 - d_{max}}{2}$. Thus the worst case complexity of word-pair generation is:

$$O(n \cdot d_{max}^2) \tag{15}$$

□

2) CANDIDATE SUBSET GENERATION

Given m total word-pairs from the corpus. Generation of the power set is $m \cdot 2^m$. From Formula 15, the total complexity is $(n \cdot d_{max}^2) + n \cdot d_{max}^2 \cdot 2^{(n \cdot d_{max}^2)}$. d_{max} is constant and for short text $n \gg d_{max}$. Thus *w.r.t.* number of documents the complexity is:

$$O(n \cdot 2^n) \tag{16}$$

□

In the following, we define a new data structure and algorithm for the generation of the candidate subsets and the selection of the subsets S in polynomial time such that when the algorithm completes, the required set of subsets C to our problem is $C := \mathcal{S}$.

B. IMPROVED GREEDY ALGORITHM

We define a new data structure called *global pairs*. Global pairs is a set of elements $\{Gp[i]\}$. Each element $Gp[i] = (ws_i, wd_i, cost_i)$ contains ws_i which is the set of words $set(d_i')$; wd_i which is a word-pair dictionary where each element of the dictionary is a word-pair *i.e.* $wd_i[j] = \{\boldsymbol{w}_j : count\}$ is the word-pair $\boldsymbol{w}_j$ and its count in the document d_i' and; $cost_i$ is the total word-pair spread of wd_i *i.e.* $cost_i := \mathcal{W}_{\mathcal{S}}(wd_i)$. Gp is sorted according to the cost. An element $Gp[i]$ is used to generate the final transformed document d_i'.

Algorithm 2 DATM

Input: A set of documents D; the number $\mathcal{K}$ of transformed documents
Result: A set of transformed documents D'

Step 1 ▷ Generate word-pairs from each document d_i
Initialization *Global Pairs* $Gp \leftarrow \emptyset$;
for $d_i \in D$ **do**
  Calculate the word-pairs $\boldsymbol{w}$ in each d_i ie $Pairs(d_i)$
  //Update Gp with $\boldsymbol{w}$
  if $\boldsymbol{w} \in Gp.wd$ **then**
    $Gp.wd[\boldsymbol{w}].count \leftarrow Gp.wd[\boldsymbol{w}].count + count(\boldsymbol{w}, d_i)$
  else
    Add a new element $(ws, wd, cost)$ with $wd = \{\boldsymbol{w} : count(\boldsymbol{w}, d_i)\}$ to Gp
Step 2 ▷ perform foldback on Gp to satisfy $\mathcal{K}$
Call procedure Foldback($C, \mathcal{K}$): $Foldback(Gp, \mathcal{K})$
Step 3 ▷ Generate documents from Gp
forall the $Gp[i]$ **do**
  $d'_i \leftarrow \biguplus(Gp[i].wd)$ ▷ Merge all word-pairs pairs with multiplicity
return $D' = \{d'_1, \cdots, d'_{\mathcal{K}}\}$.

Algorithm 2 summarises the steps in *DATM*. In step 1, the initialisation step, Gp is first initialised to an empty set $\emptyset$. For every document $d \in D$, all word-pairs $\boldsymbol{w} \in Pairs(d)$ are calculated and added to Gp if the word pair $\boldsymbol{w} \notin Gp$, else if $\boldsymbol{w} \in Gp$ then the count $Gp.wd[\boldsymbol{w}].count$ is updated.

In step 2, the foldback step, we recombine some of the elements in Gp to satisfy the constraint on the number of expected transformed documents. Since each transformed document d'_i is derived from the word pair dictionary wd_i of the element $Gp[i]$, the foldback step is performed using $Gp.wd$ instead of S as previously seen in *Foldback* procedure. Further, Gp is sorted according to cost, so finding the least cost-effective elements in Gp is straight forward.

In step 3, the document generation step, each transformed document d'_i is generated from the *global pairs* data structure by merging all the word-pairs in the dictionary wd of each element of Gp. The following example illustrates the working of the algorithm.

Example 3: Consider 3 documents with the following word-pairs $d_1 = (\boldsymbol{w}_1 : 2, \boldsymbol{w}_4 : 3)$, $d_2 = (\boldsymbol{w}_1 : 5, \boldsymbol{w}_2 : 1)$ *and* $d_3 = (\boldsymbol{w}_1 : 1, \boldsymbol{w}_2 : 1, \boldsymbol{w}_3 : 4)$ *and let* $\max_d \sum_{\boldsymbol{w}} m(\boldsymbol{w}, d) = 10$. *The word-pair spread for each document is calculated as* 0.87, 0.74 *and* 0.8 *for* d_1, d_2 *and* d_3 *respectively. The total is* 2.41.

In ***step 1*** *of Algorithm 2, Gp is initialised as* $Gp[1] = (\{set(\boldsymbol{w}_1)\}, (\boldsymbol{w}_1 : 8), 0.36)$, $Gp[2] = (\{set(\boldsymbol{w}_3)\}, (\boldsymbol{w}_3 : 4), 0.84)$, $Gp[3] = (\{set(\boldsymbol{w}_4)\}, (\boldsymbol{w}_4 : 3), 0.91)$, *and* $Gp[4] = (\{set(\boldsymbol{w}_2)\}, (\boldsymbol{w}_2 : 2), 0.96)$. *In* ***step 2***, $S_{lst} := Gp[1]$ *is combined with* $S_{mst} := Gp[4]$ *to form a new element* $Gp[1+4] = (\{set(\boldsymbol{w}_1, \boldsymbol{w}_2)\}, (\boldsymbol{w}_1 : 8, \boldsymbol{w}_2 : 2), 0.32)$. *Together* $Gp[1 + 4]$, $Gp[2]$ *and* $Gp[3]$ *have a total word-pair spread of* 2.07. *For a* $\mathcal{K} = 3$, *this completes the algorithm and the elements in Gp form the new corpus* D' *in* ***step 3***.

Algorithm 2 has a time complexity of $O(n \cdot d_{max}^2 + m^2)$ where n is the number of input documents; $d_{max} \in D$ is the longest document in the corpus; and m is the number of word-pairs from the corpus. Summarily, Step 1 has a worst case complexity of $O(n \cdot d_{max}^2)$ from generating the word-pairs; Step 2 has a worst case complexity of $O(m^2)$ from sorting the word pairs according to cost and choosing valid combinations and; step 3 has a worst case complexity of $O(m)$ when $\mathcal{K} = m$. It is important to note that Algorithm 2 is also an H_n approximation of the optimal solution due to Lemma 4.

We note that, the proposed solution can be built on top of any generative model. In the following, we present the empirical study of our proposed technique.

TABLE 1. Summary of datasets.

Dataset	Corpus Size	Avg. Doc. Length	Vocab. Size
Airline	14622	55.30 ± 24.48	10676
Maccas	2198	328.03 ± 305.58	7694
News*	500	1049.32 ± 1106.85	14398

VI. EMPIRICAL STUDY

Four benchmark algorithms were compared, namely PTM [17], BTM [19], LDA [7]; and DATM.[8] Commonly used existing implementations via Tomotopy libraries[9] for PTM and LDA, and Biterm libraries[10] for BTM were used. Remember that DATM is a data transformation approach which can be applied on top of any topic modelling technique. In our empirical study, we exemplarily use LDA and PTM denoted as DATM(LDA) and DATM(PTM).

Following the convention in [7], [16], [19], and [17], we set the parameters $\alpha = 0.1$, $\beta = 0.01$, and additionally $\lambda = 0.1$ and $P = 1000$[11] for PTM. For DATM, we set $\mathcal{K} = |D|$. We also set the number of topics k to the true number of topics in each dataset, and a Gibbs sampling iteration of 2000 to guarantee convergence. All implementations and experiments were done on a macOS with 2.6 GHz, 6-core Intel Core i7 processor and 16GB 2400 MHz DDR4 memory.

Three datasets that exhibit different characteristics were chosen to test the rigour of our data agnostic approach. The datasets are summarised in Table 1. Airline dataset[12] are customer tweets about their experience with a major U.S. airline in 2015. 62.76% of the tweets are categorised into one of nine possible topics. The Maccas dataset [12] are consumer reviews on Yelp about McDonalds. The reviews are

[8]DATM is available on github: https://github.com/mbewong/DATM_Public
[9]https://bab2min.github.io/tomotopy/v0.12.1/en/
[10]https://github.com/markoarnauto/biterm
[11]When $P > |D|$ we set $P = |D|$.
[12]Retrieved from https://www.figure-eight.com/data-for-everyone/ on 16/Mar/2018

TABLE 2. Correctness of topics found.

	Recall					Precision					F1				
	LDA	PTM	BTM	DATM (LDA)	DATM (PTM)	LDA	PTM	BTM	DATM (LDA)	DATM (PTM)	LDA	PTM	BTM	DATM (LDA)	DATM (PTM)
Airline	0.9	**1**	0.9	**1**	0.9	0.9	**1**	0.9	**1**	0.9	0.9	**1**	0.9	**1**	0.9
Maccas	**0.5**	**0.5**	**0.5**	0.4	**0.5**	**0.5**	**0.5**	0.4	0.4	**0.5**	**0.5**	**0.5**	**0.5**	0.4	**0.5**
News*	0.0	0.0	0.0	**0.2**	**0.2**	0.0	0.0	0.0	**0.2**	**0.2**	0.0	0.0	0.0	**0.2**	**0.2**

assigned 0 or more labels from 8 possible topics. The News* dataset is a random sample of 500 documents from 5 topics from the well-known 20Newsgroup dataset,[13] especially chosen to represent a smaller corpus with high variability in document lengths. Together, these 3 dataset represents 3 well encountered scenarios *viz* a large corpus of short documents with relatively stable length across the corpus (Airline); a medium sized corpus with medium sized documents and high variability in document length (Maccas); and a small corpus with long documents but very high variability in the document length. All datasets were pre-processed with NLP steps such as stop-word removal, stemming and lemmatization.

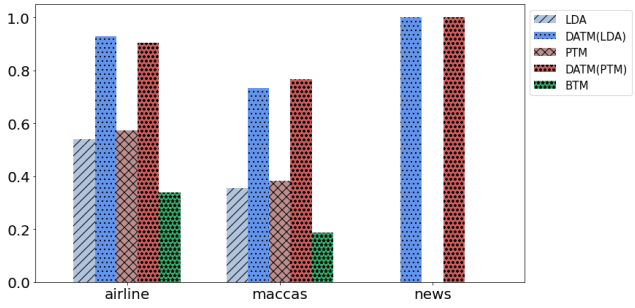


FIGURE 6. Average purity of topics found.

A. RELIABILITY OF TOPIC MODELS

This evaluation demonstrates how reliable the topics generated are. Reliability is evaluated via purity and consistency of topics. Purity is our proposed measure of the clarity with which the topics found express the true topics. We propose a measure of purity which compares the found topics to the true topics. If a found topic z matches many true topics $\hat{z}$ the found topic is said to be less pure *i.e.* $Purity(z) = \frac{1}{Count_{\forall \hat{z}}(Sim(z^*,\hat{z}^*)\geq c)}$, where z^* (resp. $\hat{z}^*$) is the most probable set of words from topic z (resp. $\hat{z}$), $Sim(a, b)$ is a distance-based similarity measure,[14] and c is a cut-off constant set to 0.6. Our measure of purity differs from from [16] since it considers all possible topic matches, but [16] only considers the highest match.

Figure 6 shows that DATM significantly improves the average purity of LDA and PTM on all datasets. This implies that, topics found using DATM, have higher clarity than the other benchmark techniques. We note that in the News* dataset, LDA, PTM and BTM failed to identify any true matching topics and thus their purity is 0. We see that not only does DATM improve the purity of LDA and PTM but it also improves the recall for challenging dataset such as News*. This is evident in the recall evaluation in Table 2 and further discussed in Section VI-B.

Consistency is measured by applying the topic model to the same dataset 5 times *i.e.* given a dataset D and a topic model T, a set of topic-sets $\{T(D)_1, \cdots, T(D)_5\}$ are generated. The average purity is calculated between each $T(D)_i$ and $T(D)_{i+1}$ for all iterations and plotted as boxplot in Figure 7. In Figure 7, DATM clearly shows an improvement of the consistency results. Figure 7 shows DATM has a median purity score of above 0.8 and a minimum score above 0.7. These are well above the competitors PTM, LDA and BTM. The results imply that when DATM is applied in many iterations on the same dataset, it is more likely that the user will get consistent topics than the existing benchmark techniques. This finding is also corroborated in Figures 7b and 7c. It is worth pointing out that while in Figure 7c, the maximum values are the same for LDA, PTM and DATM, the interquartile range for DATM is much tighter.

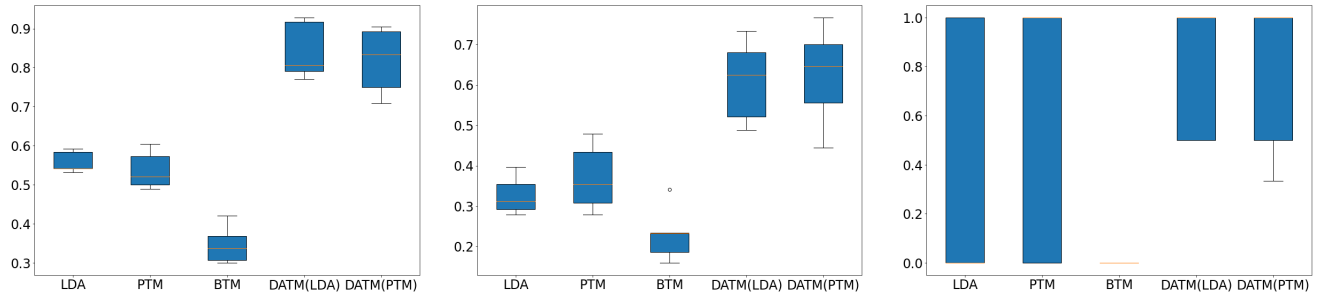
B. EFFECTIVENESS OF TOPIC MODELS

Effectiveness demonstrates how correct and coherent the topics generated are. We evaluate effectiveness using recall, precision, F1 and coherence scores. To calculate recall, precision and F1, we use the known topic labels to extract the probable key words as the ground truth. Using $(Sim(z^*, \hat{z}^*) \geq c)$, we calculate the matches to estimate the true (and false) positives, and true (and false) negatives. Table 2 which is the results shows similar results for all techniques for a given dataset. The Airline dataset yields the best results while News* dataset yields the worst. This finding is in keeping with the theoretical analysis that larger corpus size plays a positive role in effective topic modelling [8]. DATM is shown to be beneficial, particularly for the challenging News* dataset.

To further investigate the effectiveness of our approach, we evaluate the topic coherence of the topic models. Topic coherence is a well known metric for evaluating topic models. However they are neither reasonable nor accurate for evaluating short and informal texts [16], [17], [60]. Thus, we adopt coherence to evaluate the models on News* dataset. In particular, we use the C_V coherence measure proposed by [61]. We also use normalised pointwise mutual

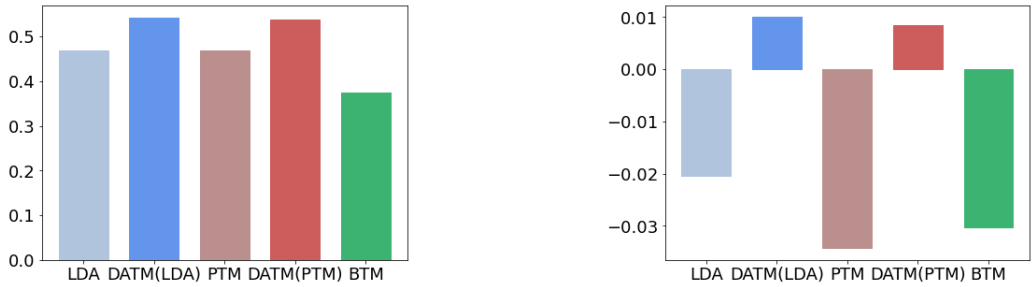
[13] Retrieved from https://www.kaggle.com/crawford/20-newsgroups on 6/June/2019

[14] In this work a sequence matcher from the python difflib module is used.



(a) Airline Dataset (b) Maccas Dataset (c) News* Dataset

FIGURE 7. Consistency of topic models.



(a) C_V (b) $C_{UCI}(NPMI)$

FIGURE 8. Coherence of topics.

information (NPMI) version of C_{UCI} [62] used in [16] and [17]. In both Figure 8a and 8b, DATM improves the coherence values. It is important to note that our aim is to observe the relative improvement of the coherence rather than the absolute values. In C_V, the values are in the range $[0, 1]$. We observe moderate values around 0.5 with DATM showing an improvement over the other techniques. We note that in C_{UCI}, the values are normalised to the range $[-1, 1]$. LDA, PTM and BTM show a negative coherence in Figure 8b, which is indicative of poor topics. DATM seems to improve these values.

C. QUALITY OF TOPIC MODELS

To assess the quality of the topics generated, we inspect some of the topics generated from the Airline dataset.[15] We randomly select 4 topics from the airline dataset namely *flight booking problems*, *late flight*, *cancelled flight* and *customer service issue*. We then consider all the topics generated by the techniques with the best matching values to the true topics. For illustration, we only present the top ten probable words found for each technique in Figure 9. We notice that, all techniques identify key words (in bold) which intuitively relate to the true topics. It is important to point out that while DATM improves, *reliability* and *effectiveness* of the topic models, it does not diminish the quality of the topic word representations.

[15]We choose the airline dataset because it has the best results across all techniques.

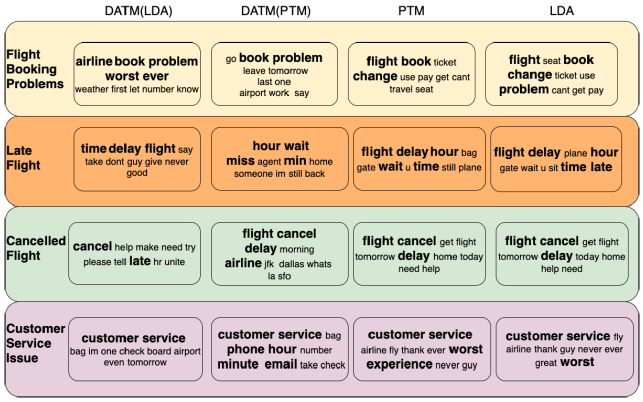


FIGURE 9. Quality of topics.

VII. CONCLUSION

Topic modelling is an important data mining technique which has broad applications in health, cyber-technology, business, and policy making. While several topic modelling techniques have been developed, their growing applications show some drawbacks in existing techniques. These include the lack of consistency in topics, challenges in dealing with mixtures of long and short documents, and failing underlying assumptions of existing techniques within specific contexts.

In this work, we presented a novel data transformation approach for improving the reliability and effectiveness of existing generative topic models. Our novel data transformation approach is driven by the observation that, topic

concentration within documents is critical for effective topic modelling. We have shown both theoretically and empirically that topic concentration within a document is related to the word co-occurrence. Therefore, by enhancing the word co-occurrence structure we can improve topic models. We modelled this as an optimisation problem and proposed an efficient solution DATM. DATM, has been demonstrated not only to be a powerful technique, but it can also be used in conjunction with other existing topic modelling techniques such as LDA and PTM.

In our future work, we aim to evaluate how DATM can be integrated into neural topic models to reduce their resource requirements whilst improving their effectiveness. In particular, by transforming the input data, we may be able to reduce the training time on word-embeddings as well as the requirement for large training datasets. We will also extend our work into a sequential scenario, by investigating updatable DATM for detecting topic evolution over time. This will have applications in the detection of changes in public sentiment and topical discussions which will have significant implications on law enforcement, business and recommendation systems.

REFERENCES

[1] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? Disruptive event detection using Twitter," *ACM Trans. Internet Technol.*, vol. 17, no. 2, pp. 1–26, 2017.

[2] J. Dupuy, A. Guille, and J. Jacques, "Anchor prediction: A topic modeling approach," in *Proc. Companion Web Conf.* New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1310–1318.

[3] N. Ramakrishnan, "'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani, Eds. New York, NY, USA, Aug. 2014, pp. 1799–1808.

[4] F. Valle, M. Osella, and M. Caselle, "A topic modeling analysis of TCGA breast and lung cancer transcriptomic data," *Cancers*, vol. 12, no. 12, p. 3799, Dec. 2020.

[5] D. M. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[6] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, Dec. 2020, Art. no. 101582.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[8] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," in *Proc. 31th Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 190–198. [Online]. Available: http://proceedings.mlr.press/v32/tang14.html

[9] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proc. 1st Workshop Social Media Anal.*, Jul. 2010, pp. 80–88.

[10] D. C. Zhang and H. W. Lauw, "Variational graph author topic modeling," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, A. Zhang and H. Rangwala, Eds. Washington, DC, USA, Aug. 2022, pp. 2429–2438.

[11] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," in *Proc. 29th Conf. Uncertainty Artif. Intell.* A. E. Nicholson and P. Smyth, Eds. Bellevue, WA, USA: AUAI Press, 2013.

[12] D. Alvarez-Melis and M. Saveski, "Topic modeling in Twitter: Aggregating tweets by conversations," in *Proc. 10th Int. Conf. Web Social Media*, Cologne, Germany, 2016, pp. 519–522.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential Twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, Feb. 2010, pp. 261–270.

[14] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in *Proc. 33rd Eur. Conf. Adv. Inf. Retr. (ECIR)*, 2011, pp. 338–349.

[15] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, p. 889.

[16] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation," in *Proc. 34th Int. Joint Conf. Artif. Intell.*, Q. Yang and M. J. Wooldridge, Eds. Buenos Aires, Argentina, 2015, pp. 2270–2276. [Online]. Available: http://ijcai.org/Abstract/15/321

[17] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, "Topic modeling of short texts: A pseudo-document view," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 2105–2114.

[18] X. Li, C. Li, J. Chi, and J. Ouyang, "Short text topic modeling by exploring original documents," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 443–462, 2018.

[19] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1445–1456.

[20] X. He, H. Xu, J. Li, L. He, and L. Yu, "FastBTM: Reducing the sampling time for biterm topic model," *Knowl.-Based Syst.*, vol. 132, pp. 11–20, Sep. 2017.

[21] X. Hu, H. Wang, and P. Li, "Online biterm topic model based short text stream classification using short text expansion and concept drifting detection," *Pattern Recognit. Lett.*, vol. 116, pp. 187–194, Dec. 2018.

[22] J. Huang, M. Peng, P. Li, Z. Hu, and C. Xu, "Improving biterm topic model with word embeddings," *World Wide Web*, vol. 23, no. 6, pp. 3099–3124, Nov. 2020.

[23] Q. Zhu, Z. Feng, and X. Li, "GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4663–4672.

[24] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, "Learning topics in short texts by non-negative matrix factorization on term correlation matrix," in *Proc. SIAM Int. Conf. Data Mining*, May 2013, pp. 749–757.

[25] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proc. World Wide Web Conf.*, P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis, Eds. Lyon, France, 2018, pp. 1105–1114.

[26] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: A simple but general solution for short and imbalanced texts," *Knowl. Inf. Syst.*, vol. 48, no. 2, pp. 379–398, Aug. 2016.

[27] F. Wang, R. Liu, Y. Zuo, H. Zhang, H. Zhang, and J. Wu, "Robust word-network topic model for short texts," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell. (ICTAI)*, San Jose, CA, USA, Nov. 2016, pp. 852–856.

[28] M. Jiang, R. Liu, and F. Wang, "Word network topic model based on Word2Vector," in *Proc. IEEE 4th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Bamberg, Germany, Mar. 2018, pp. 241–247.

[29] N. Zhong and D. A. Schweidel, "Capturing changes in social media content: A multiple latent changepoint topic model," *Marketing Sci.*, vol. 39, no. 4, pp. 827–846, Jul. 2020.

[30] A. Rortais, F. Barrucci, V. Ercolano, J. Linge, A. Christodoulidou, J. P. Cravedi, R. Garcia-Matas, C. Saegerman, and L. Svevnjak, "A topic model approach to identify and track emerging risks from beeswax adulteration in the media," *Food Control*, vol. 119, Jan. 2020, Art. no. 107435.

[31] B. Liu, P. Zhang, T. Lu, and N. Gu, "A reliable cross-site user generated content modeling method based on topic model," *Knowl.-Based Syst.*, vol. 209, Dec. 2020, Art. no. 106435.

[32] H. Chen, H. Yin, X. Li, M. Wang, W. Chen, and T. Chen, "People opinion topic model: Opinion based user clustering in social networks," in *Proc. 26th Int. Conf. World Wide Web Companion*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds. 2017, pp. 1353–1359.

[33] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Enhancing topic modeling for short texts with auxiliary word embeddings," *ACM Trans. Inf. Syst.*, vol. 36, no. 2, pp. 1–30, Apr. 2018.

[34] J. He, Z. Hu, T. Berg-Kirkpatrick, Y. Huang, and E. P. Xing, "Efficient correlated topic modeling with topic embedding," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, Aug. 2017, pp. 225–233.

[35] W. Liang, R. Feng, X. Liu, Y. Li, and X. Zhang, "GLTM: A global and local word embedding-based topic model for short texts," *IEEE Access*, vol. 6, pp. 43612–43621, 2018.

[36] L. Shi, G. Cheng, S.-R. Xie, and G. Xie, "A word embedding topic model for topic detection and summary in social networks," *Meas. Control*, vol. 52, nos. 9–10, pp. 1289–1298, Nov. 2019.

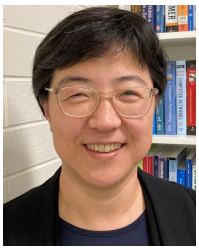
[37] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, and J. Zobel, Eds. Pisa, Italy, Jul. 2016, pp. 165–174.
[38] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "The dynamic embedded topic model," 2019, *arXiv:1907.05545*.
[39] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 439–453, Dec. 2020.
[40] L. Liu, H. Huang, Y. Gao, Y. Zhang, and X. Wei, "Neural variational correlated topic modeling," in *Proc. World Wide Web Conf.*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. Francisco, CA, USA, May 2019, pp. 1142–1152.
[41] T. Lin, Z. Hu, and X. Guo, "Sparsemax and relaxed Wasserstein for topic sparsity," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, J. S. Culpepper, A. Moffat, P. N. Bennett, and K. Lerman, Eds. Melbourne, VIC, Australia, Jan. 2019, pp. 141–149.
[42] T. Doan and T. Hoang, "Benchmarking neural topic models: An empirical study," in *Proc. ACL/IJCNLP*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. 2021, pp. 4363–4368.
[43] M. Zhang and J. Li, "A commentary of GPT-3 in MIT technology review 2021," *Fundam. Res.*, vol. 1, no. 6, pp. 831–833, Nov. 2021.
[44] Y. Yang, F. Wang, J. Zhang, J. Xu, and P. S. Yu, "A topic model for co-occurring normal documents and short texts," *World Wide Web*, vol. 21, no. 2, pp. 487–513, Mar. 2018.
[45] X. Li, A. Zhang, C. Li, L. Guo, W. Wang, and J. Ouyang, "Relational biterm topic model: Short-text topic modeling using word embeddings," *Comput. J.*, vol. 62, no. 3, pp. 359–372, Mar. 2019.
[46] Y. Liu and M. Piccardi, "Topic-based unsupervised and supervised dictionary induction," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 3, pp. 1–21, Mar. 2023.
[47] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–35, 2022.
[48] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
[49] H. M. Wallach, D. Mimno, and A. Mccallum, "Rethinking LDA: Why priors matter," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22, 2009, pp. 1973–1981.
[50] M. D. Hoffman, D. M. Blei, and F. Bach, "Online Learning for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.
[51] K. Bache, D. Newman, and P. Smyth, "Text-based measures of document diversity," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, I. S. Dhillon, Y. Koren, R. Ghani, T. E. Senator, P. Bradley, R. Parekh, J. He, R. L. Grossman, and R. Uthurusamy, Eds. Chicago, IL, USA, 2013, pp. 23–31, doi: 10.1145/2487575.2487672.
[52] C. Ricotta and L. Szeidl, "Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index," *Theor. Population Biol.*, vol. 70, no. 3, pp. 237–243, Nov. 2006.
[53] E. H. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, Apr. 1949.
[54] N. G. Hall and D. S. Hochbaum, "A fast approximation algorithm for the multicovering problem," *Discrete Appl. Math.*, vol. 15, no. 1, pp. 35–40, Sep. 1986.
[55] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman, 1979.
[56] S. Rajagopalan and V. V. Vazirani, "Primal-dual RNC approximation algorithms for (multi)-set (multi)-cover and covering integer programs," in *Proc. IEEE 34th Annu. Found. Comput. Sci.*, Nov. 1993, pp. 322–331.
[57] Q.-S. Hua, D. Yu, F. C. Lau, and Y. Wang, "Exact algorithms for set multicover and multiset multicover problems," in *Proc. Int. Symp. Algorithms Comput.* Cham, Switzerland: Springer, 2009, pp. 34–44.
[58] S. G. Kolliopoulos and N. E. Young, "Approximation algorithms for covering/packing integer programs," *J. Comput. Syst. Sci.*, vol. 71, no. 4, pp. 495–505, Nov. 2005.
[59] S. G. Kolliopoulos, "Approximating covering integer programs with multiplicity constraints," *Discrete Appl. Math.*, vol. 129, nos. 2–3, pp. 461–473, Aug. 2003.
[60] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Edinburgh, U.K., 2011, pp. 262–272. [Online]. Available: https://aclanthology.org/D11-1024/
[61] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proc. 10th Int. Conf. Comput. Semantics*, K. Erk and A. Koller, Eds. Potsdam, Germany: University of Potsdam, 2013, pp. 13–22. [Online]. Available: https://aclanthology.org/W13-0102/
[62] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Human Lang. Technol., Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, 2010, pp. 100–108. [Online]. Available: https://aclanthology.org/N10-1012/

MICHAEL BEWONG (Member, IEEE) received the Ph.D. degree from the University of South Australia. He is currently a Senior Lecturer of computing with Charles Sturt University, Australia. His research interests include the data science applications in cyber security, text analytics, data sharing, machine learning, and privacy. He has authored top tier articles in his research area.

JOHN WONDOH received the Ph.D. degree in computer and information science from the University of South Australia (UniSA), in 2018. He was a Research Fellow with Charles Sturt University. He is currently a Research Fellow with the Future Energy Exports CRC, UniSA. His research interests include machine learning, business process management, and interoperable analytics.

SELASI KWASHIE received the Ph.D. degree in computer and information science from the University of South Australia. He is currently a Senior Lecturer with the Artificial Intelligence and Cyber Futures Institute, Charles Sturt University. His works cover mining data dependencies, entity resolution, and access management. His research interests include database theory, data science, and enterprise systems security.

JIXUE LIU received the Ph.D. degree in computer science from the University of South Australia. He is currently an Associate Professor with the University of South Australia. He has published widely in databases and artificial intelligence. His research interests include machine learning, data analytics in texts and time series, integrity constraint discovery, entity linking, fairness computing, privacy in data, XML functional dependencies, and data integration and transformation.



LIN LIU received the bachelor's and master's degrees in electronic engineering from Xidian University, China, and the Ph.D. degree in computer systems engineering from the University of South Australia (UniSA). She is currently a Professor with UniSA. Her research interests include data mining, machine learning, causal inference, and bioinformatics.

JIUYONG LI (Member, IEEE) received the Ph.D. degree in computer science from Griffith University, Australia. He is currently a Professor with the University of South Australia. His research was supported by many Australian Research Council and industry funded projects. His research interests include data mining, machine learning, and bioinformatics.

MD. ZAHIDUL ISLAM is currently a Professor of computer science with the School of Computing, Mathematics, and Engineering, Charles Sturt University, Australia. His main research interests include data mining, classification, and clustering algorithms, missing value imputation, data cleaning and preprocessing, privacy preserving data mining, privacy issues due to data mining on social network users, and the applications of data mining in real life.

DAVID KERNOT received the Ph.D. degree in political science and international relations from the National Security College, Australian National University. In Australian National University, he examined linguistic markers of cognitive decline in people, including depression and anxiety and focused on identifying mathematical tipping points that might indicate self-radicalization in lone actor terrorists. He is currently with the Department of Defence, Defence Science and Technology Group.