Aggregated Topic Models for Increasing Social Media Topic Coherence

Stuart J. Blair[*], Yaxin Bi (✉)[†] and Maurice D. Mulvenna [‡]
School of Computing
University of Ulster at Jordanstown
Co Antrim, BT37 0QB, UK

Abstract

This research presents a novel aggregating method for constructing an aggregated topic model that is composed of the topics with greater coherence than individual models. When generating a topic model, a number of parameters have to be specified. The resulting topics can be very general or very specific, which depend on the chosen parameters. In this study we investigate the process of aggregating multiple topic models generated using different parameters with a focus on whether combining the general and specific topics is able to increase topic coherence. We employ cosine similarity and Jensen-Shannon divergence to compute the similarity among topics and combine them into an aggregated model when their similarity scores exceed a predefined threshold. The model is evaluated against the standard topics models generated by the latent Dirichlet allocation and Non-negative Matrix Factorisation. Specifically we use the coherence of topics to compare the individual models that create aggregated models against those of the aggregated model and models generated by Non-negative Matrix Factorisation, respectively. The results demonstrate that the aggregated model outperforms those topic models at a statistically significant level in terms of topic coherence over an external corpus. We also make use of the aggregated topic model on social media data to validate the method in a realistic scenario and find that again it outperforms individual topic models.

Keyword: Topic models, data fusion, topic coherence, ensemble methods, social media

1 Introduction

In the modern era of digital technology and with the advent of big data there is unrivalled access to masses of data that would have been unimaginable in the past. One of the challenges faced is how to extract the underlying information from these masses of data. A particular interesting development of the Internet has been the emergence of social networks, specifically microblogs. Many Internet users have been abandoning traditional methods of online communication such as blogs and newsgroups, in favour of social networks that enable microblogging, for instance, Twitter. These microblog platforms have enabled millions of users to quickly and concisely express opinions about anything from products to politics. For this

[*]Blair-S4@email.ulster.ac.uk
[†]y.bi@ulster.ac.uk (corresponding author)
[‡]md.mulvenna@ulster.ac.uk

reason these microblogs have become an invaluable source of information for many companies and institutions to gauge consumer opinion and help shape future product development or marketing campaigns.

Textual data is difficult to analyse due to their varied linguistic characteristics and semantics. A method that attempts to identify the underlying topical structure of textual data is topic models such as Latent Dirichlet Allocation (LDA) [1]. Topic models are a type of statistical and unsupervised model that can quickly discover the latent topics within large corpora of documents. When generating a topic model by LDA, a number of parameters have to be set, which have effects on the output of topics. Without prior knowledge of the corpus being modelled, setting a small number of topics will result in very broad topics that contain mostly common words and stopwords. However, when the number of topics is set too high, the models generated will be very granular and overfitted to the corpus. Other parameters, such as α and β Dirichlet prior values in LDA, are also crucial in generating the good quality of the model [2].

To prevent the very general or very specific topics that could be generated using non-optimal initial parameters, we propose a novel method of aggregating topic models. With our method, a user needs to define a set of parameters, and multiple topic models are generated using these parameters; the topics that are found to be similar amongst these models are then aggregated. The main contribution of this proposed approach is that the aggregated models are able to increase topic coherence. This has the advantage of allowing granular topics, which might only be produced in a model with many topics, to have a presence in a topic model that has a small number of topics and is more representative to the corpus as a whole. The proposed method is also advantageous as it requires no changes to the underlying topic model's generative method. This makes implementation of this method more convenient.

In order to examine the overall coherence of the topics generated in the new aggregated model, we propose to use two intrinsic and extrinsic measures. The first coherence measure allows us to assess the coherent extent that topic models accurately represent the content of a corpus used to generate the topic models based on word co-occurrences (WCO). The extrinsic measure allows for the examination of the generated topics against a corpus of general English documents to ensure that the topics are generally coherent in daily language, which is similar to how a human observing the topics would decide whether they are coherent or not. Moreover, a statistical significance test has been calculated to examine how the aggregated model is statistically different from the base model.

1.1 Research Goal and Contributions

The main goal of this research is to construct an aggregated topic model that produces topics with greater topic coherence than the base models used to create the aggregated topic model. In this sense, topic coherence refers to evaluation methods that analyse words in each topic to ensure that they would make sense together from a human-like perspective. This is in contrast to other statistical methods of topic model evaluation that may not necessarily reflect human evaluation. When generating topic models, a resulting model may produce very granular topics or it could produce very general topics populated with words common across all documents, which depends on how parameters are set in topic modelling. This research shows the theoretical framework of aggregating models, which is to combine several topic models in an ensemble-like approach with the goal of increasing overall topic coherence. The work has significant practical impact as it can be used directly in any topic modelling

systems to increase topic coherence, especially noisy domains, such as social media.

The aggregated topic model method has been evaluated in four experiments. The first two experiments makes use of a set of 2246 Associated Press articles that were used in the original LDA paper. The third experiment uses a set of 300,000 Tweets captured during the final Presidential Debate of the 2016 election campaign. Using these datasets a series of experiments were carried out by creating a series of topic models with varying parameters, such as α priors, β priors, and number of topics. Using these models aggregated topic models were constructed for each experiment, and the effect of altering the similarity thresholds in constructing the aggregated model was observed. When calculating topic coherence, an external corpus − the full set of English Wikipedia articles − was used, which is part of calculating coherence. The last experiment is a comparative study on the aggregated topic models with the Non-negative Matrix Factorisation with the same setting as in the first two experiments.

This research makes the following contributions:

1. The study introduces a novel method for aggregating multiple topic models in an ensemble-like way to create an aggregated topic model that contains topics with greater semantic coherence than individual models.

2. The aggregated model is constructed on the basis of cosine similarity and Jensen-Shannon divergence when exceeding a similarity threshold.

3. The experiment results show that the aggregated model outperforms standard topic models and Non-negative Matrix Factorisation at a statistically significant level in terms of topic coherence.

4. The aggregate topic model has been evaluated on social media data to validate the method in the third 2016 American Presidential Debate scenario, demonstrating the competitiveness of the proposed method in real work applications.

The remainder of this research article is structured as follows. Section 2 provides a literature review of work related to this area and an overview of topic coherence. In Section 3, an overview of standard topic models, such as LDA, is given. Next, in Section 4, the theoretical framework of aggregated topic models is introduced, including analysis of similarity measures, and the full algorithms for finding similar topics and combining these similar topics. Section 5 contains the three experiments conducted, this includes initial parameters used and tuning the aggregated model for each of the experiments. In Section 6 the results of these experiments are evaluated and the significance of each result is presented. Finally, in Section 7, the work is concluded and a discussion of the aggregated topic models is presented.

2 Related Work

The main difference between the current work and the existing studies discussed in this section is that the method presented in this research is focused on the aggregation process of topics after the models have been generated. This is different from other methods that are often used in aggregating text by various factors, such as author or hashtag, before the model is generated in order to create larger documents [3]. The advantage of the method described here is that this method does not rely on arbitrary manipulations of the model structures

of underlying topic models or input data, thereby producing a less complex model structure that requires no additional context specific information such as seed words or part-of-speech tagging. The remainder of this section reviews how related approaches work in the context of an ensemble-like method and topic modelling.

There has been some studies in respect of combining multiple topic models and topic model ensembles. One method proposed in the area of aggregating multiple topic models is not directly related to textual data but is used on medical data for predicting a patient's risk of disease [4]. In that ensemble, component models were trained with respective datasets, including poverty level, gender, age and race. The topics discovered in each model are then aggregated into a single matrix ($\phi_{k=1,\ldots,K}$), where ϕ_k is a topic and K is the total number of topics. This aggregated topic matrix is then used to derive a distribution ($\theta_{d=1,\ldots,D}$) over new patients to predict the disease risk they have, where θ_d is a topic distribution from a set of data D.

Another significant piece of study on combining topic models is to concern the utilisation of document-word tuples co-occurring in documents and their assigned topics [5]. That method assumes an element vector T where each element is a topic assignment of a document-word tuple (d_i, w_i), word w_i from document d_i. The corpus is then divided into a set of sub-corpora, where each sub-corpus is represented by T_i that can be merged into one vector T. That is then used to derive topic distributions over documents. That method has been introduced as both LDA and latent semantic analysis (LSA) ensembles. The evaluation conduced on both real and synthetic data demonstrates that the LDA ensemble outperforms the LSA in terms of perplexity, however, the performance of the LSA ensemble is better than that of the LDA in terms of efficiency.

A few studies have been conducted on the application of classic ensemble methods to topic models. The boosting ensemble method has been applied to generating topic models with good results and generalisation ability as LDA has the ability to map features to topic space rather than word space [6]. Another interesting work is to integrate a supervised component into LDA, for instance, through defining one-to-one correspondence between LDA's latent topics and classification labels [7], or incorporating a supervised hidden Markov model into LDA, resulting in a further enhancement of the boosting method for topic models [14].

A method for generating a pachinko allocation model (PAM) has also been proposed in [15]. That method utilises correlations between topics to create a tree-like structure, in which leaves are words from the vocabulary and interior nodes represent the correlation between the child nodes. Another similar work is related to hierarchical topic models [16]. In that model a document is generated by creating a path from the root node to a leaf node, where the root is a very general word and the further the tree is traversed, the more specific words along the path get. The method assumes that the corpus can be organised into a hierarchical structure of topics and also has the advantages of easily accommodating the growth of corpora and having nonparametric priors. That work is similar to the study published in [17].

Studies have also been conducted to bring together two separate LDA models in a two-fold method, one model for aspects and one for sentiment [18]. That method was also extended further to allow for multiple aspects to be present in a sentence as the initial version assumed only one aspect per sentence [19]. The extended two-fold model which used the Jaccard index to identify the number of aspects in a document outperformed the original two-fold model, producing more relevant results for the documents modelled but at the expense of there being slightly more results produced.

The idea of using similarity measures to compare aspects of topic models has been studied

in the past years. Cosine similarity has been used to measure the similarity among the top words in topics generated using two different methods [20], while Jensen-Shannon divergence has been used to compare the similarity of topic distributions over documents [21]. The work in this paper differs from those previous works by using the similar topics by combining them to produce new topics; the previous studies used the similarity measure as an evaluation tool. In combining the topics, new topics are generated that take aspects from different models; this results in a new set of topics that may be more coherent on the whole. Taking aspects from different models refers to how words that appear in specific topics may be introduced into a general topic to change the underlying topic structure and increase that topic's coherence.

A method for modelling consensus topics across a range of contexts has been proposed [22]. That method implements a co-regularisation framework to create pseudo-documents using a centroid-based regularisation framework to make topics from different contexts agree with each other based on a set of general topic distributions. It allows for context specific topics to be bridged together by the general topics. That method outperformed the standard LDA in terms of topic coherence when modelling Tweets.

A system called collaborative topic regression has also been proposed [23]. The proposed system has been used with social trust ensembles to provide a method for social media content recommendation. This system uses a modified version of LDA and takes into account social influence from a user's trusted friends for recommending content. It performs this function by using a similarity metric to assess friends' similar interests to recommend content.

Most of topic modelling so far has an underlying assumption, that is each topic discovered contains a characteristic anchor word and it does not appear in the other topics. In [48], the authors proposed an anchor-free topic model based on a matrix of word-topic probability mass functions and a matrix of the topic-topic correlation derived via minimizing the determinant of the topic-topic correlation matrix. Likewise Dirichlet Multinomial Mixture (DMM) topic model assumes that each piece of short texts is generated by a single topic. To make DMM to fit in more general cases, the utilisation of general word semantic relations in word embeddings has been proposed in the topic inference process [47]. The word embeddings is incorporated into the topic inference process by the generalised GPU model, which effectively accesses background knowledge external to the given short text corpus and tackles the data sparsity issue.

More recently artificial neural networks have become state-of-the-art methods for building language modelling on textual corpora. Particularly Long Short-Term Memory (LSTM) neural network has the advantage of discovering both long and short patterns from data and alleviate the problem of vanishing gradient in training a recurrent neural network (RNN) [44]. In opposed to LSTM, a RNN-based topic model, called TopicRNN, has been proposed in [46]. TopicRNN captures local (syntactic) dependencies using an RNN and global (semantic) dependencies by latent topics in an end-to-end learning fashion, instead using pre-trained topic model features as an additional input to the hidden states and/or the output of the RNN.

Three different neural structures for constructing topic distributions have been evaluated in [45], including the Gaussian Softmax distribution, the Gaussian Stick Breaking distribution, and the Recurrent Stick Breaking process, all of these structures are conditioned on a draw from a multivariate Gaussian distribution. The combination of neural structures and conventional probabilistic topic models provides parameterisable distributions over topics, thereby allow the models to be trained by backpropagation in the framework of neural variational inference, scaled to large data sets, and easily conditioned on any available contextual information.

A similar method to our approach is a Non-negative Matrix Factorisation (NMF) based method [9, 24]. That method, however, does not directly deal with improving coherence but rather stabilizing the output of the NMF that has a tendency to be more unstable than probabilistic methods like LDA. In that sense, stability almost equates to model reproducibility. Similar to the method proposed in this paper the model begins with a series of base models being created, the outputs are then combined into a single topic-term matrix (similar to work presented previously in [5, 8]) before non-negative matrix factorization is performed again.

The closest piece of previous study to our work is the self-aggregation process for short text in the model generation stage of a topic model [25]. That process naturally integrates clusters of text before the topic modelling process begins. Importantly, it has a fundamental assumption that each document only contains one latent topic. This allows for the creation of larger pseudo-documents to be used in the generative procedure of topic models.

3 Topic Coherence Measures

Topic coherence can be defined as how interpretable a topic is based on the degree of relevance between the words within the topic itself. The topic coherence measures used in this work aims to evaluate the quality of topics from a human-like perspective. Considerable studies have been carried out in the evaluation of statistical topic models, however, it has been found that those methods are not always reflective of how a human views the topics [26]. Consequently, it was revealed that metrics based on word co-occurrences and a Mutual Information Approach are more representative of how a human would approach topic evaluation [27]. It is for this reason that we use these word co-occurrence and the mutual information approach in this work.

There are two main ways to evaluate the quality of topic models. Firstly, a statistical measure can be used to estimate the probability of a series of held out documents using a trained model. The disadvantage of evaluating models in that manner is that it does not account for the human interpretability of the topics generated. For instance, if a model overfits the data, the topics may not make much sense when a human examines them but the probability of held out documents may be quite high [26]. A separate set of evaluation methods have also been proposed, which use the actual topics generated by the models and assess if the words within the topics belong together in a coherent manner.

The two main approaches of handling the data when evaluating a model statistically are to either train on some of the data and then get the probability of heldout documents given the latent topics discovered during training of the model; or to split each document in half and train the model with the first half of the document and then get the probability of the second half of the document given the latent topics discovered from training on the first half of the documents, this is known as document completion evaluation [28].

Another evaluation method for topic models is the empirical likelihood measure; that evaluation method is used in the popular topic modelling library, MALLET. A different approach present in the MALLET library is left-to-right evaluation that produces similar results to the empirical likelihood but is less prone to overestimation [29]. Additionally, an evaluation method of utilising the harmonic mean [30] has been proposed but it has been discovered to drastically overestimate the probability of heldout documents [28]. It should be noted that despite widespread use due to ease of implementation, that method has been a source of criticism, even from its original authors [30].

Although those methods are useful for getting an insight into how the model performed as a whole, in this work we do not focus on the statistical methods for evaluating topic models. We are more interested in the coherence of the latent topics that the models output. Topic coherence measures are used to quantify the similarity degree of the latent topics discovered by a topic model from a human-like perspective to identify a high degree of semantic coherence of topic models. In simple terms, topic coherence assesses such a kind of topic extent through computing word co-occurrences and mutual information, which could reflect how people would perceive this problem. Although there have been many obscure types of topic coherence measures and unifying topic coherence measures proposed [31], this research focuses on the measure which are most popular among the literature in the field, which are detailed in the following section.

3.1 Extrinsic Topic Coherence Measures

Extrinsic coherence measures can be split into direct and indirect approaches. Direct approaches require the use of an external corpus to calculate the observed coherence whereas indirect approaches will use some other test, such as identifying an intruder word in a topic. In all cases, extrinsic coherence is a useful measure of how coherent a topic would be in daily language without any context.

One such direct approach to gauging topic coherence is to utilise the pointwise mutual information (PMI) [27]. This way is extrinsic in nature as it computes PMIs over the top N words in a sliding window with an external corpus such as a Wikipedia dump file (a plain text version of every article present on Wikipedia). The calculation can be seen in (1).

$$PMI(w_i, w_j) = log \frac{p\,(w_i, w_j)}{p\,(w_i)\,p\,(w_j)} \tag{1}$$

In (1) $p(w_i, w_j)$ is the probability of two words co-occurring in a document, as seen in (2); $p(w_i)$ and $p(w_j)$ is the probability of word w_i and word w_j occurring in the document, respectively, as seen in (3). In (2) and (3) D_{ext} represents the number of documents in the external corpus that contain either one or both words, depending on the calculation. The external corpus is a set of general documents which allow for the calculation of coherence based around general English usage.

$$p(w_i, w_j) = \frac{D_{ext}(w_i, w_j)}{D_{ext}} \tag{2}$$

$$p(w_i) = \frac{D_{ext}(w_i)}{D_{ext}} \tag{3}$$

Research on the PMI has shown that it tends to have a high degree of correlation with a human evaluator's assessment of topic coherence, however, a normalized version of the PMI approach is able to produce an even higher level of correlation with human judgement (NPMI) [32]. It has been shown that for social media data, using a large collection of random Tweets as external data and a modified version of PMI finds high coherence, however, if such a large collection of Twitter data is not available then using Wikipedia as external data with a PMI based approach is most closely aligned with human judgement [33]. The use of the PMI with a general external corpus allows for calculating how frequently words in the topic occur

together. Because of a general corpus, it can be loosely seen as how a human would interpret the topics from everyday language.

Another method for calculating topic coherence is referred to as distributional semantics [34]. This method uses a similarity measure such as cosine similarity to evaluate a vector of the top N words in a topic and then weight this result using the PMI between the words. This method also introduces a variable to put more emphasis on word relationships with a higher PMI.

Another metric that allows for a measure of topic coherence is normalised Google distance (NGD) [35]. Unlike PMI, which is a measure of information, NGD allows for the similarity or dissimilarity of two words/terms to be assessed using the number of pages returned by a Google search as the probability of words appearing on their own or co-occurring together; this is then used to determine the distance between the terms. Although the method has Google in its name, it works with any indexed database or search engine, for example, Yahoo or Wikipedia. Due to the reliance of an external data source, NGD is an extrinsic coherence measure. Using search engines or databases allows for a better idea of term coherence in everyday use of language. The calculation for NGD can be seen in (4), where $f(x)$ and $f(y)$ is the number of pages containing search term x and y, respectively; $f(x, y)$ is the number of pages where both x and y occur together; N is the number of pages indexed by the search engine or database; and κ is a smoothing factor. The bounds of NGD is $0 \leq \ NGD \leq \infty$, where 0 indicates the two terms only occur together and are therefore coherent, and as the NGD value approaches ∞ the two terms are further apart.

$$NGD(x, y) = \frac{\max\left(\log f(x), \log f(y)\right) - \log f(x, y)}{\log N\kappa - \min\left(\log f(x), f(y)\right)} \tag{4}$$

Equation (4) shows how to find the NGD between two terms. Equation (5) shows the process to find the average coherence for the top N terms in a topic, where K is the number of terms to be considered in the topic.

$$N\bar{G}D = \frac{\sum_{i}^{K}\sum_{j}^{K} NGD(x_i, y_j)}{K} \tag{5}$$

An example of an indirect extrinsic approach to calculating topic coherence is to assess the detection of intruder words in a topic [32], this is essentially the inverse of previous work to detect the word that was most representative of a topic [36]. This method works by finding association features for the top N words in a topic (intruder word inclusive) and then combining the results using SVM^{rank} (support vector machine) to discover the intruder words. It was discovered that this method achieves a high level of correlation with human evaluators. The disadvantage of this method is that it requires manual placing of intruder words in the topics.

3.2 Intrinsic Topic Coherence Measures

Intrinsic coherence measures show how well a topic represents the corpus from which it was modelled. Intrinsic measures utilize word co-occurrences in documents from the corpus used to train the model [37]. The feature of the intrinsic method allows us to better judge the coherence of a topic based on training documents. The scale of the measure is in a range of $0 - 1$, if a measuring value is closer to 1, that means that the model has correctly identified

words that co-occur frequently in documents as a topic, but this cannot guarantee that they make semantic sense or that they are interpretable by a human; it simply means that the topics represent the data known to be in the corpus [38]. Using this method allows for the identification of poor topics for which word intrusions tests do not account. Given the top N words in a topic, the word co-occurrence can be calculated as seen in equation (6) [37].

$$WCO = \sum_{j=2}^{N} \sum_{i=1}^{j-1} log \frac{P(w_j, w_i)}{P(w_i)} \tag{6}$$

4 Standard Topic Models

Standard topic modelling was originated with LSA, however, in the context of an information retrieval task a standard LDA is often referred as latent semantic indexing [39]. LSA utilises a document-term matrix along with a singular value decomposition method to find similar documents. But it has a fundamental assumption that words which frequently co-occur are relevant to one another. Two notable disadvantages of LSA are that the model is based on the bag-of-words representation and that it struggles with polysemy. That means that the order of words in documents is ignored and that it cannot distinguish between the different meanings of a single word. For example, *crane* can refer to both a bird as well as a piece of construction machinery.

LDA is capable of eliminating the polysemy difficulties through incorporating a probabilistic element to the model but it still has the disadvantage with the bag-of-words model, which is not able to capture sentence structures when creating the model [1].

In this study, the topic model that will be utilised when creating the base models in experiments is LDA. LDA is a generative probabilistic model that finds latent topics in a collection of documents by learning the relationship between words (w_j), documents (d_i), and topics (z_j). The data used by an LDA model is input in bag-of-words form, word counts are preserved but the ordering of the words is lost. The only observed variable in the model are the words w_j, everything else is a latent variable. The generative process for document d_i assumes the following:

- There is a fixed number of topics T.
- Each topic z has a multinomial distribution over vocabulary ϕ_z drawn from Dirichlet prior $Dir(\beta)$.
- $i \in \{1, \ldots, M\}$ where M is the number of documents in the corpus.
- $Dir(\alpha)$ is the document-topic Dirichlet distribution.

The following is the generative process for document d_i:

1. Choose $\theta_i \sim Dir(\alpha)$.
2. For word $w_j \in d_i$:
   (a) Draw a topic $z_j \sim \theta_i$.
   (b) Draw a word $w_j \sim \phi_{z_j}$.

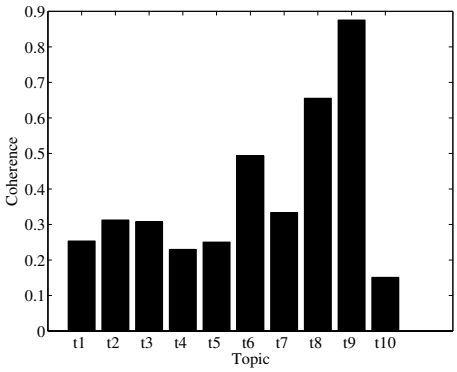


Figure 1: Topic coherence for a standard topic model

It is important to have a baseline result to compare with the results of aggregated topic models. Therefore a standard LDA topic model is created and the coherence of its topics is assessed. Specifically, the dataset used for creating a baseline model is a set of Associated Press articles along with the settings of 10 topics, 2,000 iterations of Gibbs sampling, and priors of $\alpha = 50/t$ and $\beta = 0.01$ as used in [40]. The coherence measure used is PMI with an external corpus of the English Wikipedia.

Figure 1 shows the coherence for each topic in the base topic model created by LDA. The mean coherence is 0.386 which is quite low. Topics t_8 and t_9 have a coherence far higher than the average, and topic t_{10} has a coherence far below the average. These results will serve as a baseline for comparison in later experiments.

5 Aggregated Topic Models

Prior to generating a topic model by LDA, a set of parameters need to be defined. On the one hand if setting a small number of topics, LDA could produce very broad topics, whereas if the number of topics is set too large, the topics generated could be very granular and overfitted to the corpus. However on the other hand, if one of these parameters results in very different topics to other models' topics, they are unlikely to be combined with other topics, thereby making them not have an effect on the aggregated topics.

In order to help alleviate the problem of very general or very specific topics that could be generated using non-optimal initial parameters, we propose a novel aggregating method for combining topic models. This method allows a user to define a set of different parameters and with them to generate multiple topic models. The topics that are found to be similar amongst these models are then aggregated.

One of the main problems with aggregating topic model outputs in an ensemble-like style is that unlike conventional classifier ensembles that have a finite set of possible classes $(C_1, \ldots, C_n)$, topic models have an infinite number of topic outputs that are unknown until the models have been created. For this reason, our proposed aggregation method is able to adjust to the multitude of outputs it may face.

In the area of text analysis there exists many methods for measuring similarity, such as cosine similarity, Pearson correlation coefficient, and the Jaccard index.

The Jaccard index can be used to find similar topics by simply calculating the similarity coefficient between the top N words in two given topics. A high value from the result of the

Jaccard index indicates that there is indeed some similarity between the topics. However, the downside is that a threshold for similarity needs to be set via introspection as there is no fool proof method of statistically assessing a similarity threshold.

Previously there has been research into using Jensen-Shannon divergence and Kolmogorov-Smirnov divergence to assess the similarity of topic probability distributions [41] within the topic model's ϕ distribution, which is the word distribution in each topic.

The Kolmogorov-Smirnov test [42], specifically in this work the two-sample Kolmogorov-Smirnov test [43] uses the method shown in (7).

$$D_{n,m} = \sup_x |F_n(x) - F_m(y)| \tag{7}$$

Testing $D_{n,m}$ (where n and m are the sizes of two distributions x and y, and F_n and F_m are the empirical distributions of the x and y values, respectively) allows for the evaluation of a null hypothesis that states that x and y are samples from the same underlying distribution. Proving that the null hypothesis is true allows for the assumption that two topics are very similar. Despite the usefulness of using a two-sample Kolmogorov-Smirnov test in this situation, it has been decided that it is not a viable method for finding similar topics. Although this decision may seem contradictory to what has been discussed, some initial tests using the two-sample Kolmogorov-Smirnov test gave disappointing results due to the two-sample Kolmogorov-Smirnov test needing a critical value to be calculated. When this critical value was calculated for the experiments, it resulted in the need for an exceptionally high similarity value between the two distributions, whereas other methods allow for more flexibility in setting the similarity threshold.

To perform the aggregation cosine similarity and Jensen-Shannon (JS) divergence will be used to assess the similarity of topics' ϕ distributions. The ϕ distribution (φ) is a $T * V$ stochastic matrix where each row is a topic (T) and each column is a non-zero real number probability for a word from the vocabulary (V). Both methods will then be evaluated for performance.

The cosine similarity has more flexibility in setting a similarity threshold and is also not invariant to shifts in input as opposed to measures such as Pearson correlation coefficient which is invariant to input shifts. The upper and lower bounds of the cosine similarity are 1 for complete similarity and 0 for complete dissimilarity. The process for aggregating topics with the cosine similarity is described as follows. Firstly, the user needs to define a set of parameters that will be used to generate the base models. A threshold for similarity will then have to be set using a grid search with training data sets to see which threshold can produce the most coherent topics on average. Although this may be seen as computationally expensive, on modern hardware these parameter tuning processes using a subset of training data are relatively quick to conduct and easy to run in parallel. Each topic from each of the other models is then compared to the base topics in a pairwise way in order to examine their similarity. If the cosine similarity of the two topics is above the threshold they will be then combined, which is to combine the similar topics via calculating the mean probability of each word in the ϕ distributions. It should be noted that the number of topics in the base model does not increase or decrease, nor does the number of words in the topic as the alphabet for each model is the same.

Equation (8) shows the combination process based on the cosine similarity where $\hat{\varphi}_k$ is an aggregated topic, n is the number of similar topics, M is the number of models, T_i is the

number of topics in a model, $\varphi_{(i,j)}$ is the ϕ distribution for topic T_j in model M_i, φ_x is the xth ϕ distributions from the base model, and γ is the similarity threshold.

$$\hat{\varphi}_k = \frac{\sum_{i=1}^{M}\sum_{j=1}^{T_i}}{n} \begin{cases} \varphi_{(i,j)}, & if \quad \frac{\varphi_{(i,j)} \cdot \varphi_x}{\|\varphi_{(i,j)}\| \|\varphi_x\|} \geq \gamma \\ 0, & otherwise \end{cases} \tag{8}$$

JS divergence allows for the symmetric measurement of similarity between distributions. Using the base 2 logarithm, the JS divergence has the bounds $0 \leq D_{JS}(P \parallel Q) \leq 1$ where 0 indicates complete similarity and 1 is complete dissimilarity. The JS divergence is a symmetrised and smoothed version of Kullback-Leibler (KL) divergence, using the average KL divergence for each ϕ distribution to the average of each ϕ distribution. It is shown in equation (9) where P and Q are distributions and M is the average of distributions P and Q.

$$\begin{gathered} D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \\ M = \frac{1}{2}(P + Q) \end{gathered} \tag{9}$$

The process for using JS divergence to create aggregated topics can be seen in equation (10) using the same notation as equation (8). The main difference is that the JS divergence result should be $\leq \gamma$ as opposed to cosine similarity where the result should be $\geq \gamma$.

$$\hat{\varphi}_k = \frac{\sum_{i=1}^{M}\sum_{j=1}^{T_i}}{n} \begin{cases} \varphi_{(i,j)}, & if \quad D_{JS}(\varphi_{(i,j)} \parallel \varphi_x) \leq \gamma \\ 0, & otherwise \end{cases} \tag{10}$$

The process for aggregating topic models can be seen in Algorithm 1. In this algorithm K is the set of aggregated topics, T is the set of base topics, SW is the sliding window size, γ is the similarity threshold, and $\hat{\varphi}$ is an aggregated topic.

5.1 Choosing Similarity Threshold

An important aspect of the proposed method for aggregating topic models is the choice of similarity threshold. The overall problem attempting to be solved can be viewed as optimising the semantic topic coherence by searching the optimal similarity threshold and sliding window size. The sliding window size is directly related to measuring coherence as it sets the window size for calculating word probabilities. For example, if the word co-occurrence probability for word w_i and word w_j is calculated using a sliding window of size 50 words, then as long as the words occur at least once in the 50 word window it will count as the words having co-occurred, irrelevant as to whether they are in different sentences or paragraphs. However, if a lower window size such as 10 is used, it is stricter as it limits where the words can co-occur. This allows for more confidence that the words actually occurred together in the same context.

In this paper a grid search will be used over a set of similarity thresholds and a set of sliding window sizes. A small subset (10%) of the full dataset will be used for searching the optimal values. The grid search will then allow for topics to be aggregated and the coherence calculated using the Cartesian product of the set of similarity thresholds and set of sliding window sizes. For example, if the set of similarity thresholds $Y = \{0.1, 0.2, 0.3\}$ and sliding

Algorithm 1 Aggregating Topic Models

procedure AGGREGATETOPICS($T, SW, \gamma, \hat{\varphi}, K$)
  $K = \varnothing$
  for $t_1 \in T$ **do**
    for $t_2 \in T$ **do**
      if $t_1 \neq t_2$ **then**
        if $metric = CS$ **then**
          if $\frac{\varphi_{t_1} \cdot \varphi_{t_2}}{\|\varphi_{t_1}\| \|\varphi_{t_2}\|} \geq \gamma$ **then**
            $K \cup \hat{\varphi}$
          end if
        else if $metric = JS$ **then**
          if $D_{JS}(\varphi_{(i,j)} \parallel \varphi_x) \leq \gamma$ **then**
            $K \cup \hat{\varphi}$
          end if
        end if
      end if
    end for
  end for
end procedure

window sizes $Z = \{10, 20, 30\}$, then the set of parameters generated by the Cartesian product $Y \times Z$ will be tested.

The algorithmic version of the process for choosing the similarity threshold and sliding window size is visible in Algorithm 2 where Γ is the set of similarity thresholds to be tested, SW is the set of sliding window sizes to be tested, opt_{sw} is the current optimal sliding window size, opt_{γ} is the current optimal similarity threshold, max is used to track the maximum coherence found, T is the subset of the dataset used for testing, $\hat{\varphi}$ is an aggregated topic, and K is the set of aggregated topics.

Although methods such as Bayesian optimisation can be used to optimise parameters, it is unnecessary for this task which, due to its nature, can be easily parallelised regardless suffering from the curse of dimensionality. This makes grid search a feasible option without overcomplicating the problem by using more complex methods.

6 Experiments

This section details the four experiments performed using the aggregated topic model. Experiments one and two were experiments designed to show the feasibility of aggregated topic models and prove their effectiveness when different topic model parameters were changed. Experiment three shows a real world application of the aggregated topic model. In this experiment we applied the aggregated topic model to Tweets about the third presidential debate. The last experiment is to compare the aggregated model with the algorithm Non-negative Matrix Factorisation with the same setting as in Experiments 1 and 2.

In order to show the effectiveness of aggregated topic models two initial experiments were performed. A number of variables need to be decided on before running a topic model, including the number of topics, the α Dirichlet prior, and the β Dirichlet prior. In these

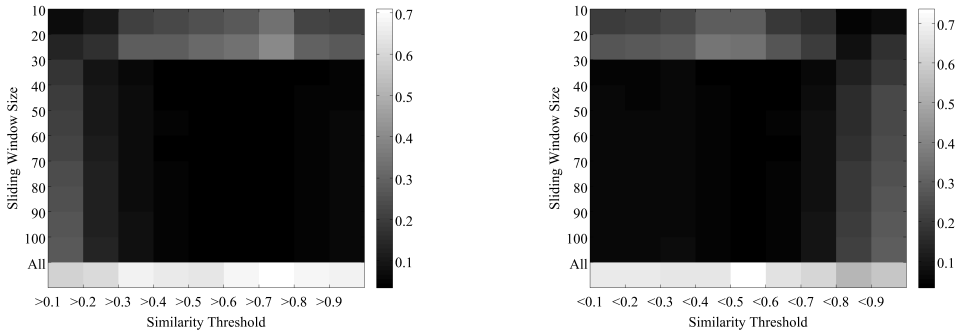
Algorithm 2 Similarity Threshold Selection

procedure SIMILARITYTHRESHOLD($\Gamma, SW, opt_{sw}, opt_{\gamma}, max, T, \hat{\varphi}, K$)
  $max = 0$
  $optSim = 0$
  $optSW = 0$
  for $\gamma \in \Gamma$ **do**
    for $s \in SW$ **do**
      $K = \varnothing$
      for $t_1 \in T$ **do**
        for $t_2 \in T$ **do**
          if $t_1 \neq t_2$ **then**
            if $metric = CS$ **then**
              if $\frac{\varphi_{t_1} \cdot \varphi_{t_2}}{\|\varphi_{t_1}\| \|\varphi_{t_2}\|} \geq \gamma$ **then**
                $K \cup \hat{\varphi}$
              end if
            else if $metric = JS$ **then**
              if $D_{JS}(\varphi_{(i,j)} \parallel \varphi_x) \leq \gamma$ **then**
                $K \cup \hat{\varphi}$
              end if
            end if
          end if
        end for
      end for
      if $\bar{K} \geq max$ **then**
        $opt_{\gamma} = \gamma$
        $opt_{sw} = s$
      end if
    end for
  end for
end procedure

experiments, models with differing variables were created and their outputs aggregated to see if it can increase topic coherence.

Each experiment involved the use of LDA with 2,000 iterations of Gibbs sampling to generate the topic models, determination of the similarity threshold, and comparisons of how the aggregated model competes with the base models used to create the aggregated model. The topic coherence test makes use of the extrinsic PMI. The reference corpus used for the extrinsic test is the English Wikipedia. This corpus has over five million articles and the average of article length in the Wikipedia dump was 133.07 words. Therefore it is a good reference of general English language for comparison. The assessment of the intrinsic coherence test has also been conducted to measure the degree to which topics capture the structure of the underlying corpus.

The data used to generate the models in the following experiments is a set of Associated Press articles used in the eponymous LDA paper, and supplied with David Blei's lda-c



(a) Cosine Similarity (b) Jensen-Shannon Divergence

Figure 2: Average topic coherence for aggregated models using different similarity thresholds and sliding window sizes for models with different numbers of topics

package[1]. The corpus contains 2246 documents and 34977 tokens after removal of stopwords.

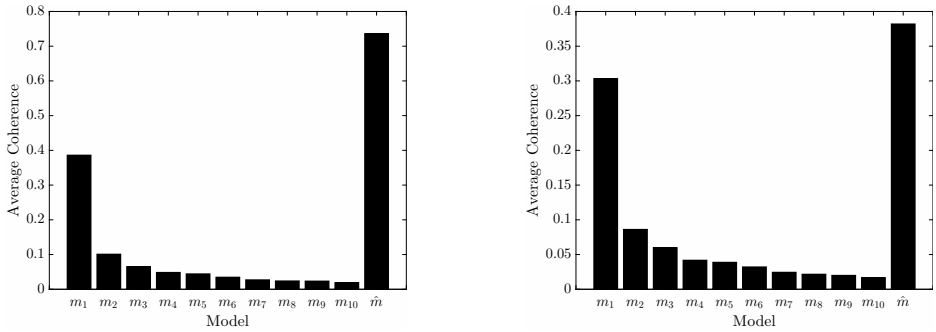
6.1 Experiment 1: Models with Different Topic Numbers

The first experiment consists of creating ten models each with a different number of topics $T = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and an α prior of $50/|T|$. Using different numbers of topics allows for representations of the corpus at different granularity levels, for example, 10 topics provides very general topics overview and 100 provides a very specific set of topics. An aggregated model that combines the similar topics from multiple models using different numbers of topics could produce a more coherent set of topics due to the output containing general topics complemented with more specific topics.

The first step in generating an aggregated model is to choose the similarity threshold at which similar topics will be combined. A grid search is used on small development sets of data at different similarity thresholds, starting at 0.1 and increasing to 0.9 in increments of 0.1. The sliding window size can also be changed at intervals of 10 words to 100 words, as well as using the whole document as the window. The results of these grid searches are presented in Figure 2a and Figure 2b for cosine similarity and Jensen-Shannon divergence, respectively. This experiment shows the optimal similarity threshold according to PMI is > 0.7 for cosine similarity and < 0.5 for Jensen-Shannon divergence. Note that cosine similarity thresholds are in the form $> n$ as the value for complete similarity is 1, whereas Jensen-Shannon divergence thresholds are $< n$ as complete similarity is 0. This shows that Jensen-Shannon divergence is more lenient in the topics that it aggregates, resulting in many more similar topics being combined, meanwhile allowing for the combination of general and specific topics. Cosine similarity has a higher similarity threshold meaning that not as many topics will be combined, however, Jensen-Shannon divergence achieves a higher topic coherence.

Following the tuning experiments, the full experiment was run using Jensen-Shannon divergence as the similarity measure and a similarity threshold of < 0.5, the results of this experiment for extrinsic and intrinsic coherence tests can be seen in Figures 3a and 3b, respectively. It should be noted that when the coherence of base models and aggregated models are compared, the same sliding window size is used for each model.

[1]Available at: `http://www.cs.columbia.edu/~blei/lda-c/`



(a) Extrinsic Coherence

(b) Intrinsic Coherence

Figure 3: Average intrinsic and extrinsic coherence for topics in the base models and aggregated model for models with different numbers of topics

In this experiment m_1 is the base model, $m_2 - m_{10}$ are the other models to be compared, and $\hat{m}$ is the aggregated model. Any model can be the base model, the fact that m_1 was chosen in this experiment is arbitrary. Also, if a different model was chosen as the base, the same topic similarity comparisons would be made; the only difference of using m_1 over the other models in this case is the number of topics in the final aggregated model. As Figure 3a shows, the aggregated model has an extrinsic PMI value of 0.75, this is much higher than any of the model used to create it. This shows that the aggregated model's topics are much more coherent based on general English language. The aggregated model also has the highest intrinsic coherence. This means the aggregated model's topics have been complemented with additional relevant topic words leading to topics that are more representative of the corpus.

This experiment resulted in some noticeable difference between the base model topics' top words and the aggregated model's top words. A comparison between the base model and aggregated model is visible in Tables 1 and 2, respectively. In t_1 the aggregated model has additional words including "Nicaragua" and "Contra"; this supplements the words from the base model, "united" and "states". It would be logical to connect these words through the Nicaraguan Revolution, when the United States supported the Contra forces in a campaign against the Sandinista National Liberation Front. Another major change can be seen in t_7 where the aggregated model contains more words about medical research and disease, whereas the base model includes some less relevant words such as "children", "percent" and "space". Additionally, t_8 sees the addition of the words "index" and "exchange"; this makes it more obvious that this topic is about stock markets and finance. The aggregated model also allows for more subtle changes such as the addition of Jackson in t_6, which refers to Jesse Jackson, the main opponent of Michael Dukakis in the 1988 Democratic presidential primaries.

6.2 Experiment 2: Models with Different Alpha Priors

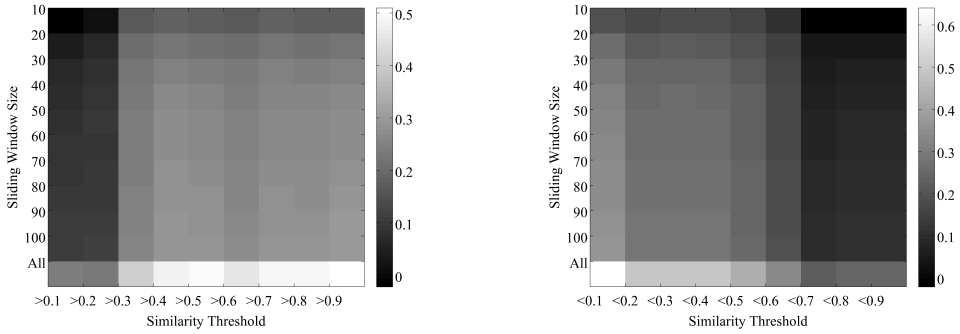
The second experiment consists of creating ten models each with a different α Dirichlet prior value $\alpha = \{1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0\}$ and fixed number of topics $T = 10$. Using different α Dirichlet priors will have a noticeable effect on topic distribution. A high α value means that documents are likely to have a mixture of many topics with no single topic being dominant. A low α value results in very few (and in some cases, only one) topics being in the document.

Table 1: Base model topics for models with different numbers of topics

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
government	air	soviet	million	police	bush	health	percent	court	year
united	miles	united	company	people	president	people	market	case	school
states	people	states	billion	government	house	children	year	attorney	years
military	officials	gorbachev	year	south	dukakis	percent	million	judge	time
aid	fire	president	workers	killed	campaign	study	prices	trial	people
panama	area	union	corp	army	bill	report	billion	state	don
china	city	east	president	africa	state	program	dollar	charges	world
president	flight	government	based	party	senate	aids	rose	police	mrs
year	plane	west	business	city	democratic	years	oil	prison	show
rights	state	war	federal	violence	congress	space	stock	law	students

Table 2: Aggregated model topics for models with different numbers of topics

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
government	fire	soviet	company	police	bush	aids	percent	court	year
aid	people	united	million	people	dukakis	health	market	case	years
military	miles	president	corp	killed	campaign	disease	dollar	trial	people
rebels	area	officials	billion	government	president	drug	stock	judge	day
states	officials	gorbachev	stock	city	house	study	year	attorney	time
united	water	states	based	army	jackson	medical	prices	charges	home
nicaragua	reported	union	companies	reported	bill	virus	trading	prison	family
panama	north	government	business	today	senate	research	index	district	don
contra	southern	meeting	offer	violence	republican	blood	rose	state	back
president	air	told	president	injured	democratic	hospital	exchange	jury	life

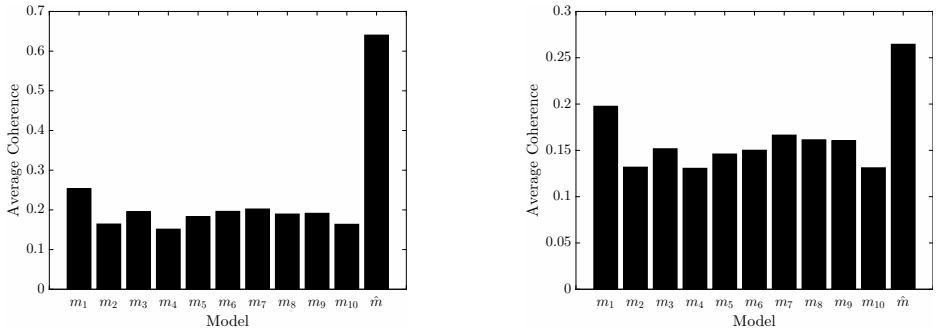


(a) Cosine Similarity

(b) Jensen-Shannon Divergence

Figure 4: Average topic coherence for aggregated models using different similarity thresholds and sliding window sizes for models with different α priors

As with the first experiment, the first step in generating the aggregated model is to choose the similarity threshold at which similar topics will be combined. The same method of grid search will be used for this experiment. The results of these grid searches are presented in Figure 4a and Figure 4b for cosine similarity and Jensen-Shannon divergence, respectively. This experiment shows the optimal similarity threshold according to PMI is > 0.9 for cosine similarity and < 0.1 for Jensen-Shannon divergence. Interestingly, this experiment is much more stringent in the similarity of topics before aggregation will take place. Experiment 1



(a) Extrinsic Coherence

(b) Intrinsic Coherence

Figure 5: Average intrinsic and extrinsic coherence for topics in the base models and aggregated model for models with different α priors

was much more lenient in the topics it would combine by using lower similarity thresholds. The results in this experiment mean that only very similar topics will be combined. This may lead to less of a change in the base model topics as fewer topics will be aggregated.

Following the tuning experiments, the full experiment was run using Jensen-Shannon divergence as the similarity measure and a similarity threshold of < 0.1, the results of this experiment for extrinsic and intrinsic coherence tests can be seen in Figures 5a and 5b, respectively. It should be noted that when the coherence of base models and aggregated models are compared, the same sliding window size is used for each model.

As in Experiment one, this experiment denotes m_1 as the base model, $m_2 - m_{10}$ as other models for comparison, and $\hat{m}$ as the aggregated model. As Figure 3a shows, the aggregated model has an extrinsic PMI value of 0.7, which is much higher than any of the models used to create it. This also shows that the aggregated model's topics are much more coherent based on general English language. The aggregated model also has the highest intrinsic coherence. This means the aggregated model's topics have been complemented with additional relevant topic words leading to topics that are more representative of the corpus.

In terms of how the underlying topics changed in the aggregated model, there are not as many changes as in Experiment 1. However, the few changes that occur improving topic coherence by a noticeable amount. For example, in m_1 there is a topic about Mikhail Gorbachev, the Soviet Union and the United States. In the aggregated model, this topic is supplemented with the words "east" and "Germany", making the topic more clearly about the Berlin wall and tensions between the West and East towards the end of the Cold War. The other major difference between base model topics and aggregated topics is in one about finances. The base model contains units of money such as "million" and "billion"; as well as words to do with the workforce, such as "workers" and "business". The aggregated model's equivalent topic also contains the words "industry" and "company".

This experiment is interesting as its topics are less changes than Experiment 1, but the few changes result in noticeable increases in topic coherence. This could be because some topics in the base model are quite specific, but are generalised more in the aggregated model.

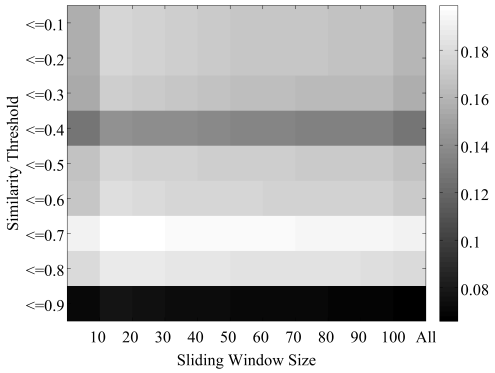


Figure 6: Average topic coherence for different similarity thresholds and sliding window sizes using social media data

6.3 Experiment 3: Aggregated Topic Models for Social Media

The concept of aggregated topic models has been validated in Experiments 1 and 2, now it can be evaluated over social media data to replicate a real use case. In this experiment 2,000 Gibbs sampling iterations were performed and a β Dirichlet prior of 0.01 was used. Ten different models were generated with each having a different number of topics and a different α prior which reflects this. Models t_1-t_{10} have the following topic numbers $T = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ and the following α priors $\alpha = \{5.0, 2.5, 1.67, 1.25, 1.0, 0.83, 0.71, 0.63, 0.56, 0.5\}$. This will allow for the evaluation of how aggregating diverse social media models affects coherence.

On 19th October 2016, the third Presidential Debate between Democratic nominee Hillary Clinton and Republican nominee Donald Trump took place at the University of Nevada, Las Vegas. The debate lasted 90 minutes and had six topics split roughly into 15 minute segments. The topics chosen by the chair were on debt and entitlements, immigration, economy, Supreme Court, foreign hot spots, and fitness to be president. This debate was the climax of lengthy campaigns which were not without scandal and dishonesty from both parties. The candidates provoked dissent and discord amongst the American population and this was reflected on Twitter. During the debate 300,000 Tweets were captured using various hashtags and keywords used by supporters of each nominee. These can be seen in Tables 3.

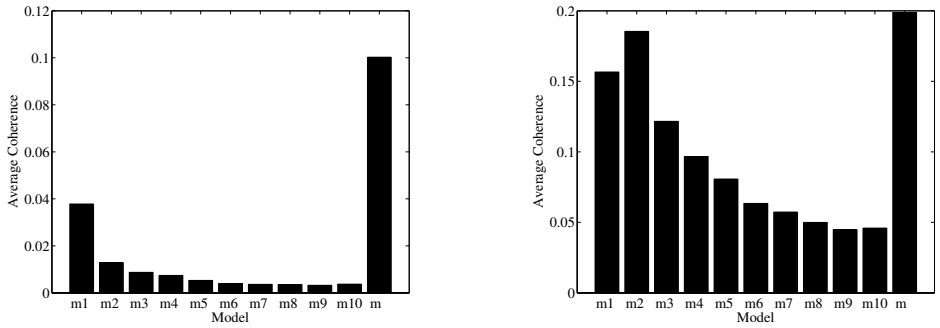
Following the same process as in the previous experiments, the first task was to perform tuning experiments to find the optimal similarity threshold and sliding window size. Since Jensen-Shannon divergence was the best performing similarity metric in both tuning experiments it was also use here. The results can be seen in Figure 6. The legend of this figure details the coherence with red being more coherent and blue being less coherent. This shows that $<= 0.7$ is the optimal threshold, meaning that the model is not too strict about which topics to aggregate, therefore many topics being aggregated.

In the tuning experiment, the full experiment was run with a similarity threshold of $<= 0.7$ and a sliding window size of 20. The results are presented in Figures 7a and 7b for extrinsic coherence and intrinsic coherence, respectively.

Figure 7a shows that the extrinsic coherence of the aggregated model increase greatly compared to the base models used to generate it. This means that the aggregated topics are a better representation of being coherent in general English language, however, this level of

Table 3: Keywords used for Tweet Collection

Trump	Clinton
#AmericaFirst	#ImWithHer
#ImWithYou	#LoveTrumpsHate
#MAGA	#NeverTrump
#TrumpTrain	#Clinton
#MakeAmericaGreatAgain	#ClintonKaine16
#TrumpPence16	#ClintonKaine2016
#TrumpPence2016	#DNC
#Trump	#OHHillYes
#AltRight	#StrongerTogether
#NeverHillary	#VoteDems
#deplorable	#dems
#TeamTrump	#DirtyDonald
#VoteTrump	#HillaryClinton
#CrookedHillary	#Factcheck
#LatinosForTrump	#TrumpedUpTrickleDown
#ClintonFoundation	#ClintonKaine
#realDonaldTrump	#WhyIWantHillary
#LawAndOrder	#HillarysArmy
#pepe	#CountryBeforeParty
#DebateSideEffects	#TNTweeters
#WakeUpAmerica	#UniteBlue
#RNC	#p2
#tcot	#ctl
	#p2b



(a) Extrinsic Coherence (b) Intrinsic Coherence

Figure 7: Average intrinsic and extrinsic coherence for topics in the base models and aggregated model for social media data

coherence is quite low as expected due to the modelled corpus being quite domain specific. The more specific a corpus is, the harder it is to have a high extrinsic coherence as the extrinsic reference corpus have a lot of general English terms not specific to the corpus.

Figure 7b shows the intrinsic coherence increases slightly compared to the base models. This means that the aggregated model is a slightly better representation of the underlying corpus that is being modelled. However, there is not much difference between model m_2 and the aggregated model.

In terms of changes between the base model and the aggregated model, Tables 4 and 5 show the topics before and after, respectively. As can be seen t_1 in the base model is mainly about Trump and his affinity for "Putin" and "Russia", whereas in the aggregated model this topic also has the words "woman" and "issues", referring to the number of women who came out before the debate to allege that Trump had made inappropriate advances on them. Importantly, t_2 has the addition of the word "amnesty". This is associated with the word "illegals" in the base model topic and represents Clinton's desire to grant amnesty to many illegal immigrants. Topic t_{10} also shows that the aggregated model has the ability to filter noise out of topics; in the base model the string "skhbwg6aq3" appears but is not present in the aggregated model.

6.4 Experiment 4: Comparison with the NMF algorithm

A NMF algorithm utilises linear algebra to factorise a matrix V into two matrices, W and H where $V = WH$. These matrices have the property that they only have positive or zero elements, means they are non-negative [9]. In the context of topic modelling, V is the document-term matrix where terms are generally represented by $tf\text{-}idf$ (term frequency–inverse document frequency) values for example, W is the term-feature matrix, and H is the feature-document matrix. In this case the feature-document matrix describes the topics found in a textual corpus.

At its core, NMF approximates the dot product of W and H through iterations, resulting in the product of V. This iterative process is repeated until a specified amount of iterations is reached or the approximation error converges to a certain point. Additionally, l_2 regularisation loss is used to prevent weights in the matrices from becoming too large.

In this experiment, the same datasets used in Experiments 1 and 2 have been used again with NMF, producing a new set of topics. As in the previous experiments, both intrinsic and extrinsic coherence are calculated. Intrinsic uses the corpus the documents were generated from and extrinsic uses the English Wikipedia. All other parameters for testing are also the same as previous experiments. For fairness, 10 topics were generated for each dataset as this is the same amount of topics that were previously generated using the aggregated topic model.

NMF was configured to run for a maximum of 2000 iterations if approximate error did not converge. Initialisation was performed using non-negative double singular value decomposition [10] with a multiplicative updater solver.

7 Evaluation of the Experimental Results

The experimental results reveal that the aggregating models increase the coherence of topics. Figures 3a, 3b, 5a and 5b show that the model with the lowest number of topics or highest α prior (m_1 from both experiments) are normally the most coherent topic but after aggregation, the aggregated topic is the most coherent. This could be because m_1 is usually the most general model, therefore when evaluated extrinsically the words would have a high probability of co-occurring as they are not specific. What is also interesting is the fact that the aggregated

Table 4: Base model topics for social media data

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
trump	hillary	america	states	debate	clinton	people	obama	corrupt	women
donald	clinton	make	president	tonight	hillary	support	money	change	debate
time	trump	great	borders	final	taxes	vote	country	future	puppet
putin	american	ready	drug	foundation	back	abortion	don	leader	wall
bad	tax	question	deport	twitter	mosul	ion	world	nation	won
lies	didn	election	open	live	plan	don	run	woman	street
russia	put	state	border	watching	lost	doesn	give	weak	presidential
wrong	pay	campaign	fact	hands	one	supporter	economic	oligarchy	rights
talk	liar	response	wallace	proof	wasn	two	rigged	defendable	skhbwg6aq3
war	talking	edu	lords	skin	amendment	four	haiti	nasty	proud

Table 5: Aggregated model topics for social media data

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
trump	hillary	great	america	debate	clinton	support	obama	corrupt	women
donald	clinton	america	president	tonight	hillary	abortion	money	future	debate
talking	american	make	tonight	ll	back	ion	country	leader	won
woman	tax	ready	deport	puppet	mosul	baby	don	nation	rights
issues	put	question	drug	hillary	tonight	hillary	world	weak	retweet
putin	dollars	states	border	america	taxes	supporter	run	oligarchy	presidential
nasty	illegals	response	hillary	taxes	lost	page	give	defendable	respect
god	amnesty	edu	america	obama	america	bi	economic	change	care
clinton	thousands	campaign	ll	wall	ll	term	rigged	uninformed	defend
speaks	provide	confirmed	fact	live	fact	babies	haiti	democracy	final

model also has the highest intrinsic coherence, meaning that combining elements of more specific models into the general model allows for a greater representation of the modelled corpus.

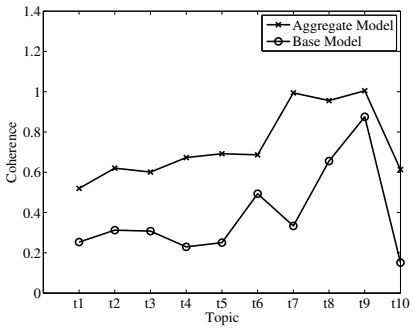
However, it was found that to maximise coherence the sliding window size had to be set to the size of the document being analysed. Using the full document size is not detrimental to results as the average document length is 133.07 words, which is only 33.07 words more than the second highest average coherence sliding window size of 100.

7.1 Significance Testing

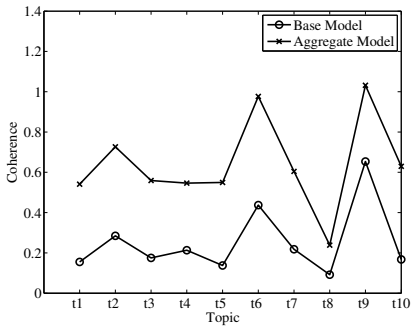
A paired t-test was performed to compare the topic coherence for the base model and best aggregated model from each experiment by accounting for the difference between base models and aggregated models, as well as the mean values, in order to ascertain if there is a statistically significant difference in topic coherence after aggregating the similar models. The critical value is $\alpha = 0.05$.

7.1.1 Experiment 1

There is a significant difference in the topic coherence for the base model ($\mu = 0.386$, $\sigma = 0.179$) and aggregated model ($\mu = 0.736$, $\sigma = 0.224$); $t(9) = 7.173$, $P = 0.0000523$. These results suggest that aggregating the output of multiple topic models can increase the topic coherence. The comparison between coherence for each topic in the base model and aggregated model can be seen in Figure 8a. Interestingly, Figure 8a also shows that the two models are



(a) Comparison of topic coherences for the base model and aggregated model for models with different topic numbers



(b) Comparison of topic coherences for the base model and aggregated model for models with different α priors

Figure 8: Comparison of topic coherences

somewhat positively correlated with a Pearson's correlation coefficient of $\rho = 0.73$. This suggests that each topic in the model has had the same scale of topic coherence improvement.

7.1.2 Experiment 2

There is a significant difference in the topic coherence for the base model ($\mu = 0.253$, $\sigma = 0.17$) and aggregated model ($\mu = 0.64$, $\sigma = 0.229$); $t(9) = 12.03$, $P = 0.00000075$. These results suggest that aggregating the output of multiple topic models can increase the topic coherence. The comparison between coherence for each topic in the base model and aggregated model can be seen in Figure 8b. Again, the two models are somewhat positively correlated with a Pearson's correlation coefficient of $\rho = 0.911$. This suggests that each topic in the model has had the same scale of topic coherence improvement.

7.1.3 Experiment 3

There is a significant difference in the topic coherence for the base model ($\mu = 0.157$, $\sigma = 0.219$) and aggregated model ($\mu = 0.177$, $\sigma = 0.227$); $t(9) = 4.53$, $P = 0.0014$. The results presented above showing that there is an increase in coherence and this is backed by a statistically significant difference, this result suggests that aggregating the output of multiple topic models for social media can increase the topic coherence at a statistically significant level.

7.1.4 Experiment 4

The coherence of these NMF topics is calculated and compared to the coherence of the topics discovered using the aggregated topic model. Due to the inherent nature of NMF one cannot directly create aggregated topics, as the words in the topics do not have well-defined probabilities. This happens because the NMF algorithm uses the Frobenius norm as the objective function for loss. Probabilities can be extracted if the generalised Kullback-Leibler divergence [11] is used for loss, however, this is equivalent to the older, superseded probabilistic latent semantic indexing [17, 13] which pre-dates LDA and NMF.

Table 6: Intrinsic and extrinsic coherence for topics generated using non-negative matrix factorisation compared to aggregated topic models for Associated Press dataset

Topic	Intrinsic	Extrinsic
Topic 1	0.01	0.08
Topic 2	0.37	0.13
Topic 3	0.15	0.16
Topic 4	0.22	0.01
Topic 5	0.25	0.06
Topic 6	0.19	0.12
Topic 7	0.22	0.26
Topic 8	0.2	0.01
Topic 9	0.28	0.04
Topic 10	0.13	0.17
Mean Coherence	0.202	0.106
Mean Aggreagted Model Coherence	**0.39**	**0.75**

The results for the Associated Press dataset can be seen in Table 6. This table shows the intrinsic and extrinsic coherence for all topics generated by NMF, along with the mean coherence and mean aggregated topic model coherence. As it shows, intrinsically, both models performed quite low, however, the aggregated topic model was more coherent by almost double NMF's coherence. The most interesting result is how much more coherent the aggregated model was extrinsically compared to NMF, more than seven times more coherent. From empirical observation it appears that NMF gives higher weight to fewer words in topics compared to the LDA models used to create the aggregated topic model. Additionally, it appears as though NMF does not capture advanced lexical devices such as polysemy as good as LDA. These could be contributing factors to the lower coherence score.

The results for the presidential debate social media dataset can be seen in Table 7. Again, in this domain the aggregated topic model outperformed NMF but not by as large a difference as the previous experiment most likely due to noise in the data. It should be noted that the zero coherence scores presented are in fact extremely small decimal values. Intrinsically, both NMF and the aggregated topic model perform similarly with the aggregated topic model outperforming NMF by only 0.021. This shows that the aggregated topic model's topics better capture the underlying topics of the dataset. The aggregated topic model also outperformed NMF extrinsically, this time by 0.081. The lower score of NMF this time is most likely due to the top terms in topics being heavily influenced by noise in the data, leading to this noise not matching anything in the English Wikipedia. The aggregated topic model helps alleviate this problem by bringing similar terms from other topics in to replace noise which is not prevalent in other topics, leading to a higher coherence.

8 Discussion and Conclusion

To summarise, the proposed aggregated topic model method was tested in three experiments (experiments one and two served as a proof-of-concept and experiment three as a real world example). All experiments showed the aggregated topic model improved topic coherence by

Table 7: Intrinsic and extrinsic coherence for topics generated using non-negative matrix factorisation compared to aggregated topic models for social media dataset

Topic	Intrinsic	Extrinsic
Topic 1	0	0
Topic 2	0.65	0.03
Topic 3	0.15	0.05
Topic 4	0.11	0
Topic 5	0.28	0.03
Topic 6	0.01	0.01
Topic 7	0.03	0.01
Topic 8	0.12	0.01
Topic 9	0.2	0.01
Topic 10	0.24	0.03
Mean Coherence	0.179	0.019
Mean Aggreagted Model Coherence	**0.2**	**0.1**

a statistically significant amount.

This work proposes a novel solution for aggregating topic models that can improve the coherence of the topics produced. The experiments conducted demonstrate that the coherence has been improved through aggregating topic models. The experiments show that the coherence is improved after creating a number of models with different numbers of topics or different parameters and applying the aggregation technique. The experimental results provide an insight into a conjecture of the improvement that when models are created with different numbers of topics, they create a mix of general, as well as more focused, specific sets of topics as the number of topics increases. The advantage of this is that the aggregated models have more general topics which lead to the aggregated model being more representative of the corpus it was generated from as shown by the intrinsic coherence results. It is also observed that Jensen-Shannon divergence generally gives better results than cosine similarity. This could be because Jensen-Shannon divergence assesses if two distributions were drawn from the same underlying distribution rather than simply assessing similarity, as is the case with cosine similarity.

The results of the proof-of-concept experiment two were also interesting as despite having fewer changes in the aggregated model than the first experiment, there was a noticeable difference in coherence. This suggests that aggregation allows for more general topics, and that any form of generalisation results in a higher topic coherence.

We also showed that this work can be used successfully in the social media domain. We demonstrated that it works well at increasing the topic coherence and adding additional words to topics which make them more coherent. Additionally, the aggregated method has the feature of being able to filter out noise from topics. Despite the experimental results showing an increase in coherence, it was not at a statistically significant level.

It is important to note that although the top N words in a topic may not appear to change much in some cases; the underlying ϕ distribution of the topics (topic-word distribution) will change after the aggregated model is formed.

The proposed aggregation technique shows that it outperforms standard topic models in

topic coherence, but the method can still be improved, for example, by clustering or bagging the corpus into subsets of data and generating base models using these subsets, which could then be used for generate aggregated topic models. The topics generated from these subsets when aggregated could provide a good mix of general topics, as well as specific topics. This work could also be furthered by creating aggregated topics from different types of topic models.

The comparison between NMF and aggregated topic models demonstrate that the aggregated topic model outperforms NMF in terms of coherence both extrinsically and intrinsically on both datasets. Both modelling methods performed quite similarly intrinsically showing that they both capture the underlying topical structure of datasets well, however, the extrinsic results are extremely different. The aggregated topic model strongly outperforms NMF extrinsically. This reveals how the aggregated topic model brings similar terms into a topic from other similar topics to displace potentially noisy terms, thereby increasing coherence extrinsically which demonstrates that the topic should be coherent in daily English language.

Another important area of further work is to present the base models and aggregated models to humans and have them to rank the topics based on human's perception. This will allow for examining the correlation of the coherence of the aggregated model with human opinion.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: `http://dl.acm.org/citation.cfm?id=944919.944937`

[2] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno, "Evaluation Methods for Topic Models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, New York, USA, 2009, pp. 1105–1112.

[3] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: ACM, 2013, pp. 889–892. [Online]. Available: `http://doi.acm.org/10.1145/2484028.2484166`

[4] A. K. Rider and N. V. Chawla, "An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2013, pp. 333–340.

[5] Z. Shen, P. Luo, S. Yang, and X. Shen, "Topic Modeling Ensembles," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 1031–1036.

[6] W. Yongliang and G. Qiao, "Multi-LDA Hybrid Topic Model with Boosting Strategy and its Application in Text Classification," in *33rd Chinese Control Conference*, 2014, pp. 4802–4806.

[7] X. Li, J. Ouyang, X. Zhou, Y. Lu and Y. Liu. Supervised labeled latent Dirichlet allocation for document categorization. Applied Intelligence (2015) 42: 581–593.

[8] J. Qiang, Y. Li, Y. Yuan, W. Liu. Snapshot ensembles of non-negative matrix factorization for stability of topic modeling. Applied Intelligence (in press).

[9] S. Sra and I. S. Dhillon. Generalized Non-negative Matrix Approximations with Bregman Divergences, Advances in Neural Information Processing Systems, 2006, , 283–290.

[10] C. Boutsidis, and E. Gallopoulos. SVD based Initialization: A head start for Non-negative Matrix Factorization, Pattern Recognition, 2008, 41 (4), 1350–1362.

[11] S. Kullback, Solomon, Information Theory and Statistics, Courier Corporation, 1997

[12] T. Hofmann, Thomas Probabilistic Latent Semantic Analysis. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999, 289–296.

[13] C. Ding and T. Li, Tao and W. Peng. On the Equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing, Computational Statistics & Data Analysis, 2008, 52(8), 3913–3927.

[14] ——, "Modeling Texts in Semantic Space and Ensemble Topic-Models via Boosting Strategy," in *34th Chinese Control Conference*, 2015, pp. 3838–3843.

[15] W. Li and A. McCallum, "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 577–584.

[16] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical Topic Models and the Nested Chinese Restaurant Process," in *Advances in Neural Information Processing Systems 16*, 2004, pp. 17–24.

[17] T. Hofmann, "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data," *IJCAI*, vol. 99, pp. 682–687, 1999.

[18] N. Burns, Y. Bi, H. Wang, and T. Anderson, "A Twofold-LDA Model for Customer Review Analysis," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, pp. 253–256.

[19] ——, "Extended Twofold-LDA Model for Two Aspects in One Sentence," in *Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2012, pp. 265–275.

[20] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A Biterm Topic Model for Short Texts," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 1445–1456.

[21] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 363–371.

[22] J. Tang, M. Zhang, and Q. Mei, "One Theme in All Views: Modeling Consensus Topics in Multiple Contexts," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 5–13.

[23] H. Wu, K. Yue, Y. Pei, B. Li, Y. Zhao, and F. Dong, "Collaborative Topic Regression with social trust ensemble for recommendation in social media systems," *Knowledge-Based Systems*, vol. 97, pp. 111–122, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705116000216

[24] M. Belford, B. M. Namee, and D. Greene, "Ensemble Topic Modeling via Matrix Factorization," in *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science, {AICS} 2016, Dublin, Ireland, September 20-21, 2016.*, 2016, pp. 21–32. [Online]. Available: http://ceur-ws.org/Vol-1751/AICS{_}2016{_}paper{_}36.pdf

[25] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and Sparse Text Topic Modeling via Self-Aggregation," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 2270–2276.

[26] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems*, 2009, pp. 288–296.

[27] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, 2010, pp. 100–108.

[28] H. M. Wallach, D. Mimno, and A. Mccallum, "Rethinking LDA: Why Priors Matter," in *NIPS*, 2009.

[29] H. M. Wallach, "Structured Topic Models for Language," Ph.D. dissertation, Universty of Cambridge, 2008.

[30] M. Newton and A. Raftery, "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society*, vol. 56, no. 1, pp. 3 – 48, 1994.

[31] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: ACM, 2015, pp. 399–408. [Online]. Available: http://doi.acm.org/10.1145/2684822.2685324

[32] 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*, Gothenburg, 2014, pp. 530–539.

[33] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Topics in Tweets: {A} User Study of Topic Coherence Metrics for Twitter Data," in *Advances in Information Retrieval - 38th European Conference on {IR} Research, {ECIR} 2016, Padua, Italy, March 20-23, 2016. Proceedings*, 2016, pp. 492–504. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-30671-1{_}36

[34] N. Aletras and M. Stevenson, "Evaluating Topic Coherence Using Distributional Semantics," in *Proceedings of the 10th International Conference on Computational Semantics, {IWCS} 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, 2013, pp. 13–22. [Online]. Available: http://aclweb.org/anthology/W/W13/W13-0102.pdf

[35] R. Cilibrasi and P. Vitanyi, "The Google Similarity Distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[36] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best Topic Word Selection for Topic Labelling," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 605–613.

[37] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[38] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring Topic Coherence over Many Models and Many Topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 952–961. [Online]. Available: `http://dl.acm.org/citation.cfm?id=2390948.2391052`

[39] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. [Online]. Available: `http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490`

[40] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5228–5235, 2004.

[41] N. Aletras and M. Stevenson, "Measuring the Similarity between Automatically Generated Topics," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, 2014, pp. 22–27.

[42] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[43] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd ed. Chapman & Hall/CRC, 2000.

[44] Soutner, D., and Müller, L. Application of LSTM neural networks in language modelling. In International Conference on Text, Speech and Dialogue. pp. 105–112. Springer, Berlin, Heidelberg.

[45] Miao, Y., Grefenstette, E., and Blunsom, P. (2017). Discovering discrete latent topics with neural variational inference. arXiv preprint arXiv:1706.00359.

[46] Dieng, A. B., Wang, C., Gao, J., and Paisley, J. (2016). Topicrnn: A recurrent neural network with long-range semantic dependency. arXiv preprint arXiv:1611.01702.

[47] Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2017). Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings. ACM Transactions on Information Systems (TOIS), 36(2), 11.

[48] Fu, X., Huang, K., Sidiropoulos, N. D., Shi, Q., and Hong, M. (2018). Anchor-Free Correlated Topic Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence.