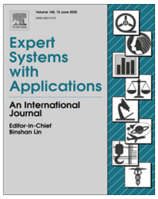
Expert Systems with Applications 161 (2020) 113709

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa



Ensemble topic modeling using weighted term co-associations

Mark Belford [*], Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 24 February 2020
Revised 17 June 2020
Accepted 29 June 2020
Available online 8 July 2020

Keywords:
Topic modeling
Ensemble learning
Evaluation
Word embeddings
Interpretation

ABSTRACT

Topic modeling is a popular unsupervised technique that is used to discover the latent thematic structure in text corpora. The evaluation of topic models typically involves measuring the semantic coherence of the terms describing each topic, where a single value is used to summarize the quality of an overall model. However, this can create difficulties when one seeks to interpret the strengths and weaknesses of a given topic model. With this in mind, we propose a new ensemble topic modeling approach that incorporates both stability information, in the form of term co-associations, and semantic similarity information, as derived from a word embedding constructed on a background corpus. Our evaluations show that this approach can simultaneously yield higher quality models when considering the produced topic descriptors and document-topic assignments, while also facilitating the comparison and evaluation of solutions through the visualization of the discovered topical structure, the ordering of the topic descriptors, and the ranking of term pairs which appear in topic descriptors.

1. Introduction

The goal of topic modeling is to uncover semantic structures, referred to as topics, from a corpus of documents. There are many popular topic modeling algorithms, including probabilistic techniques such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). More recently, matrix factorization methods have also been applied to discover topical structure (Arora, Ge, & Moitra, 2012; Kuang, Choo, & Park, 2015), with the most prominent being Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999). Topic modeling algorithms have been applied to a variety of tasks, including sentiment analysis (Lin & He, 2009), text mining (Pauca, Shahnaz, Berry, & Plemmons, 2004), and the analysis of large collections of news articles (Jacobi, Van Atteveldt, & Welbers, 2016).

After generating one or more topic models on a given corpus, the subsequent interpretation and evaluation of these models can itself be a difficult task. Common approaches to topic model evaluation, such as *topic coherence* (Newman, Lau, Grieser, & Baldwin, 2010; Mimno, Wallach, Talley, Leenders, & McCallum, 2011), and *topic stability* (De Waal & Barnarde, 2008), are typically applied in order to produce a single score which summarizes the "quality" of the overall model. However, these measures tend to capture different aspects of the model, and are rarely used in conjunction with one another. As a result, varying levels of quality can be reported when evaluating the same models. This also leads to ambiguities around the interpretation of the resulting models, especially when comparing the outputs from different algorithms or comparing solutions comprising of different numbers of topics. These issues around interpretation are potentially magnified when we consider ensemble approaches to topic modeling, which involve the generation and combination of many different models (Belford, Mac Namee, & Greene, 2018).

To address these issues, we propose a new ensemble topic modeling approach based on Weighted Term Co-associations (WTCA), which is interpretable in the sense that the strengths and weaknesses of different topic models and individual topics can be readily identified. In this approach, the inherent stability of topic modeling solutions that have been generated over a large number of runs is captured by considering the extent to which pairs of terms are repeatedly associated with the same topic. We augment this information by incorporating weights for pairs of terms based on their corresponding semantic similarity, also known as coherence, as derived from a pre-trained word embedding model. This allows us to incorporate rich background knowledge from a large reference corpora using an efficient low dimensional representation. While the use of pre-trained word embeddings has been widely adopted in classification tasks, its use in topic modeling has primarily focused on probabilistic techniques such as LDA (Xie, Yang, & Xing, 2015; Yang, Downey, & Boyd-Graber, 2015; Das, Zaheer, & Dyer, 2015; Nguyen, Billingsley, Du, & Johnson, 2015; Moody, 2016). There has also been initial work using these

[*] Corresponding author.
E-mail addresses: mark.belford@insight-centre.org (M. Belford), derek.greene@ucd.ie (D. Greene).

https://doi.org/10.1016/j.eswa.2020.113709
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

embeddings to evaluate the coherence of topic models (O'Callaghan, Greene, Carthy, & Cunningham, 2015; Ding, Nallapati, & Xiang, 2018).

By using this weighted term co-association method, a set of more interpretable *ensemble topic descriptors* can be extracted, which is complemented by a visual heatmap of the discovered topical structure. This provides a visual representation of how the underlying data is represented for a given value of *k*, provides a ranking of the discovered topic descriptors and also allows for the exploration of the best and worst pairs of terms in each topic, which in conjunction aids in the interpretation of large numbers of topics by a user. Previous work has rarely seen the use of pre-trained embeddings in the context of NMF topic modeling or the combination of ensemble topic modeling with word embeddings. Also, while this paper focuses on the use of WTCA in conjunction with NMF, our approach is not tied to a specific topic modeling paradigm, due to utilizing the produced topics, and not the underlying weights or probabilities of the model. In this sense our proposed approach is algorithm *agnostic*.

To validate our new ensemble approach, we apply it to a large, diverse set of corpora, and compare its performance to a number of topic modeling approaches. Through several evaluation tasks, we show that our approach yields more coherent solutions, while also producing more accurate document-topic assignments. We provide a reference implementation of the proposed WTCA method for use by other researchers working in the area of topic modeling[1].

The remainder of the paper is structured as follows. In Section 2 we summarize relevant work in the areas of topic modeling, word embeddings, topic model evaluation, and model visualization. In Section 3 we discuss the methodology of our proposed ensemble topic modeling approach before presenting an evaluation of our approach in Section 4. Finally we present our conclusions in Section 5.

2. Related work

2.1. Topic modeling

Topic modeling, which aims to discover meaningful semantic structures in a corpus, originates from early work on Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), where Singular Value Decomposition was applied to a term-document matrix to uncover underlying themes in the data. The goal of topic modeling is to identify *k* topics, each represented as a ranked list of the top *t* associated terms, often referred to as a *topic descriptor*. These descriptors are typically presented as the main output of a model. Topic modeling algorithms also represent each document in a corpus as a combination of these *k* topics with varying degrees of association, where a *primary topic* can be assigned to each document by choosing the corresponding topic for which it has the highest association score. By applying this process to all documents, we can naturally produce a disjoint partition of primary topic assignments.

There has been considerable previous research in the area of probabilistic topic modeling. In these approaches, topics are viewed as a probability distribution over words, while the documents are composed as a mixture of the topics. LDA (Blei et al., 2003) is the most popular probabilistic technique, and a number of related inference algorithms exist, such as Markov chain Monte Carlo and variational inference. One of the most popular inference approaches, implemented as part of the *Mallet* software package (McCallum, 2002), uses fast Gibbs sampling. More recently, work on alternative topic modeling algorithms, such as NMF (Lee & Seung, 1999), have become prominent for discovering topics in corpora of text documents (Arora et al., 2012; Kuang et al., 2015). In the context of textual data, NMF can be viewed as a dimensionality reduction technique, where the goal is to approximate the original $n \times m$ document-term matrix $\mathbf{A}$, as the product of two non-negative factors $\mathbf{W}$ and $\mathbf{H}$. The rows of $\mathbf{H}$ correspond to the k topics, where each topic is associated with the m terms from the corpus vocabulary, weighted using non-negative values. If we order these rows (*i.e.* terms) in a descending fashion, we can construct the corresponding ranked topic descriptors for each of the k topics. The columns of $\mathbf{W}$ hold the membership weights for each document with respect to the k topics.

2.2. Word embeddings

The generation of *word embeddings* is a process which entails mapping the vocabulary of a corpus to a vector space representation. Unlike traditional bag-of-words models which yield extremely sparse vectors, word embedding vectors are typically dense and low dimensional. One popular approach for creating word embeddings, word2vec, is a two-layer neural network which can efficiently learn high quality vectors from very large datasets (Mikolov, Chen, Corrado, & Dean, 2013).

The word2vec approach offers two different model architectures for generating word embeddings. Both are based on training the respective model using information about co-occurring words over a large number of documents. The first of these is known as the Continuous Bag-of-Words model in which a target word is predicted by it surrounding context words. The second approach is known as the Continuous Skip-gram model which is the inverse of the previous methodology in which the surrounding context words are predicted based on the target word. In both of these models a user-defined window size is used to select the context words. Once a word2vec model has been trained, the similarity of words can be calculated as the pairwise cosine similarity between their corresponding vectors. Words that are similar with regards to their contextual usage should be embedded closely together in the space (O'Callaghan et al., 2015).

2.3. Topic model evaluation techniques

2.3.1. Topic coherence

Topic descriptors are traditionally the primary output of a topic model. These are typically presented to the user in tabular form, or passed to a downstream application. As such, it is important that they are of high quality and are interpretable. One popular approach to quantify their quality is through the measurement of the semantic interpretability of the terms, known as *topic coherence*. While originally cast as a human evaluation task (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009), one popular automated technique that has inspired a number of different approaches makes use of pairwise term co-occurrence counts relative to an external reference corpus. One such example of this, proposed by Newman et al. (2010), calculates co-occurrences with respect to a large Wikipedia reference corpus, which are then used in a Pointwise Mutual Information measure to calculate coherence scores for pairs of terms in a given topic descriptor. To generate an overall coherence score for the topic, the authors suggest calculating the arithmetic mean of all pairwise scores. A similar measure was also proposed by Mimno et al. (2011). However, in this approach the term co-occurrence counts are instead calculated based on whether they appear within the same document, and the authors also suggest that they are calculated by using the same corpus that the topic model was trained on.

Other techniques have calculated topic coherence through the use of word embeddings. For instance, O'Callaghan et al. (2015)

[1] See https://github.com/MarkBelford/co-association

proposed an intra-topic coherence measure that uses a word embedding to calculate pairwise cosine similarity values for terms in a topic's descriptor. A topic-level coherence score is produced by averaging these similarities across all pairs of terms appearing in the descriptor. An overall model score is given by the average coherence across all k topics.

2.3.2. Topic stability

Topic models frequently employ random initialization for the initial values of their respective topic-term and document-topic assignment matrices. These initial values can have a large influence on the resulting model due to the discovery of different local solutions between consecutive runs, which leads to an inherent *instability* in both probabilistic and factorization-based topic modeling approaches (Belford et al., 2018). In other words, if we repeatedly apply the same topic modeling algorithm to the same corpus, or data sampled from the same source, then we frequently achieve different sets of results between consecutive runs. The implications of this in practice include the ordering of terms changing within a descriptor, topics appearing intermittently across different runs, and variations in the primary topic assigned to each document.

Instability in topic modeling, specifically with regards to using the probabilistic distributions of LDA has previously been investigated. Steyvers and Griffiths (2007) measure the dissimilarity of the topics present in each model using a symmetric Kullback–Leibler distance measure that considers the associated topic-term distributions, before using these distances to calculate an overall stability score. Another similar stability measure (De Waal & Barnarde, 2008) was also proposed, however this approach calculates the stability of topics by considering the correlation between the document-topic probability distributions.

Previous work regarding instability with respect to matrix factorization techniques, specifically NMF, has also been investigated. Greene, O'Callaghan, and Cunningham (2014) proposed a term-centric measure, Average Jaccard, that can be used to evaluate the stability between either individual pairs of topic descriptors or pairs of topic models, and can also be employed for model selection tasks. Belford et al. (2018) also discussed the issue and proposed a number of measures to quantify stability that can be utilized for both probabilistic and matrix factorization approaches.

2.4. Topic model visualization

While topic model vizualization has previously been used as an aid to explore the structure of large collections of text documents (Chaney & Blei, 2012; Liu et al., 2012; Gretarsson et al., 2012), it can also be used as part of the topic model evaluation process. One such approach, *Termite* (Chuang, Manning, & Heer, 2012), takes a term-centric view, where the objective is to produce a compact and interpretable visualization of the topic-term matrix, constructed from the underlying probabilities generated by LDA. The authors propose a saliency measure to ensure that only the most informative terms are incorporated into the final visualization. The rationale is that terms appearing in multiple topics will tend to provide less information than those that are more distinct, and topics tend to have a long tail of low probability terms over the corpus vocabulary. It is also possible to perform a seriation step on this topic-term matrix visualization which clusters the data to extract inherent structure, which is useful to further enhance the end user's interpretation of the data. Another popular topic model visualization tool, *LDAvis* (Sievert & Shirley, 2014), embeds topics in a two dimensional-space using multidimensional scaling, where the inter-topic distance is taken into account. As with Termite, this tool also incorporates a term relevance measure which is used to choose which terms to present from each topic to produce an interpretable visualization. A further clustering step of this two dimensional visualization using k-means is also possible to observe the impact of terms within these grouped topics.

While the majority of previous topic modeling visualization research has focused primarily on the topic-term distributions, Murdock and Allen (2015) proposed a framework called *Topic Explorer*, which instead models inter-document and topic document relationships. In this case, the user can choose a focal topic or document to study, and produce a ranked list of documents that are most similar, with respect to their Jensen–Shannon distance. This allows for the further interpretation of the topic modeling solution. Velcin et al. (2018) proposed a topic visualization tool known as *Readitopics*, in which a number of automatic topic labelling and coherence methodologies are applied for interpretation and evaluation. Previous visualization approaches in this area have focused primarily on probabilistic topic models, although (Kim, Kang, Park, Choo, & Elmqvist, 2016) developed methods which were applicable to topic models generated via hierarchical matrix factorization.

3. Methodology

In this section we outline our proposed Weighted Term Co-association (WTCA) method which leverages both term stability and semantic similarity information, to support the further investigation and interpretation of topic models. While we focus on its use in the context of matrix factorization, WTCA could also be used in the context of LDA, as it only relies on the use of topic descriptors, rather than the underlying weights or probabilities generated by a given algorithm.

3.1. Weighted term co-associations

Previous work in the area of ensemble clustering has looked at the idea of measuring the *co-association* between items (e.g. documents) in a dataset based upon their repeated co-assignment across multiple clustering runs (Strehl, 2002). This effectively results in an emergent measure of stability between pairs of items. Co-associations are either calculated simply based on the proportion of ensemble members for which a pair of items are assigned together, or by applying some weighting function with respect to this count (Berikov & Pestunov, 2017). The resulting co-associations are typically represented by a symmetric matrix, where a value close to 1 indicates a high level of certainty that two items should be assigned to the same cluster, while a value closer to 0 strongly suggests that they should be assigned to different clusters.

Following on from this work, and motivated by the topic model validation issues previously discussed, we propose a co-association strategy which focuses on terms rather than documents. When considering many runs of a topic modeling algorithm generated on the same corpus of documents for a given value of k, the repeated appearance of pairs of terms in topic descriptors provide a useful indicator of stable solutions (Belford et al., 2018). Therefore, a key aspect of WTCA involves identifying pairs of consistently-appearing terms across many runs. This co-association information is then augmented by leveraging semantic similarity information from a word embedding model which allows for the incorporation of coherence knowledge derived from the embedding.

For a given value of k, we generate a collection of r "base" topic models, each corresponding to the output of a single run of randomly-initialized NMF on the document-term matrix representation of the corpus. These runs could either be generated on the full corpus or, following methods used in ensemble clustering to generate diversity (Strehl, 2002), we could generate each run on a different random sample of documents. When selecting the value of r care should be taken to ensure that many diverse models are produced. However, it is also the case that as the value of r increases so does the computational cost of the approach, with eventually diminishing returns due to the likelihood of producing many similar models. For our upcoming experiments we find that $r = 100$ strikes a good balance between promoting diversity and reducing computational cost. It should also be noted that due to the generation and integration steps of our proposed ensemble approach, it will naturally be slower than existing randomly initialized techniques. However, the goal of WTCA is to produce more robust and informative models which is a worthwhile tradeoff with respect to computational expense.

From all r topic models, we then construct the set T consisting of all v unique terms that appear in all topic descriptors across these models. Next, we construct a $v \times v$ symmetrical term co-association matrix **C**, where an entry C_{ij} provides a count of the number of times that a pair of terms i and j have appeared together in a topic descriptor. These counts are normalized with respect to the total number of models r. This provides a measure of the stability of the association between these terms. An illustration of the structure of **C** is given in Fig. 1.

To incorporate additional semantic information, we use a word embedding model generated on an appropriate background corpus. From this, we construct a new $v \times v$ symmetrical term similarity matrix **S**, where the entries S_{ij} correspond to the cosine similarity score between the embedding vectors for all pairs of terms in the set T. An illustration of the structure of **S** is given in Fig. 1. A new *weighted term co-association matrix*, denoted **L**, can be derived by multiplying the original term co-association matrix **C**, and the term similarity matrix **S**, as seen in Eq. (1). The rationale for the weighting scheme is that, if two terms appear frequently together and have a high similarity score, then the pair will have a higher value which indicates that they are more stable and coherent. However, if the terms have a low co-association value and/or low semantic similarity, then this indicates that they are unstable and/or have low coherence. Formally, the calculation is given as:

$$L_{ij} = C_{ij} \times S_{ij} \equiv coassoc(i,j) \times sim(i,j) \quad (1)$$

Not only does the matrix **L** provide us with information about the stability and coherence of pairs of terms in a collection of topic models, but next we show that it can also be used to extract a set of *k ensemble topic descriptors* and further aid in their interpretability. One potential concern about the weighted term co-association matrix **L** would be if its composition was identical to the term co-association matrix **C**, which would occur if all values in **S** had a cosine similarity of 1. This would suggest that the semantic similarity information provided by the word embedding would have no effect on the end result. However, in practice this is unlikely, as the cosine similarity scores between the set of T represent a varied distribution of values calculated based on the dense embedding vectors, rather than raw co-occurrence counts.

3.2. Ensemble topic descriptors

To generate each ensemble topic, we first choose the pair of terms with the highest weighted co-association score in **L** as a pair of seed terms, so long as they have not been used as the seed for a previous ensemble topic. While the number of terms in the topic is less than the total desired length of the descriptor t, we iterate through all of the terms in T that are not currently in the ensemble topic, and add the term that has the highest weighted co-association score with respect to all of the terms that are already in the topic, based on the scores in **L**. This process is then repeated until k ensemble topics have been generated. It is important to note that the ordering of the terms in each topic is determined

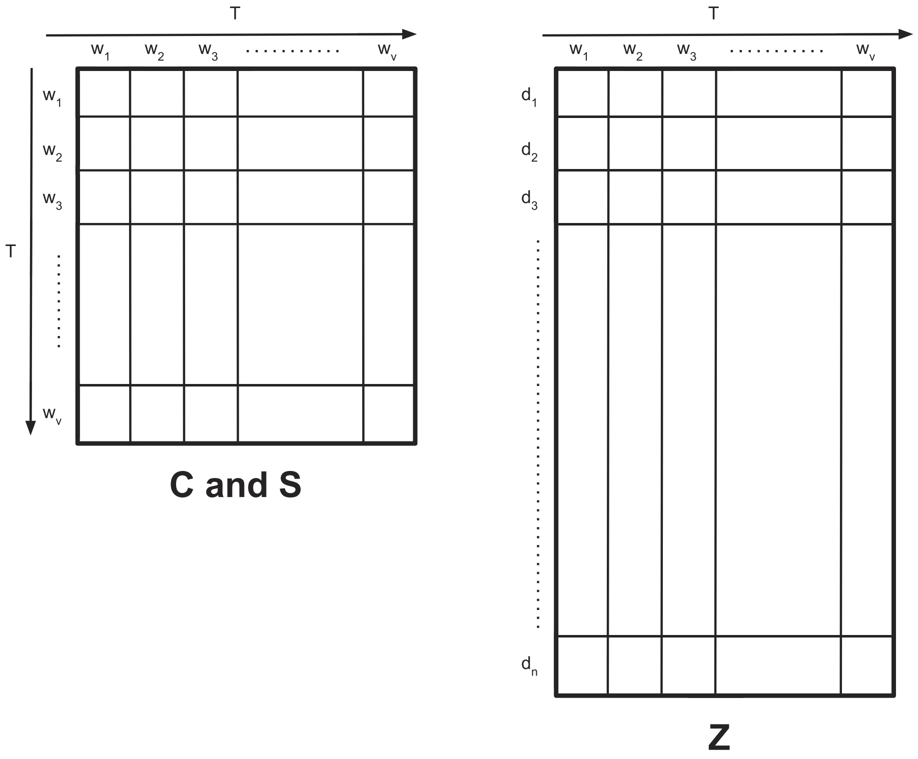


Fig. 1. On the left, an illustration of the structure of the $v \times v$ term co-association matrix **C** and the $v \times v$ term similarity matrix **S**, that are used to create the weighted term co-association matrix **L**. On the right, an illustration of the $n \times v$ document-term co-association matrix **Z** used to extract the primary ensemble topic assignment for each document.

by the order in which the clusters are constructed. A full overview of this process is given in Fig. 2.

The value of t is commonly set to a relatively low value, such as 10 (Lau & Baldwin, 2016), to improve topic interpretation. In initial experiments we investigated the effects of increasing the value of t, however, we observed that this produces more generic ensemble topics with many overlapping terms which can have a significant impact on the resulting coherence scores.

It is important to note that in our upcoming experiments that we focus on a static value of k for each corpus, which is based on associated ground truth information. We produce models using the "correct" number of topics as we wish to evaluate the validity of our proposed approach with respect to these labels. However, in practice a range of k values can be utilized in automated evaluation processes as seen in previous work (Belford et al., 2018) and for manual human interpretation based on expert knowledge.

These ensemble topics allow for the strengths and weaknesses of a solution to be identified through the visualization of the discovered topical structure, the ranking of these ensemble descriptors, the exploration of the best and worst pairs of terms for each topic, and the extraction of topic level stability and coherence scores, which will be discussed in further detail in upcoming sections.

3.3. Ensemble document-topic assignments

Once a set of ensemble topic descriptors has been generated, we can produce ensemble document-topic weights, which reflect the strength of association between each document and these k topics. Using the previously generated collection of r base topic models and the corresponding r disjoint partitions of primary topic assignments for each document, we first build an $n \times v$ document-term co-association matrix $\mathbf{Z}$. Each entry of the matrix Z_{ij} provides a count of the number of times that term j has been present in the primary topic descriptor assigned to document i over all r base runs, based on the information contained in the r partitions. These counts are then normalized by the number of base runs in which a document has appeared. This value is usually r if the base models were all generated on the entire corpus. However, this value can be less than r if document sampling was applied to generate each base model. An illustration of the structure of $\mathbf{Z}$ can be seen in Fig. 1.

Using this document-term co-association matrix $\mathbf{Z}$, along with the previously generated ensemble topic descriptors, we are then able to produce a primary ensemble topic assignment for each document. To generate these assignments we first construct an empty $n \times k$ document-topic matrix $\mathbf{W}$, where each column of the matrix will store the weights of the n documents for a given ensemble topic. To generate each column of $\mathbf{W}$, we identify the t columns from the document-term co-association matrix $\mathbf{Z}$, that correspond to the terms in the associated ensemble topic descriptor. We then compute the mean of these vectors of weights to produce a final column of document weights in $\mathbf{W}$. This column is then normalized by the number of terms t in the ensemble topic descriptor. Once fully constructed, $\mathbf{W}$ can be utilised like a traditional document-topic matrix and a final disjoint partition of primary ensemble topic assignments can be extracted. An overview of this process can be seen in Fig. 3.

3.4. Ranking ensemble topics

With the set of ensemble topics identified, it is important to be able to further quantify their individual quality and produce an ordered ranking of the descriptors, to aid users in discovering which topics are performing well while also facilitating easier interpretation for larger numbers of topics. We can identify the quality of a given pair of terms by extracting their previously-calculated pairwise weighted co-association score from $\mathbf{L}$. By repeating this process for all unique pairs of the t terms in an ensemble topic, we can calculate a Mean Descriptor Score (MDS) which will serve as our quality score for a single ensemble topic:

$$MDS\,Score = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} L_{ij} \qquad (2)$$

MDS scores can be calculated for each ensemble topic descriptor, and a ranking of these descriptors can be achieved by sorting these scores in descending order.

3.5. Ranking term pairs

When ranking a set of descriptors, it is possible to identify topics with low quality scores. When this occurs it would be beneficial to gain a further understanding as to why this is occurring. As previously discussed, for any given pair of terms in T, we can extract their corresponding weighted co-association score from $\mathbf{L}$. This process can be repeated for all unique pairs of terms in a topic descriptor and a ranked list of these pairwise quality scores can be presented to the user to facilitate a further investigation into why a topic is performing poorly. This is useful from not only a topic quality perspective but it also provides an insight into how the word embedding model represents the data and can also serve as an indicator to identify pairs of terms with weighted co-association scores that differ from user expectations or domain knowledge.

```
1. For each ensemble topic, in the range $i = 1$ to $k$:
    1. Choose the pair of terms with the highest score according to $\mathbf{L}$ as the seed for the $i$-th ensemble topic, such that this pair has not been previously used as a seed for a previous topic.
    2. While the number of terms in the $i$-th ensemble topic $< t$:
        (a) From all terms in $T$ that are not already in the current ensemble topic, select the term with the highest similarity with respect to all terms already in the ensemble topic, based on the scores in $\mathbf{L}$.
        (b) Add the selected term to the $i$-th ensemble topic.
```

Fig. 2. Summary of the method for extracting k ensemble topics from the weighted term co-association matrix $\mathbf{L}$.

```
1. Construct an empty $n \times k$ document-topic matrix $\mathbf{W}$.
2. For each column in $\mathbf{W}$, in the range $i = 1$ to $k$:
    1. Identify the $t$ corresponding columns from the document-term co-association matrix $\mathbf{Z}$ that correspond to the terms in the $i$-th ensemble topic descriptor.
    2. Compute the mean of these column vectors to produce the $i$-th column of document weights in $\mathbf{W}$.
    3. Normalize this column by the number of terms $t$ in the ensemble topic descriptor.
3. Extract the disjoint partition of primary ensemble document-topic assignments from $\mathbf{W}$.
```

Fig. 3. Summary of the method for extracting a disjoint partition of primary ensemble document-topic assignments using the ensemble topic descriptors.

3.6. Ensemble topic stability and coherence scores

It is also possible using the matrices **C** and **S**, to extract a mean stability and coherence score for an ensemble topic descriptor. For each pair of terms in a topic we can extract their pairwise scores from a given component matrix. By repeating this process for all unique pairs of terms in the descriptor we can calculate an overall stability and coherence score as seen in Eqs. (3) and (4) respectively. This is useful for the interpretation of solutions, and can be used to determine if one or both of the component scores are impacting a descriptors MDS score.

$$\textit{Topic Stability} = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} C_{ij} \tag{3}$$

$$\textit{Topic Coherence} = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} S_{ij} \tag{4}$$

3.7. Model exploration

Fig. 4 provides a summary of all previously described methodological steps that are utilised in the WTCA ensemble topic modeling approach, including generating the base models, extracting the ensemble topic descriptors and the partition of primary ensemble document-topic assignments, quantitatively ranking the topics and their term pairs, before finally calculating descriptor stability and coherence scores.

Once the steps in Fig. 4 are complete, we shift our focus to the exploration of the outputs of the process. We can reorder the rows and columns (*i.e.* set of T terms) of the weighted term co-association matrix, where this reordering is based on the previously generated ranking of the ensemble topics, and the ordering of the terms in these descriptors. This allows us to represent **L** as a heatmap, where the underlying block structure corresponds to the ensemble topics, which are displayed in descending order of quality, allowing a user to gain a further understanding of the topical structure. In the heatmap, the saturation of each cell indicates the weighted co-association scores for a pair of terms, with a darker saturation corresponding to a higher score, which represents a pair of terms that exhibit higher inherent stability and coherence across r topic modeling solutions.

1. For a given corpus:
    1. Generate r base runs of a chosen topic modeling approach on the corpus.
    2. Construct the $v \times v$ term co-association matrix $\mathbf{C}$, using the previously generated r base sets of topic descriptors.
    3. Construct the $v \times v$ term similarity matrix $\mathbf{S}$ using a given word embedding model.
    4. Compute the weighted term co-association matrix $\mathbf{L}$ as the product of $\mathbf{C}$ and $\mathbf{S}$.
    5. Extract the ensemble topic descriptors from the weighted term co-association matrix, as described in Figure 2.
    6. Extract the partition of primary ensemble document-topic assignments, as described in Figure 3.
    7. Rank the topic descriptors, based on their MDS scores as seen in Equation 2.
    8. Rank the best and worst pair of terms in each topic descriptor, as described in Section 3.5.
    9. Generate the average stability and similarity score for each ensemble topic descriptor, using Equation 3 and Equation 4 respectively.

Fig. 4. Summary of the Weighted Term Co-association (WTCA) method.

An example of this heatmap and the associated ranked topics used to reorder **L** can be seen in Fig. 5 and Table 1 respectively, where ensemble topics have been generated using the ground truth number of topics, $k = 4$, on the *guardian-2005* dataset. In this example we identify four distinct and relatively high quality topics, corresponding to "technology", "business", "football" and "politics". This block structure suggests that while the topical structure of all descriptors are generally quite good, some pairs of terms which appear together, such as "mr" and "election" in the last ensemble topic, have a much lower weighted co-association score, as indicated by the low saturation of the cells.

4. Evaluation

For our evaluations, we conduct two distinct experiments. In Experiment 1 we examine the extent to which our WTCA method is able to produce ensemble topic models which accurately reflect the ground truth categories in the datasets in question. Subsequently in Experiment 2 we investigate the ability of WTCA to produce coherent ensemble topic descriptors. Finally, we further demonstrate the versatility of WTCA by providing a case study in which the discovered topical structure is investigated and evaluated in further detail.

4.1. Datasets

For our experiments we collected 1,595,844 news articles using The Guardian API[2], published from 2004 to 2018, covering a diverse range of themes from politics to entertainment. From the complete set of news articles, we created 15 corresponding yearly datasets, where associated editorial-curated category metadata was used to assign documents into ground truth categories (*e.g.* "politics", "technology", "football"). To ensure that each category contained a suitable number of documents for the topic modeling process, we removed any categories with less than 1,000 articles for each dataset. Details of these datasets are given in Table 2.

We also used a set of four Reddit datasets[3], where user posts were collected from a variety of subreddits covering themes such as "hobbies" and "technology". The inherent information of which subreddit the posts originated from can be used as ground truth categories. We also include eight of the diverse corpora that were used in our previous topic modeling stability experiments (Belford et al., 2018). Details of these datasets can be seen in Table 3.

Prior to topic modeling, we apply standard preprocessing techniques to each dataset. This included filtering terms that appear in less than 20 documents, filtering using an English stopword list of 422 words, applying log-based Term Frequency-Inverse Document Frequency (TF-IDF) weighting, and applying document length normalization. The preprocessed versions of the datasets are made available for use[4].

4.1.1. Word embedding models

For our experiments, we use three different pre-trained word embedding models[5], each constructed using the word2vec Continuous Bag-Of-Words (CBOW) architecture, with 100 dimensions and a window size of 5. The first of these models are trained using the

[2] https://open-platform.theguardian.com

[3] Subsets of the collection provided here https://www.kaggle.com/mswarbrick-jones/reddit-selfposts

[4] Available from http://erdos.ucd.ie/co-association

[5] 4

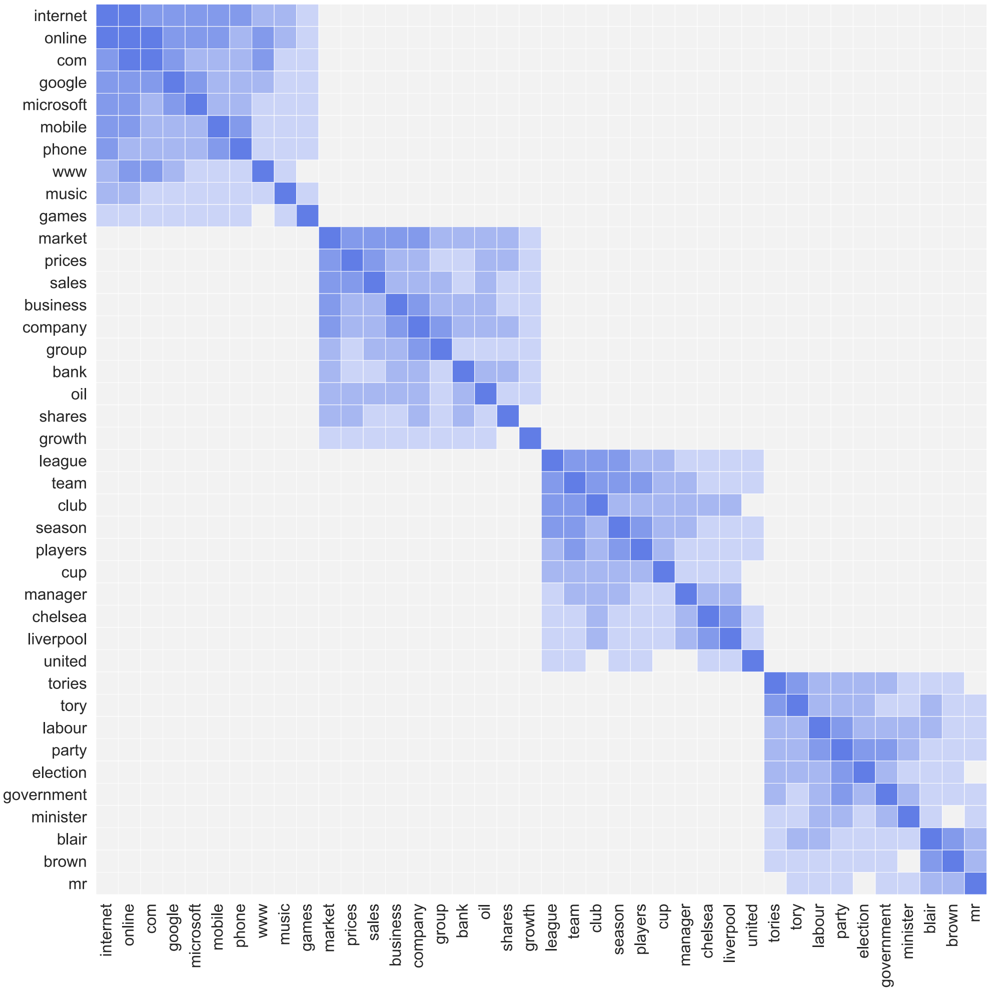


Fig. 5. Example weighted term co-association heatmap, for the *guardian-2005* dataset, that visualizes the discovered topical structure, where the saturation of each cell indicates the weighted co-association score for a pair of terms. The block structures correspond to the ensemble topics.

Table 1
Ensemble topic descriptors, ranked by MDS score, for $k = 4$ topics generated on the *guardian-2005* corpus of news articles, along with the highest and lowest scored pairs of terms.

MDS Score	Topic Descriptor	Best Pair	Worst Pair
0.4282	internet, online, com, google, microsoft, mobile, phone, www, music, games	(internet, online)	(games, www)
0.3707	market, prices, sales, business, company, group, bank, oil, shares, growth	(market, prices)	(growth, shares)
0.3657	league, team, club, season, players, cup, manager, chelsea, liverpool, united	(league, team)	(cup, united)
0.3472	tories, tory, labour, party, election, government, minister, blair, brown, mr	(tories, tory)	(election, mr)

complete Guardian corpus previously mentioned in Section 4.1. The second model was trained using a large collection of Wikipedia long abstracts collected in 2016 (Qureshi & Greene, 2018), while the third embedding was trained on a corpus of CNN and Daily Mail news articles previously compiled by Hermann et al. (2015). Details for these embeddings are given in Table 4.

4.2. Experiment 1: document assignment analysis

In this section we evaluate the accuracy of the document assignments generated by WTCA. For each dataset we can convert the document-topic assignments produced by a topic modeling approach into a disjoint partition consisting of the primary topic label assigned to each document. This partition can be compared against a separate disjoint partition of associated ground truth categories using Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002), which is a standard clustering agreement measure.

4.2.1. Experimental setup

For this experiment we compare four different topic modeling approaches:

1. NMF with random initialization, using the fast alternating least squares variant proposed by Lin (2007) and provided by the *sckit-learn* toolkit (Pedregosa et al., 2011).
2. LDA with random initialization, using fast Gibbs sampling and provided by the *Mallet* software package (McCallum, 2002).
3. KFold ensemble topic modeling for matrix factorization combined with improved initialization, as described in previous work (Belford et al., 2018).
4. The proposed WTCA method, previously described in Section 3.1.

For these approaches, there are a number of common and distinct parameters which need to be specified:

Table 2
Details of the 15 Guardian evaluation corpora used in our experiments, including the total number of documents n, number of terms m, and number of categories $\hat{k}$ in the associated "ground truth" annotations.

Corpus	n	m	$\hat{k}$
guardian-2004	18,209	20,189	5
guardian-2005	17,311	17,389	4
guardian-2006	24,338	22,485	6
guardian-2007	28,218	27,048	6
guardian-2008	36,774	30,577	8
guardian-2009	30,411	26,823	7
guardian-2010	25,164	25,422	6
guardian-2011	20,840	24,006	5
guardian-2012	28,820	28,781	7
guardian-2013	22,139	24,811	5
guardian-2014	28,774	29,116	7
guardian-2015	32,593	32,097	7
guardian-2016	30,634	31,055	7
guardian-2017	17,918	23,279	5
guardian-2018	15,334	21,520	5

Common parameters: For all approaches, the number of topics k is set to correspond to the number of ground truth categories for each dataset.

NMF parameters: For NMF with random initialization, the maximum number of iterations is set to 100 by default. A different random seed is used for each run to populate values in the initial factors **W** and **H**. This process is repeated for $r = 100$ runs.

LDA parameters: For LDA we utilise random initialization and the maximum number of iterations is set to 1000. The alpha and beta parameters are set to 5 and 0.01 respectively. This process is repeated for $r = 100$ runs.

KFold ensemble parameters: For this approach, we apply $p = 10$ rounds of $f = 10$ folds, thus also yielding a collection of 100 ensemble members for integration, with k determined as the same number of ground truth categories for each dataset. This entire process is repeated 20 times.

WTCA parameters: We integrate a collection of 100 members, which are generated via random NMF initialization where each run uses a random 80% sample of the documents. A given word embedding is used to calculate the pairwise term semantic similarity scores as part of the WTCA method, while the final number of topics k is set to be the same as the number of ground truth categories for each dataset. This entire process is repeated 20 times.

4.2.2. Discussion of results

For this experiment we use all 27 datasets previously discussed and report the mean NMI score for all runs of each topic modeling approach in Table 5, where the approach with the highest NMI score for a dataset is highlighted in bold. As the WTCA approach requires a given word embedding to generate the ensemble topic descriptors, and thus the final document-topic partition, we report the NMI scores for WTCA in combination with each of the three word embeddings previously discussed in Section 4.1.1. It is clear to see that a variant of WTCA performs the best by producing the highest partition accuracy in 21 of the 27 datasets. Our previously proposed KFold ensemble approach also performs relatively well, producing better partition scores in 5 of these datasets. It is worth noting that in a small number of cases that this difference in the NMI scores for the KFold approach can be much larger than WTCA, suggesting a better model has been found. However, the weighted co-association approach provides more information regarding the topical structure and the quality of the model, which these other methodologies do not allow for, so a small decrease in NMI might represent a reasonable trade-off. While randomly-initialized NMF performs poorly with regards to the ensemble approaches, randomly-initialized LDA performs the worst across the majority of datasets. It should be noted that these results vary slightly from previous NMI experiments (Belford et al., 2018) which were carried out on a subset of the same datasets due to a larger and more general set of stopwords being used in the preprocessing stage.

These results are also summarised in a ranking table, as seen in Table 6. In this table a score is assigned to each dataset, for each topic modeling approach, with a score of 1 representing the best performing approach. The mean of these scores for each topic modeling methodology are then reported. In this case we can

Table 3
Details of 12 evaluation datasets, including 4 Reddit corpora and 8 corpora used in previous stability experiments, including the total number of documents n, number of terms m, and number of categories $\hat{k}$ in the associated "ground truth" annotations.

Corpus	n	m	$\hat{k}$	Description
bbc	2,225	3,078	5	General news articles from the BBC from 2003.
bbcsport	737	936	5	Sports news articles from the BBC from 2003.
guardian13	6,520	10,739	6	Corpus of news articles published by The Guardian during 2013.
irishtimes2013	3,246	4775	7	Corpus of news articles published by The Irish Times during 2013.
nytimes1999	9,551	12,927	4	A subset of the New York Times Annotated Corpus from 1999.
nytimes2003	11,527	14,939	7	A subset of the New York Times Annotated Corpus from 2003.
reddit-general	15,000	5,377	15	A collection of Reddit posts from 15 general interest subreddits.
reddit-hobbies	6,000	2,381	6	A collection of Reddit posts from 6 hobby-related subreddits.
reddit-sports	9,000	3,535	9	A collection of Reddit posts from 9 sports-related subreddits.
reddit-tech	7,000	2,262	7	A collection of Reddit posts from 7 technology-related subreddits.
wiki-high	5,738	17,234	6	Subset of 2014 Wikipedia dump, where articles are assigned labels based on their high level WikiProject.
wiki-low	4,986	15,368	10	Subset of 2014 Wikipedia dump, where articles are labeled with fine-grained WikiProject sub-groups.

Table 4
Details of the three background corpora used in our experiments to generate word embedding models, including the total number of documents n and the number of terms m.

Corpus	n	m	Description
guardian15	1,595,844	557,937	Collection of 15 years of Guardian news articles.
wikipedia2016	4,899,998	1,333,306	Collection of Wikipedia long abstracts.
cnn-dailymail	312,085	243,863	Collection of CNN and Daily Mail news articles

Table 5
Comparison of document partition accuracy scores using a number of topic modeling approaches, over all 27 corpora, based on mean Normalized Mutual Information (NMI) with respect to the number of ground truth categories $\hat{k}$.

Corpus	NMF	LDA	KFold	WTCA (Wiki)	WTCA (Guard15)	WTCA (CD)
guardian-2004	0.7375	0.7044	0.7382	0.7596	**0.7606**	0.7592
guardian-2005	0.7532	0.7138	**0.7767**	0.7496	0.7496	0.7494
guardian-2006	0.7054	0.6508	0.6957	**0.7199**	0.7195	0.7186
guardian-2007	0.7177	0.6896	0.7647	0.7658	**0.7699**	0.7650
guardian-2008	0.7151	0.6724	0.7305	0.7692	0.7666	**0.7737**
guardian-2009	0.7039	0.6719	0.7307	**0.7466**	0.7429	0.7458
guardian-2010	0.7013	0.6930	0.7364	0.7348	0.7357	**0.7383**
guardian-2011	0.7551	0.7134	**0.8720**	0.7970	0.8333	0.8230
guardian-2012	0.7207	0.6776	0.7563	0.7591	**0.7599**	0.7596
guardian-2013	0.7708	0.7465	**0.8538**	0.7788	0.7829	0.7812
guardian-2014	0.7042	0.6834	0.7239	0.7340	**0.7357**	0.7356
guardian-2015	0.7053	0.6522	0.6845	0.7296	**0.7358**	0.7304
guardian-2016	0.6786	0.6534	0.6930	**0.7109**	0.6989	0.6928
guardian-2017	0.7128	**0.7315**	0.7175	0.7146	0.7139	0.7137
guardian-2018	0.6993	0.7106	0.6345	0.7346	0.7378	**0.7406**
bbc	0.7735	0.7326	0.7843	**0.7879**	0.7877	0.7876
bbcsport	0.8050	0.5581	0.8457	0.8476	0.8484	**0.8487**
guardian13	0.8178	0.7248	0.8484	0.8503	0.8503	**0.8504**
irishtimes2013	0.7249	0.6398	0.7667	**0.7678**	0.7640	0.7671
nytimes1999	0.5863	0.6188	0.6713	**0.6721**	0.6054	0.6560
nytimes2003	0.6098	0.5887	0.6034	0.6160	0.6035	**0.6183**
reddit-general	0.8706	0.6892	0.8966	0.8963	**0.8970**	0.8969
reddit-hobbies	0.8724	0.6507	0.8888	**0.8932**	0.8924	0.8922
reddit-sports	0.7602	0.4789	0.7675	0.7750	0.7703	**0.7754**
reddit-tech	0.6763	0.5028	0.6786	0.6959	**0.7007**	0.6997
wikihigh	0.7255	0.6998	**0.7443**	0.7434	0.7435	0.7432
wikilow	0.8665	0.8321	**0.8870**	0.8809	0.8441	0.8785

Table 6
Ranking of document partition accuracy scores for a number of topic modeling approaches, over all 27 corpora, based on mean Normalized Mutual Information (NMI) with respect to the number of ground truth categories $\hat{k}$.

Corpus	NMF	LDA	KFold	WTCA (Wiki)	WTCA (Guard15)	WTCA (CD)
guardian-2004	5	6	4	2	**1**	3
guardian-2005	2	6	**1**	3	4	5
guardian-2006	4	6	5	**1**	2	3
guardian-2007	5	6	4	2	**1**	3
guardian-2008	5	6	4	2	3	**1**
guardian-2009	5	6	4	**1**	3	2
guardian-2010	5	6	2	4	3	**1**
guardian-2011	5	6	**1**	4	2	3
guardian-2012	5	6	4	3	**1**	2
guardian-2013	5	6	**1**	4	2	3
guardian-2014	5	6	4	3	**1**	2
guardian-2015	4	6	5	3	**1**	2
guardian-2016	5	6	3	**1**	2	4
guardian-2017	6	**1**	2	3	4	5
guardian-2018	5	4	6	3	2	**1**
bbc	5	6	4	**1**	2	3
bbcsport	5	6	4	3	2	**1**
guardian13	5	6	4	3	2	**1**
irishtimes2013	5	6	3	**1**	4	2
nytimes1999	6	4	2	**1**	5	3
nytimes2003	3	6	5	2	4	**1**
reddit-general	5	6	3	4	**1**	2
reddit-hobbies	5	6	4	**1**	2	3
reddit-sports	5	6	4	2	3	**1**
reddit-tech	5	6	4	3	**1**	2
wikihigh	5	6	**1**	3	2	4
wikilow	4	6	**1**	2	5	3
Mean Score	4.78	5.67	3.30	**2.41**	**2.41**	2.44

quantitatively see that WTCA performs the best with regards to all 27 datasets, regardless of the word embedding used.

4.3. Experiment 2: coherence analysis

In this section we examine the coherence of WTCA with respect to associated ground truth labels.

4.3.1. Experimental setup

For this experiment we compare the same four topic modeling approaches: NMF, LDA, KFold and WTCA as previously described in Experiment 1. The parameter settings of all approaches remain the same as before.

We consider a subset of the datasets previously described in Section 4.1 for this experiment, specifically the 15 yearly Guardian datasets. The rationale for this is that one of our word embedding models, *guardian15*, is trained using the full superset of 1.5 million news articles that were collected from the Guardian API, while the remaining two word embeddings, *wikipedia2016* and *cnn-dailymail*, are also domain appropriate for these datasets. We also drop the remaining datasets as there may be temporal issues due to when some of the datasets were originally collected, and geographical language differences depending on the country of origin.

We examine the extent to which the ensemble topic descriptors produced by WTCA are coherent. We use a previously proposed topic coherence measure (O'Callaghan et al., 2015) that utilizes a pre-trained word embedding model to calculate the semantic coherence of a topic. This is an intra-topic quality measure in which the coherence of a given pair of terms is calculated as the cosine similarity of their corresponding vectors in the chosen word embedding model. This process can be repeated for each unique pair of terms in a topic descriptor and a final overall topic coherence score can be computed as the average of these pairwise scores. An overall model level score can also be calculated by computing the average of these individual topic scores. As previously mentioned, each dataset has associated ground truth annotations, in which the "correct" number of topics is known in advance. With this in mind, we investigate the coherence of the produced ensemble topic descriptors for the ground truth number of topics for each of the Guardian datasets.

It is important to note that WTCA utilizes a chosen embedding model to construct the ensemble topics. By using this coherence metric we are now also using an embedding model to calculate a final coherence score. Therefore, we also investigate the effect of using different combinations of embeddings (*i.e.* one embedding to generate the ensemble topics and another to evaluate them) in order to provide a thorough evaluation of our approach. This ensures that we are not "overfitting" the data by utilizing the same embedding for both steps.

Since we use three different embeddings in our evaluation, we separate our results into three corresponding sets, as listed in Tables 7–9 respectively. In each table, for the ground truth number of topics associated with each dataset, we first report the average coherence scores that have been calculated using the previously generated 100 runs of randomly-initialized NMF and LDA as a baseline, and then report the average coherence score of the KFold approach, all of which use the chosen embedding model to calculate coherence. We then report the results for each of the WTCA variants, where the first embedding is used to construct the ensemble topics, while the second embedding is used to calculate the coherence (*e.g.* Guard15/Wiki denotes using the *guardian15* embedding to build the ensemble topics, and the *wikipedia2016* embedding to evaluate them). In this respect, we are "training" our WTCA approach using three different embeddings, and then comparing or "testing" it on a single embedding to investigate its performance.

It is important to note that it is possible when constructing the term similarity matrix **S**, or when calculating the coherence of the final ensemble topic descriptors, that terms may not be present in

Table 7
Comparison of model coherence scores, evaluated with respect to the *guardian15* word embedding, for the ground truth number of categories $\hat{k}$ for each yearly Guardian dataset.

Corpus	$\hat{k}$	NMF	LDA	KFold	Guard15/Guard15	Wiki/Guard15	CD/Guard15
					(WTCA)	(WTCA)	(WTCA)
guardian-2004	5	0.3853	0.2981	0.3721	0.4169	0.3971	**0.4228**
guardian-2005	4	0.4019	0.3050	0.4019	0.4039	0.4028	**0.4182**
guardian-2006	6	0.4213	0.3143	0.4015	0.4440	0.4385	**0.4442**
guardian-2007	6	0.3958	0.3331	0.3862	0.4270	0.4054	**0.4322**
guardian-2008	8	0.4497	0.3548	0.4499	0.4814	0.4732	**0.4871**
guardian-2009	7	0.4485	0.3388	0.4797	0.4992	0.4881	**0.5058**
guardian-2010	6	0.4269	0.3265	0.4132	0.4593	0.4496	**0.4700**
guardian-2011	5	0.4482	0.3508	0.4917	**0.5033**	0.4829	0.4987
guardian-2012	7	0.4302	0.3341	0.4409	0.4762	0.4604	**0.4847**
guardian-2013	5	0.4408	0.3563	**0.4803**	0.4660	0.4309	0.4559
guardian-2014	7	0.4625	0.3546	0.4742	**0.4925**	0.4901	0.4904
guardian-2015	7	0.4353	0.3502	0.4337	0.4569	**0.4656**	0.4477
guardian-2016	7	0.3947	0.3504	0.4083	0.4111	**0.4236**	0.4122
guardian-2017	5	0.4276	0.3625	0.4226	**0.4434**	0.4351	0.4361
guardian-2018	5	0.3846	0.3218	0.4042	0.4110	0.3824	**0.4184**

Table 8
Comparison of model coherence scores, evaluated with respect to the *cnn-dailymail* word embedding, for the ground truth number of categories $\hat{k}$ for each yearly Guardian dataset.

Corpus	$\hat{k}$	NMF	LDA	KFold	Guard15/CD	Wiki/CD	CD/CD
					(WTCA)	(WTCA)	(WTCA)
guardian-2004	5	0.3397	0.2769	0.3300	0.3719	0.3559	**0.3825**
guardian-2005	4	0.3520	0.2786	0.3520	0.3544	0.3532	**0.3709**
guardian-2006	6	0.3725	0.2854	0.3508	0.3951	0.3895	**0.3960**
guardian-2007	6	0.3496	0.3084	0.3392	0.3822	0.3598	**0.3953**
guardian-2008	8	0.3847	0.3113	0.3896	0.4224	0.4184	**0.4347**
guardian-2009	7	0.3908	0.3064	0.4302	0.4480	0.4392	**0.4613**
guardian-2010	6	0.3865	0.2941	0.3735	0.4225	0.4122	**0.4392**
guardian-2011	5	0.3910	0.3164	0.4405	**0.4458**	0.4237	0.4418
guardian-2012	7	0.3805	0.3058	0.3935	0.4236	0.4103	**0.4432**
guardian-2013	5	0.3908	0.3225	**0.4257**	0.4119	0.3877	0.4161
guardian-2014	7	0.4241	0.3338	0.4417	0.4578	0.4658	**0.4687**
guardian-2015	7	0.3778	0.3026	0.3786	0.3973	**0.4081**	0.3981
guardian-2016	7	0.3379	0.3000	0.3475	0.3495	**0.3618**	0.3579
guardian-2017	5	0.3559	0.3037	0.3549	0.3732	0.3663	**0.3790**
guardian-2018	5	0.3103	0.2724	0.3049	0.3367	0.3165	**0.3639**

Table 9
Comparison of model coherence scores, evaluated with respect to the *wikipedia2016* word embedding, for the ground truth number of categories $\hat{k}$ for each yearly Guardian dataset.

Corpus	$\hat{k}$	NMF	LDA	KFold	Guard15/Wiki	Wiki/Wiki	CD/Wiki
					(WTCA)	(WTCA)	(WTCA)
guardian-2004	5	0.3742	0.3614	0.3671	**0.4069**	0.4031	0.4040
guardian-2005	4	0.3939	0.3415	0.3939	0.3951	0.3951	**0.4015**
guardian-2006	6	0.4084	0.3731	0.3919	**0.4449**	0.4434	**0.4449**
guardian-2007	6	0.3933	0.3788	0.3952	0.4172	0.4049	**0.4198**
guardian-2008	8	0.4148	0.3994	0.4259	0.4451	0.4444	**0.4493**
guardian-2009	7	0.3980	0.3909	0.4380	0.4477	0.4484	**0.4569**
guardian-2010	6	0.3725	0.3703	0.3668	0.3921	**0.4100**	0.4004
guardian-2011	5	0.3935	0.4011	**0.4564**	0.4421	0.4200	0.4349
guardian-2012	7	0.3769	0.3890	0.4079	0.4112	**0.4186**	0.4135
guardian-2013	5	0.3871	0.3892	**0.4286**	0.3954	0.3898	0.4025
guardian-2014	7	0.4056	0.3992	0.4180	0.4326	**0.4423**	0.4369
guardian-2015	7	0.3690	0.3770	0.3645	0.3956	**0.4079**	0.3862
guardian-2016	7	0.3543	0.3814	0.3710	0.3694	**0.3932**	0.3698
guardian-2017	5	0.3819	0.3799	0.3834	0.3937	**0.3963**	0.3918
guardian-2018	5	0.3416	0.3589	0.3406	0.3715	0.3714	**0.3750**

the vocabulary of the respective word embedding used, also known as being out-of-vocabulary. However, due to the large vocabulary size of the respective reference corpora used to construct each embedding, and these reference corpora being domain specific with respect to our chosen datasets, we find that this issue rarely occurs. In the case of our experiments, if a term is missing from an embedding vocabulary then we assign it a cosine similarity score of 0 with all other terms. However, it should be noted that there are alternative methodologies available for dealing with this out-of-vocabulary issue (Bahdanau et al., 2017).

4.3.2. Discussion of results

It is interesting to note that, in each set of results, a variant of WTCA always outperforms traditional NMF and LDA. This clearly shows that WTCA produces more coherent topics with respect to the ground truth number of topics in each dataset. It is also interesting to note that when the WTCA approach is "trained" using the *cnn-dailymail* word embedding, it frequently produces the most coherent descriptors, regardless of the "testing" embedding. One explanation for this may be due to the background reference corpus for this embedding consisting of news articles written in both British and American English, and thus the embedding is able to better capture the semantic similarities of terms. However, in the majority of cases, the difference between these WTCA variants when using a combination of different embeddings is usually minimal, suggesting that any combination of embeddings can be used and more coherent descriptors will frequently be produced than randomly initialized NMF and LDA. It is interesting to note that LDA appears to produce the least coherent descriptors for the ground truth number of topics across all of the datasets, however previous research (O'Callaghan et al., 2015) has identified this trend of LDA producing less coherent topics.

4.4. Case study

One of the advantages of WTCA is that we can gain a further understanding of the discovered topical structure and provide further explanation as to how the underlying data is being represented, rather than providing a single score for interpretation. With this in mind we perform a case study of the *guardian-2009* dataset using WTCA, for the ground truth number of topics, $k = 7$, in which we discover an interesting variance from the ground truth information, which can be investigated in further detail. In this case we utilised 100 base runs of randomly-initialized NMF, and generated the ensemble topic descriptors and evaluated their coherence using the *wikipedia2016* word embedding.

We first extract the set of ensemble topic descriptors, as seen in Table 10. It is evident that the discovered topics differ from the associated ground truth labels. Specifically, two distinct and granular business topics, related to banking and trading have been discovered (topics 4 and 6), instead of one broad "business" topic as specified by the provided labels. It is also interesting to note that the "books", and "film" topics have formed a general media topic (topic 5). By observing the ranking of the topics that are also provided in Table 10, we see that this general media topic performs worse with respect to more well defined topics such as "technology" (topic 1) due to the merging of themes. It is also apparent that the "music" topic that has been discovered is poor due to its low MDS score (topic 7). This low score can be investigated further by inspecting the average stability and coherence scores of the descriptor, as provided in Table 11. This clearly shows that the poor MDS score is due to a low inherent stability of the terms across the 100 base runs of NMF. We can further investigate the quality of the "music" topic by visualizing the discovered topical structure by

Table 10
Ensemble topic descriptors, ranked by MDS score, for $k = 7$ on the *guardian-2009* dataset. The best and worst pairs of terms are also listed.

Topic Num.	MDS Score	Topic Descriptor	Best Pair	Worst Pair
1	0.4038	google, microsoft, iphone, apple, windows, users, mobile, internet, twitter, online	(google, microsoft)	(online, windows)
2	0.3524	league, team, club, season, players, game, manager, chelsea, liverpool, united	(league, team)	(club, united)
3	0.3044	labour, party, government, election, tory, mps, minister, cameron, brown, expenses	(government, party)	(expenses, tory)
4	0.2826	bank, banks, banking, financial, government, tax, economy, treasury, lloyds, bonuses	(bank, banks)	(economy, lloyds)
5	0.2640	book, novel, story, life, books, world, film, movie, films, time	(book, novel)	(movie, time)
6	0.2316	company, group, market, shares, ftse, trading, sales, price, down, profits	(company, group)	(down, ftse)
7	0.2291	album, band, songs, music, pop, rock, song, sound, jazz, guitar	(album, band)	(jazz, song)

Table 11
Average stability and coherence scores for the ensemble topic descriptors, extracted from the corresponding **C** and **S** matrices for $k = 7$ on the *guardian-2009* dataset, where topics are ordered based on their ranking in Table 10.

Topic Num.	Topic Stability Score	Topic Coherence Score
1	0.7089	0.5756
2	0.9400	0.3734
3	0.8760	0.3451
4	0.6347	0.4206
5	0.5247	0.4700
6	0.6187	0.3802
7	0.4116	0.5439

producing a heatmap of the weighted term co-association matrix, as seen in Fig. 6. While this visualization highlights a relatively good block structure for said topic, the saturation of the cells are low due to the lower stability previously discussed. This visualization also highlights the previously discussed issue of terms being merged into a broader media topic as it shows a much poorer combined block structure.

As with other topic modeling techniques, the inspection of this dataset in practice may be an iterative process with humans interpreting the results, before either settling on the resulting model or, running the model again for different values of k based on domain or expert knowledge.

5. Conclusions

The evaluation and interpretation of one or more topic models is a difficult task, with typically a single score used to summarize the quality of a model. This is further complicated when different evaluation metrics report varying levels of quality for the same models. To address this issue we have proposed an approach to not only facilitate the generation of more robust and coherent ensemble topic descriptors, but also provide a number of useful evaluation metrics, and an approach to allow for the visualization of the topical structure, based on the inherent stability of terms and semantic similarity information provided by a given word embedding.

We have clearly shown the potential of ensemble topic modeling to generate higher-quality models, with respect to producing more coherent topic descriptors and more accurate final document-topic partitions. We have also demonstrated the potential of this approach to allow for the further interpretation and comprehension of topic modeling solutions through the visualization of the identified topical structure, the ranking of ensemble topic descriptors, the identification of the best and worst pairs of terms in a descriptor, and by providing stability and semantic similarity scores for each topic. When combined, these contributions allow for the further investigation, interpretation and comparison of different topic modeling solutions by the user, especially when the results differ from expectations. While our focus has been on

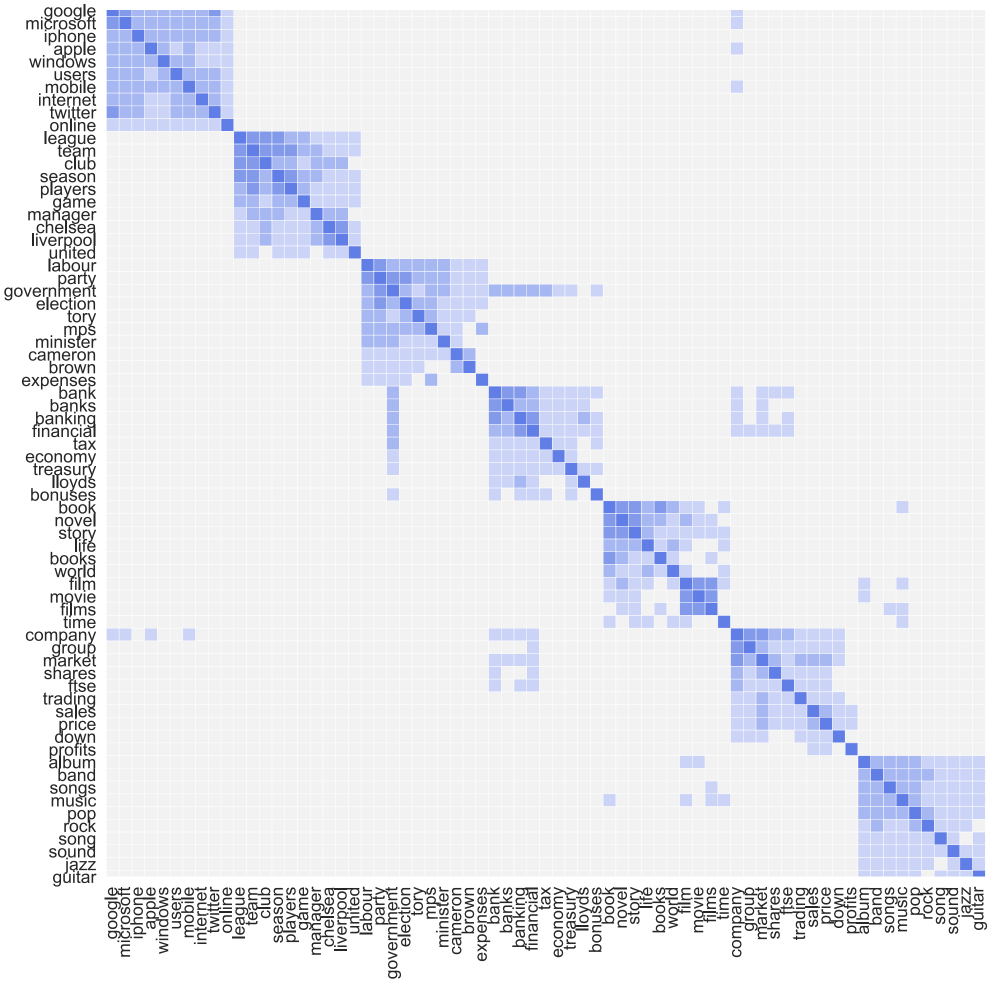


Fig. 6. A heatmap visualization of the weighted term co-association scores for the *guardian-2009* dataset, generated for $k = 7$ topics.

the use of matrix factorization algorithms to generate a collection of base models, the agnostic nature of the proposed approach means that it could be readily applied in conjunction with other types of topic modeling algorithms, due to utilizing the topic descriptors and not the underlying weights or probabilities of the model.

CRediT authorship contribution statement

Mark Belford: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Derek Greene:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by Science Foundation Ireland (SFI) under Grant No. SFI/12/RC/2289_P2.

References

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models – Going beyond SVD. In *Proc. 53rd symposium on foundations of computer science* (pp. 1–10). IEEE.

Bahdanau, D., Bosc, T., Jastrzbski, S., Grefenstette, E., Vincent, P., & Bengio, Y. (2017). Learning to compute word embeddings on the fly. arXiv preprint arXiv:1706.00286.

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications, 91*, 159–169.

Berikov, V., & Pestunov, I. (2017). Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties. *Pattern Recognition, 63*, 427–436.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. In *Proc. 6th international conference on weblogs and social media* (pp. 419–422).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: how humans interpret topic models. *Proc. neural information processing systems*, 288–296.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: visualization techniques for assessing textual topic models. In *Proc. international working conference on advanced visual interfaces* (pp. 74–77). ACM.

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. In *Proc. 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Vol 1: long papers)* (pp. 795–804).

De Waal, A. & Barnard, E. (2008). Evaluating topic models with stability. In *Proc. pattern recognition association of south africa* (pp. 79–84).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Ding, R., Nallapati, R., & Xiang, B. (2018). Coherence-aware neural topic modeling. arXiv preprint arXiv:1809.02687.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Proc. joint european conference on machine learning and knowledge discovery in databases* (pp. 498–513). Springer.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). Topicnets: visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology, 3*(2).

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Proceedings on Neural Information Processing Systems*, 1693–1701.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89–106.

Kim, M., Kang, K., Park, D., Choo, J., & Elmqvist, N. (2016). TopicLens: efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics, 23*(1), 151–160.

Kuang, D., Choo, J., & Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional clustering algorithms* (pp. 215–243). Springer.

Lau, J. H. & Baldwin, T. (2016). The sensitivity of topic coherence evaluation to topic cardinality. In *Proc. 2016 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 483–487).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proc. 18th conference on information and knowledge management* (pp. 375–384). ACM.

Lin, C.-J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation, 19*(10), 2756–2779.

Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., & Lian, X. (2012). Tiara: interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology, 3*(2).

McCallum, A. K. (2002). Mallet: a machine learning for language toolkit.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proc. conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.

Moody, C. E. (2016). Mixing Dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019.

Murdock, J., & Allen, C. (2015). Visualization techniques for topic model checking. In *Proc. 29th AAAI conference on artificial intelligence*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proc. human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108). Association for Computational Linguistics.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, 3*, 299–313.

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications, 42* (13), 5645–5657.

Pauca, V. P., Shahnaz, F., Berry, M. W., & Plemmons, R. J. (2004). Text mining using non-negative matrix factorizations. In *Proc. 2004 SIAM international conference on data mining* (pp. 452–456). SIAM.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Qureshi, M. A., & Greene, D. (2018). EVE: Explainable vector based embedding technique using Wikipedia. Journal of Intelligent Information Systems.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proc. workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Steyvers, M., & Griffiths, T. (2007). Latent semantic analysis: a road to meaning. Laurence Erlbaum, chapter Probabilistic Topic Models.

Strehl, A. (2002). Relationship-based clustering and cluster ensembles for high-dimensional data mining. Ph.D. thesis, University of Texas, Austin.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles–a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*, 583–617.

Velcin, J., Gourru, A., Giry-Fouquet, E., Gravier, C., Roche, M., & Poncelet, P. (2018). Readitopics: make your topic models readable via labeling and browsing. In *Proc. 27th international joint conference on artificial intelligence* (pp. 5874–5876).

Xie, P., Yang, D., & Xing, E. (2015). Incorporating word correlation knowledge into topic modeling. In *Proc. conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 725–734).

Yang, Y., Downey, D., & Boyd-Graber, J. (2015). Efficient methods for incorporating knowledge into topic models. In *Proc. conference on empirical methods in natural language processing* (pp. 308–317).