PAIRWISE PROXIMITY METRICS FOR TOPIC MODELLING EVALUATION BASED ON BERT EMBEDDINGS

Anonymous ACL submission

Abstract

The use of topic modelling methods is a popular way to describe natural language text with a representative set of words. In order to evaluate such methods, objective metrics such as coherence and silhouette scores are commonly used. However, it has been shown that topic assessment based on such metrics does not align well with human judgment for classical document corpora such as articles, books and server logs and, at the same time, it is still unclear how appropriate they are for dialog data. In this paper, we investigate the most commonly used topic modelling evaluation scores in terms of their alignment with human judgment in the specific area of dialog speech. We show that there is still space for improvement in the objective evaluation of topic modelling, and propose a new group of metrics, called Pairwise Proximity metrics, that are shown to align better with human judgment, when compared to coherence and silhouette scores.

1 Introduction

Topic modelling (TM) is an unsupervised machine learning technique that is commonly used to discover the latent semantic structure in a set of documents. TM methods start by scanning the set of documents in order to detect representative word or phrase patterns. These abstract structures, or clusters, are called "topics", and are represented as lists of words. TM methods are used in a variety of tasks, for example in order to organize large collections of documents, for text recommendation, and document ranking. Such solutions find numerous applications in modern technology, for example for the analysis of scientific article collections Murakami et al. (2017), collection of images Feng and Lapata (2010), Argyrou et al. (2018), videos and music and news streams Schinas et al. (2015). TM approaches are also used for genome annotation Stein (2001) and other tasks of bioinformatics Wandy et al. (2018).

Moreover, there are numerous applications of TM in the area of dialog speech recognition and understanding, particularly in the call center paradigm, as it can be used to improve speech recognition output, segment the recognized speech into topics Purver et al. (2006) and either perform quality control by checking the call center agent responses Kalitvianski et al., or help in designing dialogue responses Camilleri (2002), Wang et al. (2017), Valenti et al.. All these applications can greatly benefit the customer experience and support quality offered by call center agents.

Traditionally, most TM approaches can be categorized into two main categories, namely probabilistic topic models and latent semantic analysis. The most common paradigm in the first category is Latent Dirichlet allocation (LDA) Blei et al. (2003), which is exploited in numerous TM papers Tong and Zhang (2016), Bagheri et al. (2014), Mutanga and Abayomi (2020). In the area of latent semantic analysis the most representative example is Non-negative Matrix Factorization (NMF). NMF is particularly useful in assessing the contribution of topics to a document Torres et al. (2017), which, for example, allows to annotate documents with topics Sherstinova et al. (2020). More recently, TM approaches use word embeddings calculated with neural architectures. The authors in Angelov (2020) present a clustering approach for TM and semantic search called Top2Vec Ghasiya and Okamura (2021). The algorithm is based on the assumption that many semantically similar documents indicate an underlying topic. The document word embeddings can be calculated with alternative methods, for example Doc2Vec Řehůřek and Sojka (2010) or Universal Sentence Encoder Cer et al. (2018) or BERT Sentence Transformer Devlin et al. (2018).

Even though extensive research has been conducted in the area of TM, there are still various open problems. First, the results of TM methods are usually topics in the form of keyword lists,

which the model developers must interpret, therefore creating barriers in automation Alokaili et al. (2020). Therefore automatic topic interpretation, aligning with human judgement is still an open area. Second, most statistical TM methods and metrics still present a poor agreement with human judgement in topic understanding. Finally, developing a topic model for dialogue speech is different from topic models for large text corpora. To develop a dialogue thematic model, a wider range of methods, both known and proposed by other scientists, are used.

In this paper, we focus on telephone dialogue speech and we investigate alternative methods to measure the consistency of topic prediction within documents and human judgement. We evaluate various commonly used TM methods and metrics, with regards to human topic annotation. Finally, we propose a new set of metrics that better align with human judgement and are flexible in assessing the consistency between a set of words and a document since only pretrained embeddings are required for their calculation while the time consuming, and often complex, text preprocessing steps are not required.

The remaining of this paper is organized as follows. In Section 2 we discuss the related work in the area of objective TM evaluation metrics. We present the most important literature metrics, which we also investigate in this work, and some previous work on evaluating these objective scores. In Section 3 we present the proposed pairwise proximity scores and their statistical characteristics. In Section 4, we present all the experimental framework, the TM methods, data and human annotations that we use. Moreover, we present various comparisons between objective and subjective TM evaluation scores and show the improvement that we can get using the proposed Pairwise Proximity scores. Finally, conclusions are drawn in Section 5.

2 Objective TM evaluation metrics

The most common way to evaluate the quality of a TM approach are the *co-occurrence* based methods, which use co-occurrence statistics to estimate the semantic similarity of a topic and the documents it has been assigned to. The Coherence metric was first proposed in Wallach et al. (2009), and it is calculated as a sum of the conditional probability of a word in a topic given all other words. Based on this elementary component, more sophisticated Coherence metrics were proposed and evaluated in Newman et al. (2010). Among them, we identify the following list as the most interesting metrics for TM, that we also focus on in the current work.

- C_v is a coherence measure based on a sliding window, one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.
- C_{uci} is a coherence measure based on a sliding window and the pointwise mutual information (PMI) of all word pairs of the given top words.
- C_{umass} is a coherence measure based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure.
- C_{npmi} is an enhanced version of the C_{uci} coherence measure using the NPMI.

In a different group of evaluation metrics is the Silhouette coefficient, S_c, a metric used to calculate the goodness of a clustering technique and can also be applied for TM Wang et al. (2017). Given a cluster point i, the silhouette value is defined as:

$$S(i) = \frac{b - a}{max(a, b)}, \tag{1}$$

where a is the mean distance between i and all other points in the same class and b the smallest mean distance between i and all other points in the other clusters that i is not a member of. The Silhouette coefficient is defined as

$$S_c = \max_k \tilde{S}(k), \tag{2}$$

where $\tilde{S}(k)$ is defined as the mean of $S(i)$ over the whole dataset for k number of clusters.

Among the various automatic topic coherence metrics ten are empirically investigated in Fang et al. (2016a) in terms of their appropriateness, by comparing how closely they align with the human judgment of topic coherence. To evaluate which coherence metrics most closely align with human judgment, they conduct a large-scale empirical crowd-sourced user study to identify the coherence of topics generated by three different TM approaches upon two Twitter datasets. They also use these pairwise coherence preferences to assess the suitability of 10 topic coherence metrics

for Twitter data Peinelt et al. (2020), Fang et al. (2016b). They investigate how much the coherence metric evaluation matches the human judgment.

There are some similar papers which investigate *automatic coherence metrics* as measures of topics interpretability, but to the best of our knowledge there are no papers that investigate how well the predicted topic for each document matches with the text of the document.

3 Pairwise proximity metrics

We propose a new set of objective evaluation metrics for TM method evaluation, which is based on the BERT embeddings of the raw document and topic texts. The proposed metric, called Pairwise Proximity (PP), uses the proximity metric $M_d(T, D)$ between a pair of topic T, and a document D:

$$M_d(T, D) = \frac{1}{N} \sum_{i=1}^{N} \min_j (d_e(W_i^T, W_j^D)) \quad (3)$$

where $d_e(W_i^T, W_j^D)$ is the Euclidean distance between the BERT embeddings of i-th word of the topic T, and the j-th word of the document D. As shown in Fig 1, the values of the metric of Equation 3 range from 0 to ∞, with lower values indicating that more topic words are similar to document words. To change the dynamics and value range we define the PP score as:

$$PP(T, D) = \frac{\ln(e - 1 + \frac{1}{M_d(T,D)+1}) - \ln(e - 1)}{1 - \ln(e - 1)}, \quad (4)$$

for which the values range in the $(0, 1]$. The corresponding distribution is shown in Figure 2. $PP(T, D)$ is multiplied with the Silhouette score with values in the range $(0, 1]$ to create the metric $PP_s(T, D)$, which is experimentally proved to have better agreement with humans.

$$PP_s(T, D) = PP(T, D)\frac{S_c + 1}{2}. \quad (5)$$

It is noted here that instead on the Euclidean distance, similar metrics can be formulated using other distances, for example cosine similarity. We have investigated such alternatives and even though we found some differences in the obtained distributions, the pairwise proximity scores maintain their ability to evaluate TM methods independently from the exact distance metric used.

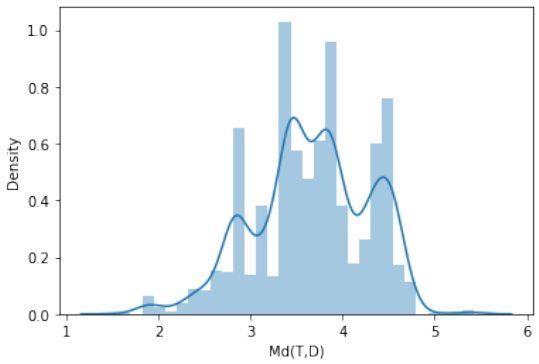

Figure 1: Distribution of $M_d(T, D)$. The values range from 0 to ∞, with larger values indicating lower similarity between the topic and the document.

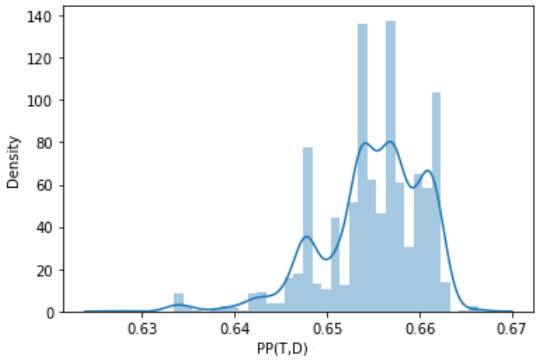

Figure 2: Distribution of $PP(T, D)$. Notice the changed dynamics and the new value range in $(0, 1]$

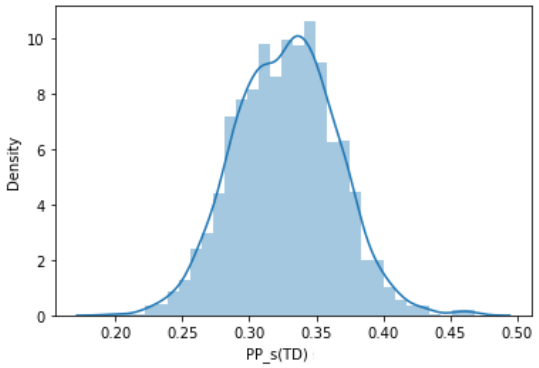

Figure 3: Distribution of $PP_s(T, D)$

4 Experiments and results

The goal of this study is to investigate the appropriateness of various objective TM evaluation metrics, and compare how closely they align with human, *i.e* subjective, evaluation of TM methods. To run these experiments, we implemented three different TM methods, widely used in the literature. In the following sections we describe the TM methods and objective metrics used, and the methodology we followed to create the human evaluations of the same TM methods.

4.1 TM methods and objective evaluation metrics

In this work we implemented three TM approaches, based on NMF, LDA and BERT embeddings (BERTclust) respectively. For LDA and NMF models we tuned the number of topics and the alpha value, based on average coherence scores C_v. The clustering approach BERTclust consists of three steps: calculating data embeddings, dimension reduction of data embeddings by UMAP technique (Uniform Manifold Approximation and Projection McInnes et al. (2018)), data clustering using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications McInnes et al. (2017)). We performed tuning of the number of components, the number of neighbors for the UMAP model and the minimal cluster size for the HDBSCAN model, based on average silhouette scores of non-outlier texts.

To evaluate and compare the selected approaches we use four different Coherence scores, namely the C_v, C_{uci}, C_{npmi} and C_{umass}, and the Silhouette coefficient, S_c. We also use the proposed proximity metrics, $PP(T, D)$ and $PP_s(T, D)$.

4.2 Datasets

For our experiments we use two datasets, both comprised transcripts of telephone conversational speech. The first is the publicly available "SWITCHBOARD" dataset Godfrey et al. (1992) and the second a proprietary dataset of ImpacTech Ltd. TM approaches BERTclust, LDA, NMF selected 14, 14 and 12 topics for 6796 documents of the ImpacTech dataset. TM approaches BERTclust, LDA, NMF selected 33, 20 and 35 topics for the 121187 documents of SWITCHBOARD dataset.

For each dataset we generate a corpus of texts with the most likely topic per document calculated by each of the three TM approaches. In this way each document has three variants of topics. In order to compare the three TM approaches, either with objective, or subjective evaluation methods, we divide the comparison task into three units: BERTclust vs. LDA, BERTclust vs. NMF and LDA vs. NMF, similar to the methodology described in Fang et al. (2016a). Each comparison unit consists of 200 tuples of randomly selected texts with topic pairs as follows:

$$U_j = \{t_i, T_i^A, T_i^B\}, i \in 0...200, j \in 1, 2, 3 \quad (6)$$

where T_i^A, T_i^B are the topics assigned to text t_i from the methods A and B respectively.

4.3 Subjective TM method evaluation

Producing graded coherence assessment of topics can be a challenging task. Therefore, we apply a pairwise preference user study to gather human judgment. We create the corpus of comparison units as described above and get multiple human annotations for each tuple. In our annotation experiments we use five annotators per tuple. Each tuple is presented to the annotators as shown in Figure 4 and the annotators are asked to select the topic that better matches the text, or any other option between both or none.

To validate the quality of the human labeled data, and ensure that we do not have significant internal inconsistencies we use the Fleiss' kappa measure Fleiss (1971). The Fleiss' kappa measure is a statistical measure for quantifying the degree of agreement between categorical ratings given by a fixed number of annotators. In addition, we use the Wilcoxon rank test Wilcoxon (1945), a non-parametric test, commonly used to compare the rank of the mean values of different data sets. The null hypothesis for the Wilcoxon test states that the median values of two datasets do not differ. The alternative hypothesis for the one-sided test states $M_1 > M_2$, where M_i is the median of group i. If the result of the test is greater than the significance level (0.05 is frequently used), then the samples are indistinguishable. If, however, the result is less than the chosen level, then they are different.

For each annotation task, *i.e* a TM method pair, we find all data than can be removed as noise using the following two conditions. First, Fleiss' kappa F_k' must be more than 0.41, which according to Landis and Koch (1977) indicates a moderate inter-annotator agreement. Secondly, the p-value of Wilcoxon test p_W for samples of human

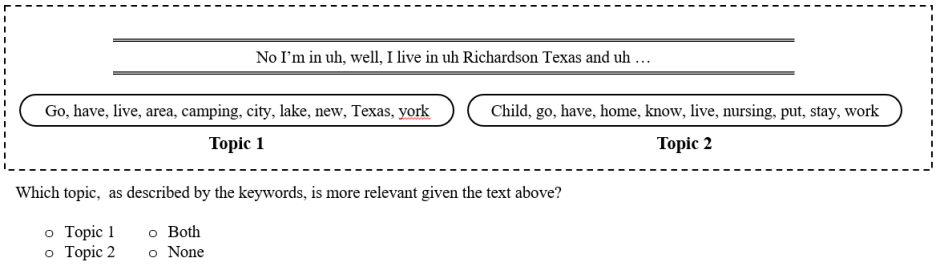

Figure 4: A single annotation task, as presented to the annotators. Each text, Topic 1, Topic 2 tuple was presented in five annotators.

		SWITCHBOARD		ImpacTech	
		F'_k	p_W	F'_k	p_W
BERTclust vs. LDA	before	0.27	$2.9\ 10^{-31}$	0.33	$1.6\ 10^{-16}$
	after	1.0	$1.2\ 10^{-7}$	0.55	$1.6\ 10^{-5}$
BERTclust vs. NMF	before	0.1	$1.6\ 10^{-7}$	0.37	$6.1\ 10^{-1}$
	after	1.0	$1.1\ 10^{-7}$	1.0	$5.3\ 10^{-7}$
LDA vs. NMF	before	0.24	$2.8\ 10^{-37}$	0.22	$1.5\ 10^{-12}$
	after	0.39	$2.5\ 10^{-6}$	0.45	$1.2\ 10^{-2}$

Table 1: The Fleiss' kappa F'_k and Wilcoxon test p-value p_w before and after noisy data removal.

scores in comparison to TM approaches must be less than 0.05, which indicates that there are no significant differences. The detected outliers are removed from the annotation sets, which as shown in Table 1, improves the quality of the subjective evaluations for all comparison units.

4.4 TM method ranking order investigation

First, we are interested to evaluate how different TM objective evaluation metrics behave in terms of ranking TM methods, and compare the ranking order according to objective scores with the ranking order according to the subjective evaluation. Starting with the comparison units described in Equation 6, we evaluate each tuple with various objective scores, and the best evaluated method across all tuples in each comparison unit is found. The ranking order is determined by the number of comparison units that each method is best evaluated for. In a very similar way, we evaluate each comparison unit according to human judgement, and create the subjective ranking order of the three TM methods as well.

To compare the different ranking orders we use the Wilcoxon rank test. The results of these experiments are presented in Table 2 for the two datasets that we use. We present the rank of each method, according to human evaluation, and according to the various subjective measure. First, we observe that, for both datasets, BERTclust is the best evaluated TM method according to subjective evaluation. However, the ranking order from our proposed $PP_s(T, D)$ metric matches exactly with the human ground-truth ranking order across our two datasets. A similar behaviour is observed for the Silhouette score.

4.5 Comparison of the objective and subjective evaluation

The distributions of human judgments and metrics are compared for each unit. Samples contain the -1/0/1 values, where "1"/"-1" represents that the topic from T1/T2 is preferred and "0" means no preference. We use the sign test to determine whether the automatic metrics perform differently than human judgments. The sign test is a statistical method to test for consistent differences between pairs of observations. Given pairs of observations for each subject, the sign test determines if one member of the pair tends to be greater than the other member of the pair. In our experiment we can interpret the null hypothesis rejection as a proof that there are differences between an objective and subjective evaluation of the same comparison unit.

We hypothesise that there are no differences between the preference data points from an objective metric and from human annotations for a comparison unit (null hypothesis), and thus we calculate the p-values reported in Table 3. Each metric gets 6 tests (3 tests from the SWITCHBOARD

SWITCHBOARD								
	Human	Rank	S_c	Rank	$PP(T,D)$	Rank	$PP_s(T,D)$	Rank
BERTclust	0.98	1^{st}	0.5	1^{st}	0.65	1^{st}	0.33	1^{st}
LDA	0.42	2^{nd}	0.49	2^{nd}	0.64	2^{nd}	0.31	2^{nd}
NMF	0.2	3^{rd}	0.48	3^{rd}	0.64	3^{rd}	0.3	3^{rd}
	C_v	Rank	C_{uci}	Rank	C_{npmi}	Rank	C_{umass}	Rank
BERTclust	0.63	1^{st}	0.15	1^{st}	0.86	1^{st}	-3.12	$2^{nd}/3^{rd}$
LDA	0.44	$2^{nd}/3^{rd}$	0.03	2^{nd}	0.16	$2^{nd}/3^{rd}$	-2.92	$2^{nd}/3^{rd}$
NMF	0.43	$2^{nd}/3^{rd}$	0.03	3^{rd}	0.15	$2^{nd}/3^{rd}$	-2.59	1^{st}

ImpacTech								
	Human	Rank	S_c	Rank	$PP(T,D)$	Rank	$PP_s(T,D)$	Rank
BERTclust	0.74	1^{st}	0.53	1^{st}	0.65	$2^{nd}/3^{rd}$	0.34	1^{st}
LDA	0.09	3^{rd}	0.47	3^{rd}	0.66	$2^{nd}/3^{rd}$	0.3	3^{rd}
NMF	0.21	2^{nd}	0.49	2^{nd}	0.66	1^{st}	0.32	2^{nd}
	C_v	Rank	C_{uci}	Rank	C_{npmi}	Rank	C_{umass}	Rank
BERTclust	0.68	1^{st}	0.2	1^{st}	0.41	1^{st}	-2.39	1^{st}
LDA	0.49	$2^{nd}/3^{rd}$	0.02	$2^{nd}/3^{rd}$	-2.24	$2^{nd}/3^{rd}$	-5.21	$2^{nd}/3^{rd}$
NMF	0.43	$2^{nd}/3^{rd}$	0.05	$2^{nd}/3^{rd}$	-0.69	$2^{nd}/3^{rd}$	-3.47	$2^{nd}/3^{rd}$

Table 2: The objective and subjective evaluation results for the investigated TM methods. The same rank for two different TM methods means that no statistical significance was found in the ranking order with the Wilconxon test.

dataset and 3 tests from the ImpacTech dataset). If $p \leq 0.05$, the null hypothesis is rejected, which means that there are differences between the preferences of the same comparison unit between a given metric and humans. We can observe, that the metric $PP_s(T,D)$ resulted in high p-values, which means that there are no statistically significant differences between the data points coming from the distribution of this metric and human evaluations.

In summary, we find that the $PP_s(T,D)$ metric demonstrates the best alignment with human preferences. This metric is convenient in that it does not require changes to text and vectors, which allows it to be applied to any set of words and text documents.

5 Conclusions

In this work, we have presented a detailed comparison of objective and subjective evaluation for different TM methods in the area of dialog speech analysis. We have proposed a set of pairwise proximity metrics for TM evaluation and found that they better agree with human judgment of TM, in evaluating three different TM methods. By using crowd-sourcing to obtain user preferences of topical coherence of topics and texts, we determined how closely each metric aligns with the human judgment. We showed that our proposed metric $PP_s(T,D)$ provided the highest levels of agreement with the human assessments.

So far, we have limited the scope of our research in a single dialog turn, and aimed to better detect the topics disregarding the whole dialog structure. As a next step, we wish to extent our scope to the analyses of the whole dialog, introducing a way to inform each TM method about topics detected in previous dialog turns. In addition, we plan to investigate how the available evaluation metrics can be extended in correctly evaluating topic selection across different dialog turns.

Acknowledgements

This work is funded by the European Union – NextGenerationEU, under Project Protocol Number INNOVATE/0719/0057.

References

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2020. Automatic generation of topic labels. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

SWITCHBOARD							
	S_c	$PP(T,D)$	$PP_s(T,D)$	C_v	C_{uci}	C_{npmi}	C_{umass}
BERTclust vs. LDA	$2.3\ 10^{-3}$	0.02	0.06	1.0	1.0	1.0	$6.1\ 10^{-5}$
BERTclust vs. NMF	$1.0\ 10^{-5}$	1.49	0.25	0.5	1.0	0.25	$1.6\ 10^{-2}$
LDA vs. NMF	0.66	0.83	0.06	0.07	0.78	0.15	0.68

ImpacTech							
	S_c	$PP(T,D)$	$PP_s(T,D)$	C_v	C_{npmi}	C_{uci}	C_{umass}
BERTclust vs. LDA	0.03	$4.6\ 10^{-6}$	0.09	0.09	0.23	0.06	0.23
BERTclust vs. NMF	0.34	0.24	1.0	0.07	0.07	0.55	$3.9\ 10^{-2}$
LDA vs. NMF	1.0	0.51	1.0	0.62	0.9	0.71	0.32

Table 3: The sign-test p-values for the proposed proximity metrics and human judgment scores for the two datasets.

Argyris Argyrou, Stamatios Giannoulakis, and Nicolas Tsapatsoulis. 2018. Topic modelling on instagram hashtags: An alternative way to automatic image annotation? In *2018 13th international workshop on semantic and social media adaptation and personalization (SMAP)*, pages 61–67. IEEE.

Ayoub Bagheri, Mohamad Saraee, and Franciska De Jong. 2014. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40(5):621–636.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Guy Camilleri. 2002. Dialogue systems and planning. In *International Conference on Text, Speech and Dialogue*, pages 429–436. Springer.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016a. Topics in tweets: A user study of topic coherence metrics for twitter data. volume 9626, pages 492–504.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016b. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Piyush Ghasiya and Koji Okamura. 2021. Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach. *IEEE Access*, 9:36645–36656.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Ruslan Kalitvianski, Emmanuelle Dusserre, and Muntsa Padró. Promises and disappointments of semantic analysis of speech-to-text applied to call center conversations in an industrial setting. *Industry Track*, page 6.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Akira Murakami, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. 'what is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277.

Murimo Bethel Mutanga and Abdultaofeek Abayomi. 2020. Tweeting on covid-19 pandemic in south africa: Lda-based topic modelling approach.

African Journal of Science, Technology, Innovation and Development, pages 1–10.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.

Matthew Purver, Konrad P Körding, Thomas L Griffiths, and Joshua B Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 17–24.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, and Pericles A Mitkas. 2015. Visual event summarization on social media using topic modelling and graph-based ranking algorithms. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 203–210.

Tatiana Sherstinova, Olga Mitrofanova, Tatiana Skrebtsova, Ekaterina Zamiraylova, and Margarita Kirina. 2020. Topic modelling with nmf vs. expert topic annotation: The case study of russian fiction. In *Mexican International Conference on Artificial Intelligence*, pages 134–151. Springer.

Lincoln Stein. 2001. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7):493–503.

Zhou Tong and Haiyi Zhang. 2016. A text mining research based on lda topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*, pages 201–210.

Johnny Torres, Alberto Jimenez, Sixto García, Enrique Peláez, and Xavier Ochoa. 2017. Measuring contribution in collaborative writing: An adaptive nmf topic modelling approach. In *2017 Fourth International Conference on eDemocracy eGovernment (ICEDEG)*, pages 63–70.

Andrew P Valenti, Ravenna Thielstrom, Felix Gervits, Michael Gold, Derek Egolf, and Matthias Scheutz. A multi-level framework for understanding spoken dialogue using topic detection.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Joe Wandy, Yunfeng Zhu, Justin JJ van der Hooft, Rónán Daly, Michael P Barrett, and Simon Rogers. 2018. Ms2lda. org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics*, 34(2):317–318.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017. A hierarchical topic modelling approach for tweet clustering. In *International Conference on Social Informatics*, pages 378–390. Springer.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80.