Relationship between Quality and Perceptual-Difference Metrics for Evaluating Topic Models

Frankland, Matthew Hafiz, Farhan Hughes, George Malcolm Jędrzejczyk, Sabina
Moir, Ross Mukhtar, Ridwan Schmieg, Mark Sparagano, Nicolas Welsh, Cailean

November 2020

1 Abstract

There exists a limited amount of research regarding evaluating topic models through the use of statistical measures. Little research has been done on the relationship between these measures when used to evaluate a topic model. This paper explores the benefits and limitations to evaluating topic models through quality metrics and perceptual-difference metrics. Topic models were generated from three different corpora, then various quality and perceptual-difference metrics were calculated, the values of which were compared and analysed for their relationships. Through this research it was concluded that topic models with similar quality metrics may vary greatly perceptually as there is no correlation between quality metrics and perceptual-difference metrics.

2 Introduction

Little work has been done in the area of statistical relationships between metrics which evaluate topic models. In particular, it is not known whether a correlation exists between perceptual-difference metrics and quality metrics for a set of topic models which have been trained on the same corpus but used differing initial conditions. As such, it is difficult to justify using solely quality metrics as evidence of a good model. This justification would be eased by demonstrating that topic models with high scores on quality metrics tend to have little perceptual-difference.

2.1 Overview of the Paper

The background section of this paper consists of research carried out regarding the quality and perceptual-difference measures. Additionally, the correlation between the different quality metrics and the correlation between the quality and the perceptual-difference metrics is also explored.

The general experiment design section presents an explanation on the data sets used for the experiments and the reasoning behind their selection. Along with that a detailed description of the configuration file used for the experiments is also talked about. Furthermore, the measures that were used for the evaluation of the experiments are also mentioned.

The experiments section is comprised of the results in the format of charts, tables and textual description from carrying out the experiments. The results obtained are discussed and compared against the hypothesis.

Finally, the paper ends with a conclusion which recaps the justification for this research while also discussing the conclusion made from the results of the experiments. The report closes with a discussion of the limitations of the research along with future work to be carried out.

2.2 Hypothesis

The null hypothesis is given as:

> For any two models with similarly high quality metrics, perceptual-difference metrics between models will be small.

2.3 Description of the Experiment

Three different experiments were carried out with three corpora of contrasting sizes to investigate the correlation between the quality and perceptual-difference metrics. A topic model generator was used for the generation of multiple topics. Most of the parameters used to generate models were kept consistent throughout the experiments, with the only changes taking place in the seed index, number of iterations carried out and the number of topics to be generated, each of which varied depending on the corpus size.

After the generation of models, quality and perceptual-difference metrics were calculated for each. These included log-likelihood, coherence, pairwise-difference and topical alignment. The parameters for the metrics remained consistent between experiments.

Finally, the resulting values obtained from the execution of the previous two programs were displayed on the Visual Report and the data was analysed and compared against the hypothesis, with the help of scatter plots and word bubbles.

3 Background

3.1 Quality Metrics

When using Latent Dirichlet Allocation (LDA) [1] to produce topic models, evaluation must be performed to choose optimal models. Early evaluation techniques consisted of purely quality metrics, such as topic coherence proposed by Newman, David, et al. [2] with an optimised variation [3] implemented in Java natural language processing tool MALLET [4], seen in formula 1.

$$coherence = \sum_{i} \sum_{j<i} log \frac{D(W_j, W_i) + \beta}{D(W_i)} \quad (1)$$

Coherence measures whether words within a topic are semantically the same. $D(W_j, W_i)$ finds the number of documents that contain both words i and j. $D(W_i)$ corresponds to documents that have the word i contained within them. Beta is the smoothing factor [5].

Log-likelihood is another metric used within MALLET that helps show the 'goodness of fit' of a model against sample data. It is used to help find how valid a topic model is by finding which model produced the highest average log-likelihood value.

3.2 Relationships of Quality Metrics

Strong correlation between metrics is important to help find reliable models. Researchers have used multiple metrics to help determine which model or method is best fit for use.

An experiment conducted by Cohen et al. [6] checked to see that their new LDA implementation, Red-LDA, performed better than the vanilla LDA implementation. They used quality metrics, log-likelihood and coherence, to test this and demonstrated that high log-likelihood also corresponded to high coherence.

Zhao et al. [7] were using both perplexity and topic coherence to demonstrate that their model, had either similar or improved results to normal topic models. Their proposed topic model consisted of topics focusing on words after being informed by word embeddings. Their results showed that low perplexity correlated with high coherence, hence also demonstrating that their proposed model showed more diverse topics with more representative words, than topic models using simply LDA.

Delamaire et al. [8] wanted to determine whether the quality of LDA output could be estimated using text similarity metrics. To achieve this they compared six different metric groups each containing at least two different variations of the group, with a seventh group containing other metrics. The results showed high correlation between several metrics such as Hellinger and Soergel distances as well as correlations between Jaccard and cosine distances.

The experiments showcased that a strong correlation between quality metrics resulted in a model or a new LDA method being significant. The quality metric relationship also helped proved that the research experiment results either improved or had similar outcomes to existing examples.

3.3 Perceptual-Difference Metrics

Upon the discovery of the unexpected negative correlation between likelihood metrics and topic coherency according to human subjects, exploration of new evaluation metrics began [9], leading to development of perceptual-difference metrics which attempt to model the way in which a human would score the coherency of topics.

Topical alignment [10], [11] encompasses all the topics from all the models given to it. It clusters the most similar topics together, and does not allow topics from the same model to appear in the same cluster, as well as prohibiting one topic from appearing in more than one cluster. To create such clusters, a distance matrix is produced by comparing the top most weighted words from each topic using distance equations such as, Manhattan distance or cosine distance. Agglomerative clustering is then performed using the distance matrix to extract clusters of most similar topics by always choosing the shortest distance. From topical alignment more metrics can be produced, which include topic size variability and cluster size variability.

Pairwise difference [12] takes a similar approach, but considers the distance between pairs of models; a distance matrix is created by normalising the weights of all topics, taking the top M words of each and finding the difference between each pair of topics with a distance function such as Manhattan or euclidean distance. An assignment algorithm, specifically the Hungarian algorithm [13], is used to find the set of points on this distance matrix which minimise the sum, with this sum being used as the distance between the two models.

3.4 Interactive Viewers

There exist numerous interactive viewers for visualising topic models in various ways, including word bubbles of topic clusters [14], trend maps of research areas [15] and evolution of the weighting of documents as the model is run for more iterations [16].

Topic Check [11] provides an interactive tool to explore the alignment of a set of topic models, with the goal of understanding how stable the models generated by a corpus are. This comparison of a set of topic models is limited to alignment-related metrics however, and while useful for gaining insight into the corpora themselves, leaves all other metrics out of scope.

Other than the tool developed for the research that our paper explores, there are currently no accessible resources which allow for the exploration of a set of topic models based on comparing their metrics; such a tool would provide a means to explore the relationships between metrics for models with different initial conditions.

3.5 Quality and Perceptual-Difference Metrics Relationship

It is currently unknown, as far as the research for this paper has been able to reach, as to whether or not there exists a relationship between the above defined quality and perceptual-difference metrics.

The existence of a negative correlation between the two would affect the confidence in topic modelling in that it would ascertain the usefulness of quality metrics as a determining factor in evaluating topic models, since models would appear to tend towards a global attractor as the quality metrics score higher.

On the other hand, if a positive or lack of correlation was observed, it would suggest that models with higher quality metric scores are likely to be perceptually different, and thus it becomes difficult justify the use of quality metrics as basis for choosing the ideal topic model. In this case, confidence in topic models is lost.

4 Results

4.1 Experimental Design

4.1.1 Data sets

Three data sets were chosen to be analysed in the following experiments in order to test results at a range of different scales. The corpora were chosen based on their size, ensuring a range of total documents and words, allowing results to be extrapolated to a range of corpora. The three data sets that were used were the BBC News corpus, a Kaggle corpus on coronavirus data, and a corpus from the machine learning Stack Exchange forum.

Each experiment used a given configuration file as a starting point. The following parameters, threads; optimisation interval; symmetric alpha; alpha-sum and beta, were fixed for each data set that was analysed.

Threads The number of threads used for each experiment was fixed at 20. This creates 20 parallel samplers, each looking at a portion of the data set before combing the resulting statistics. Mallet uses the number of threads it has available to optimize the alpha and beta value, thus results differ dependant on the number of threads assigned. The number of threads used during each experiment was fixed at a high number for consistency and fast results.

Symmetric Alpha Symmetric Alpha is not set in the configuration file. Therefore the value used is the system default, false. This results in a low alpha value placing more weight on having each document composed of only a few dominant topics during topic alignment.

Optimisation Interval Optimisation interval was initially set to 0. This stops Mallet turning on hyper parameter optimisation on every 10 iterations. Too many optimisation rounds can lead to instability, with alpha parameter going to 0. This was to be avoided on a large data set such as Stack Exchange.

Alpha-Sum The value of alpha-sum effects the initial weights given on the distribution of topics prior to any optimisation taking place. So that words did not start off with a different weight in each experiment, this value was fixed to 1.0.

Beta Mallet's beta value means that each topic has a weight/100 on the distribution prior to the size of the vocabulary being calculated. The beta value will be adjusted during optimisation by around a factor of two and therefore to keep experiments consistent, this value was kept constant.

The following factors, number of topics; seed index; and iterations, were variable for each data set that was analysed.

Number of Topics The number of topics was initially fixed to 40 however this value could be changed depending on the data set used in order to account for the size of a given corpus.

Seed index Seed index was changed to try a variety of topic models. The seed index could be increased to have more topics to work with.

Iterations A minimum iteration value was given at 400 iterations to guarantee that Topic Models did not converge too early. This would cause the same topic model to be produced on each iteration. This value could be adjusted to explore where convergence would occur for a given data set.

4.1.2 Measures

Each experiment will investigate whether there is a significant correlation between the quality metrics of topic coherence and log-likelihood, and the perceptual-difference metrics of topical alignment and pairwise-difference.

Log-likelihood and topic coherence were selected as quality measures because topic models with a high log-likelihood and topic coherence, compared with other topic models, would be the cases that would give the best perceptual-difference metrics for analysis.

4.1.3 Evaluation

Given a strong correlation in quality metrics, selected topics perceptual-difference metrics can be evaluated to determine

whether the experiments null hypothesis can be rejected or not. If aligned topics are shown to have a strong correlation in their quality metrics, the null hypothesis can be rejected. If pairwise-difference results are wide ranging even though a group of topic models have strongly correlated quality metrics, this would also reject the null hypothesis.

4.2 Experiments

4.2.1 BBC Corpus

Introduction This experiment used the smallest corpus, which contained 2,225 news articles from BBC News between 2004-2005[17]. The articles are drawn from five topic areas. As such, the number of topics in a model was chosen to be five. A range of 1-100 seed indexes were used – as opposed to the 20 specified in the experiment methodology – since the smaller corpus and number of topics allowed for more models to be processed with the same compute, and more confidence could be gained for larger samples. Each model was trained for 750 iterations, once again due to available compute, as well as to explore the idea that topic models might converge perceptually as they stabilise.

Results Three models which gave high scores for log-likelihood and coherence relative to other topic models were selected for inspection. The choice of models can be seen in Figure 1, with the choices being coloured as blue, pink, and green. Given that the model with the highest values for both coherence and log-likelihood would find itself in the top-right corner of figure 1, this specific selection included three models found within the top-right quadrant of the graph chosen for their high perceptual-differences from each other.

Since pairwise-difference is measured between pairs of models, a single model can be chosen as a base and its distance from each other model plotted on an axis. This was performed for each of the selected models, with an example shown in Figure 2 in which the spread of the models' distance from blue is visualised. While pink is seen to be of below-average distance, green is the fourth most distant model from blue.

The alignment of topics in the selected models was also explored, with word bubbles of aligned topics shown in 3 and 4. The alignment between the models is demonstrated in table 1, for which a distance threshold of 0.05 was used.

Model pair	Total Matches	Total Non-Matches	Match Rate
Blue, Pink	4	1	0.8
Blue, Green	2	3	0.4
Pink, Green	2	3	0.4

Table 1: Topical Alignment between the three selected models

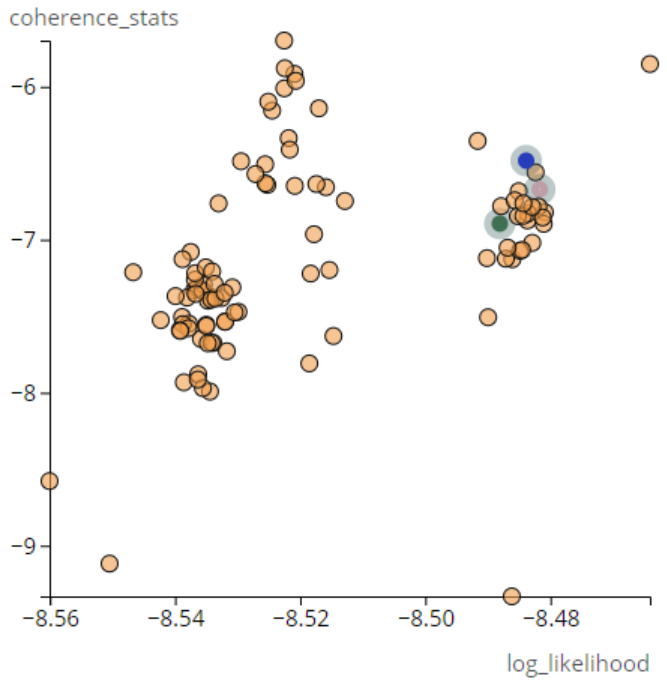

Figure 1: Scatter plot showing blue, pink and green models with high coherence and log-likelihood.

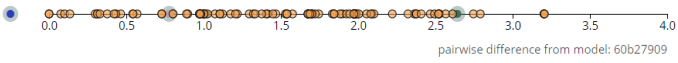

Figure 2: Pairwise-difference from blue model to pink and green models.

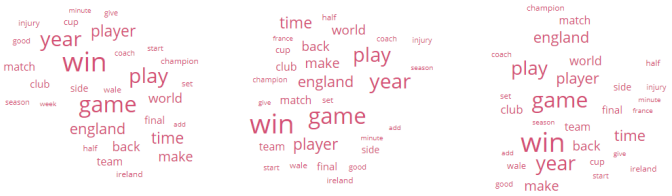

Figure 3: Example of alignment of topics in blue, pink, and green models.

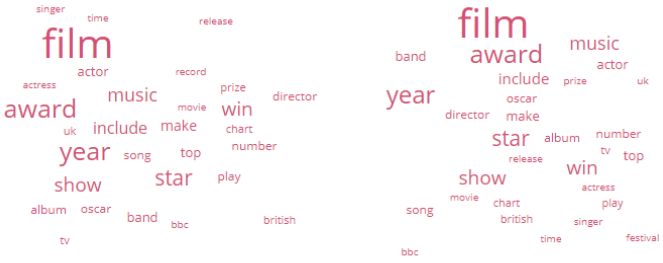

Figure 4: Example of alignment of topics in blue and green models.

The Pearson correlation coefficient, labeled as r, was used to mathematically determine the correlation between the models' attributes [18]. Additionally, the statistical p-value, was used to determine the mathematical significance of the correlation r. The degrees of freedom used in calculating p was the number of models at 100. The metrics explored for correlation included Log-likelihood, Coherence, and mean of Pairwise-Differences.

As seen in table 2, the mean of the models' pairwise-difference is positively correlated with the log-likelihood at an r of 0.409. This correlation is significant at a p-value of 0.000024 for $p < 0.05$. The other quality metric, coherence, has a low correlation with the mean of the pairwise-difference with an r of 0.197, which is statistically significant with a p-value of 0.049 for $p < 0.05$.

Metrics		Correlation	
Perceptual	Quality	r	p-value
Pairwise ($\bar{x}$)	Log-likelihood	0.409	0.000024
Pairwise ($\bar{x}$)	Coherence	0.197	0.049

Table 2: Correlation between Quality and Perceptual-Difference Metrics on the BBC corpus.

Discussion Considering pairwise difference, the blue and pink models are relatively perceptually similar at 0.78 units, below the average distance for the blue and pink models which were 1.5 and 1.4 respectively. This perceptual closeness, alongside their similar log-likelihood and coherence, supports the null hypothesis. The blue and green models demonstrate a counterexample: they have similar log-likelihood and coherence scores, however, there is an above average perceptual distance of 2.6 units between them with the average pairwise distance for the green model being 2.2. This single example provides evidence to reject the null hypothesis for the case of this corpus, though frequency of this behaviour was not explored here.

The alignment reflects the information given by the pairwise difference; while blue and pink are almost perfectly aligned with 4/5 topics matched, blue and green have only 2/5 topics aligned. This provides more evidence for the case of rejecting the null hypothesis with two perceptual-difference metrics in agreement. This is also supported by the fact that pink and green have an above average pairwise difference as well as a low alignment of 2/5.

The results from the correlation table indicate that as the quality metrics of log-likelihood and coherence increase, so too does the average perceptual difference between the topic models. There appears to be a divergence rather than the supposed convergence of topics. Since the results from this are statistically significant, the null hypothesis is rejected.

It has been ascertained for this corpus that picking a model with good scoring quality metrics does not account for the existence of another model with very similar quality metric scores which is perceptually very different, i.e. its topics do not align well and its words are weighted differently. Confidence in topic models selected by quality metrics is lost in this case.

4.2.2 Kaggle Covid 19 Corpus

Introduction This experiment used the 'kaggle_covid19_200617' corpus which was composed of the titles and abstracts from papers in english about coronaviruses, extracted on June 17th 2020. The seed ranged from 1-100 to gain a larger sample size for the experiment. A value of 5 was chosen for the number of topics due to Figure 5 highlighting the coherence getting worse as more topics were added.

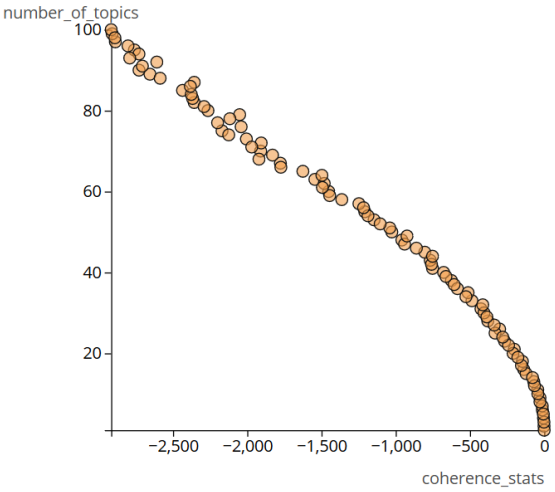

Figure 5: Kaggle topic number experiment

Results The models that produced the highest quality measures (log-likelihood and coherence) were selected for comparison. Figure 6 showcases the models selected for this experiment. All three models scored highly in coherence and log-likelihood and can be seen in the top right quadrant of the graph.

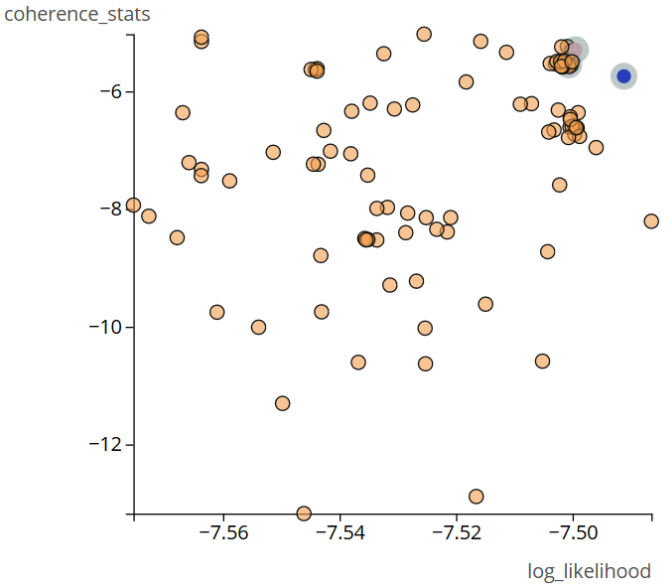

Figure 6: Kaggle selection of models showing the relation between log-likelihood and coherence

The figures below showcase the pairwise difference between the selected models. Figure 7 showcases that perceptually, the models are different.

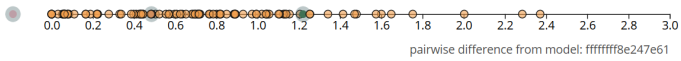

Figure 7: Kaggle using pink model as base model for pairwise comparison.

Topical alignment was also performed on the dataset and a significance value of 0.05 was used. Table 3 showcases the alignment between the three selected models and how many matches and non matches were made.

Model pair	Total Matches	Total Non-Matches	Match Rate
Blue, Pink	3	2	0.6
Blue, Green	3	2	0.6
Pink, Green	5	0	1.0

Table 3: Kaggle Topical Alignment between the three selected models

Figure 8: Kaggle showing three aligned word bubbles.

Figure 9: Kaggle showing two aligned word bubbles.

Table 4 shows a negative correlation for pairwise with coherence and log-liklihood. Both are significant as the p values are < 0.00001 for $P < 0.05$.

Metrics		Correlation	
Perceptual	Quality	r	p-value
Pairwise ($\bar{x}$)	Log-likelihood	-0.564	$<.00001$
Pairwise ($\bar{x}$)	Coherence	-0.548	$<.00001$

Table 4: Correlation between Quality and Perceptual metrics on the Kaggle corpus.

Discussion All three models were similar in terms in quality metrics however, differed perceptually. Figure 7 showcases this as the pink and green models were at a distance of 1.2 units, compared to the blue model being at a distance of 0.45 units.

In topical alignment, where a similarity threshold of 0.05 was used, the pink and the green model perfectly matched with 5/5 matches. This result matches the one seen in Figure 6, where the green and pink model have very similar log-likelihood and coherence measures. However, the blue model only matched at a rate of 0.6 to the pink and the green model. Since all the models are located within the top right quadrant of the graph and are very close in terms of quality measures, we would expect the topical alignment results to have a full match rate for all the chosen models, this is however not the case.

The correlation table highlights that as the quality metrics of log-likelihood and coherence go down, so does the perceptual difference. The results are significant as seen in Table 4.

Due to similar high quality measures being different perceptually and the significance of our results, the null hypothesis can confidently be rejected.

4.3 Stack Exchange Corpus

Introduction This final experiment used a corpus comprised of a subset of posts and replies made on the Machine Learning Stack Exchange forum, extracted on the 1st of June 2020. Having over 300,000 documents, this corpus was a good test to ensure results obtained from the previous two experiments are consistent at scale. Due to the large number of documents and words, model generation and metric calculation took multiple hours and so variation in parameters was costly in respect to time. For this reason,

a small number of models were produced using the base parameters described in 4.1 chapter 4.1, but a more thorough investigation has been carried out on these models.

Results Firstly 3 topics were chosen that were equidistant from each other in respects to mean coherence and log likelihood, pictured below in Figure 10. This model was chosen as the base because the other two models have one metric which is almost identical to the base.

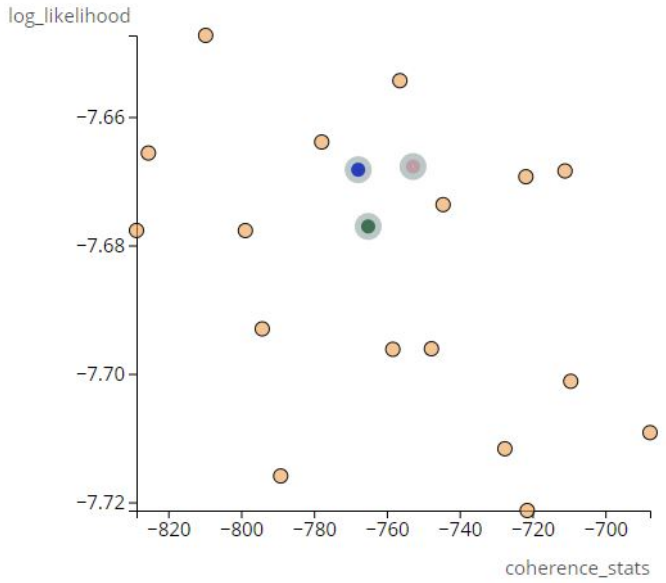

Figure 10: Stack Exchange Equidistant Models

These are as equidistant as possible whilst attempting to find models with similar and high quality metrics given the smaller number of models generated. The pairwise difference between the base model, shown as blue in 10, and model 2, pink, and model 3, green, is 20.34 and 17.01 respectively. To put this in context, the pairwise difference between model 1 and 2 is relatively poor, coming in the worst performing quartile, whilst between model 1 and 3, the pairwise difference comes in 4th best performing, and subsequently the top quartile. The number of aligned topics between the base and the compared models is similar, with only 13 and 15 matches between them, out of 40 total topics. 8 of the topics had matches across all 3 models, whilst the remaining were independently matched with the base model.

The second set of results compare 2 similarly high models in regards to log likelihood and coherence, and one distinctly different and poorer performing model.

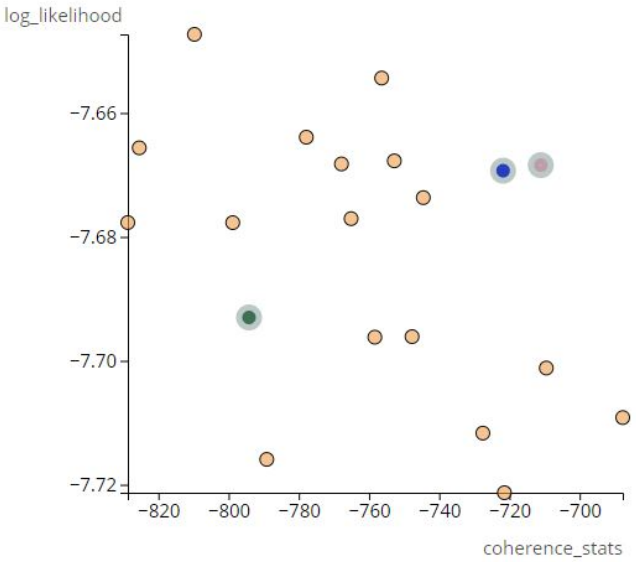

Figure 11: Stack Exchange 2 Similar 1 Diff

In this instance we find that, in comparison to the base model, the poorer performing topic model is the model with the lowest pairwise difference from the base model, whilst the very similar model is in the bottom 50% of models. This is pictured below in Figure 12

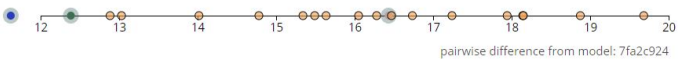

Figure 12: Stack Exchange Pairwise Difference

The comparison models both match with an equal number of topics, however specific examples of heavily weighted words in word clouds show relevant results. In Figure 13 we can see that 'Canada' appears as a primary word in the topic, however this is not reflected in the paired topic from model 3.

Figure 13: Stack Exchange Pairwise Difference

The Pearson's r values and p-values are described in Table 5.

Metrics		Correlation	
Perceptual	Quality	r	p-value
Pairwise ($\bar{x}$)	Log-likelihood	-0.146	<.0.539086
Pairwise ($\bar{x}$)	Coherence	0.311	<.154008

Table 5: Correlation between Quality and Perceptual metrics on the Stack Exchange corpus.

Discussion The first set of results give an indication that models can have similar quality metrics but different perceptual-difference metrics. Whilst the first results are "equidistant", it could indicate that similar coherence values have a greater impact on perceptual similarities in comparison to log-likelihood. However, when testing this and finding other examples of models which were equidistant in similar ways to 10, these results were not consistent. Instead it was found that a more similar log-likelihood resulted in a lower pairwise difference. There was not found to be any consistency regarding this, which suggests that a singular similar quality metric does not impact perceptual similarity within these results.

The second set of results are significant, as they demonstrate the opposite of what is posed as the hypothesis. If evidence were to support our hypothesis, we would expect the models with similar quality metrics to also be perceptually similar, and the poorer performing models to be perceptually different, when in fact the opposite is true. Lower pairwise difference with the base model did not appear to correlate to a shorter distance in respect to quality metrics between respective models. We can draw two relevant conclusion from this result. Firstly, that relying purely on pairwise difference could result in poor scoring topic models in regards to quality metrics. Secondly, choosing two models based on quality metrics may result in two perceptually different models being selected. This is a rejection of the hypothesis.

The topical alignment results would further strengthen this case, particularly when examining the example used of the word "Canada". In the base model, "Canada" is the top word in a topic, yet this doesn't appear in the matched topic with model 3 and appears in an unmatched topic in model 2. This demonstrates that words considered to define a topic in one model may not even appear in another, or do not align at all. With model 1 and 2 in the second set of results, they score similarly, but here "Canada" appears in an unmatched topic. This is similarly inconsistent with the hypothesis.

The p-values, shown in Table 5, between pairwise difference and the quality metrics used are very high, meaning the results are not significant and no statements can be made regarding correlations or otherwise between the quality and perceptual-difference metrics. This is consistent with the results found during this experiment and prohibits any conclusions to be made generally regarding all models and the relationship between quality and perceptual-difference metrics.

5 Conclusions

5.1 Project Summary

At present, there exists a multitude of mathematical metrics which aim to evaluate the quality of and perceptual distance between topic models. However, a lack of research into the statistical relationship between quality and perceptual-difference metrics had previously been conducted. In this paper, we proposed a null hypothesis in Section 2.2, and conducted three unique experiments, utilising differing corpora to generate a range of topic models with varying initial conditions, designed to reject the defined null hypothesis. The investigations conclude with, through contradiction in each case, a rejection of the null hypothesis.

The results of these experiments were inconclusive regarding correlation between quality metrics and perceptual-difference metrics for any set of topic models trained on a singular corpus with differing initial conditions.

5.2 Limitations and Future Work

Whilst this paper successfully rejects the null hypothesis described in Section 2.2 through the experiments performed and the results obtained in Section 4, there are some limitations to the work.

One such limitation is that the experiments performed only compared two quality metrics, log-likelihood and coherence, against the perceptual-difference metrics. Other quality metrics such as perplexity were not used in these experiments. Thus, any future work would include looking at using more of the quality metrics in the experiments and analysing a wider variety of models in order to reinforce or reject the results obtained in this paper.

Another limitation of the work described in this paper is the lack of a developed set of topical alignment evaluation measures. Currently the only way that topical alignment is used in the evaluation is by considering the percentage of matched topics for each of the selected models. This does not describe, in and of itself, the quality of the matches, and can be adjusted by changing the distance threshold used when calculating topical alignment. Future work for this would involve developing a set of evaluation measures extracted from topical alignment comparisons, thereby better evaluating the perceptual-differences.

Currently there are only two perceptual-difference metrics that exist, pairwise difference and topical alignment, therefore any future work would involve developing further perceptual-difference metrics and rerunning the experiments using these metrics.

Finally, whilst this paper has made some reference to the correlations that exist between quality metrics and

perceptual-difference metrics, an investigation that primarily focuses on these relationships may be useful. Such an investigation could include considering the relationship between a single type of quality metric with perceptual-difference metrics, before expanding the scope to encompass comparisons using all types of quality metrics.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108.

[3] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.

[4] A. K. McCallum, "Mallet: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.

[5] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 952–961.

[6] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, "Redundancy-aware topic modeling for patient record notes," *PloS one*, vol. 9, e87555, Feb. 2014. DOI: 10.1371/journal.pone.0087555.

[7] H. Zhao, L. Du, and W. Buntine, "A word embeddings informed focused topic model," in *Proceedings of the Ninth Asian Conference on Machine Learning*, M.-L. Zhang and Y.-K. Noh, Eds., ser. Proceedings of Machine Learning Research, vol. 77, PMLR, Nov. 2017, pp. 423–438. [Online]. Available: http://proceedings.mlr.press/v77/zhao17a.html.

[8] A. Delamaire, M. Juganaru-Mathieu, and M. Beigbeder, "Correlation between textual similarity and quality of lda topic model results," eng, in *2019 13th International Conference on Research Challenges in Information Science (RCIS)*, IEEE, 2019, pp. 1–6, ISBN: 1728148448.

[9] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Advances in neural information processing systems*, vol. 22, pp. 288–296, 2009.

[10] J. Chuang, S. Gupta, C. Manning, and J. Heer, "Topic model diagnostics: Assessing domain relevance via topical alignment," in *International conference on machine learning*, 2013, pp. 612–620.

[11] J. Chuang, M. E. Roberts, B. M. Stewart, R. Weiss, D. Tingley, J. Grimmer, and J. Heer, "Topiccheck: Interactive alignment for assessing topic model stability," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 175–184.

[12] A. Gharavi, M. J. Chantler, S. PADILLA, and T. S. Methven, *The iillusion of certainty: Ruining and restoring confidence in stochastic algorithms.*

[13] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. DOI: https://doi.org/10.1002/nav.3800020109. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

[14] P. L. Bras, A. Gharavi, D. A. Robb, A. F. Vidal, S. Padilla, and M. J. Chantler, *Visualising covid-19 research*, 2020. arXiv: 2005.06380 [cs.IR].

[15] S. Padilla, T. S. Methven, D. W. Corne, and M. J. Chantler, "Hot topics in chi: Trend maps for visualising research," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '14, Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 815–824, ISBN: 9781450324748. DOI: 10.1145/2559206.2578867. [Online]. Available: https://doi.org/10.1145/2559206.2578867.

[16] Y. Yang, Q. Yao, and H. Qu, "Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling," *Visual Informatics*, vol. 1, no. 1, pp. 40–47, 2017, ISSN: 2468-502X. DOI: https://doi.org/10.1016/j.visinf.2017.01.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2468502X17300074.

[17] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 377–384, ISBN: 1595933832. DOI: 10.1145/1143844.1143892. [Online]. Available: https://doi.org/10.1145/1143844.1143892.

[18] J. Cohen, *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, 1988.