Fall 10-1-2020

Assessing Topical Homogeneity with Word Embedding and Distance Matrices

Jeffrey M. Stanton
Syracuse University

Yisi Sang
Syracuse University

Recommended Citation

Stanton, Jeffrey M. and Sang, Yisi, "Assessing Topical Homogeneity with Word Embedding and Distance Matrices" (2020). *School of Information Studies - Faculty Scholarship*. 193.
https://surface.syr.edu/istpub/193

Assessing Topical Homogeneity with Word Embedding and Distance Matrices

Description/Abstract

Researchers from many fields have used statistical tools to make sense of large bodies of text. Many tools support quantitative analysis of documents within a corpus, but relatively few studies have examined statistical characteristics of whole corpora. Statistical summaries of whole corpora and comparisons between corpora have potential application in the analysis of topically organized applications such social media platforms. In this study, we created matrix representations of several corpora and examined several statistical tests to make comparisons between pairs of corpora with respect to the topical homogeneity of documents within each corpus. Results of three experiments suggested that a matrix of cosine distances calculated from vector summaries of short phrases contains useful information about how closely the documents within a corpus relate to one another. Both the tested summarization method and a non-parametric test for comparing cosine distance matrices appear to have utility for examining and comparing corpora containing brief texts.

Keywords

data analysis, text analysis, word embedding

Disciplines

Applied Linguistics | Databases and Information Systems | Library and Information Science | Management Sciences and Quantitative Methods

Creative Commons License

Assessing Topical Homogeneity with Word Embedding and Cosine Distance Matrices

Jeffrey M. Stanton
Hinds Hall, Syracuse University, Syracuse, NY, 13244, USA
jmstanto@syr.edu
315-443-2879

Yisi Sang
Hinds Hall, Syracuse University, Syracuse, NY, 13244, USA
yisang@syr.edu

Abstract

Researchers and practitioners use statistical tools to analyze large collections of text. Many statistical tools support quantitative analysis of documents within a corpus, while relatively few consider the statistical characteristics of whole corpora or comparisons between corpora. Statistical summaries of whole corpora and comparisons between them have possible applications in the analysis of topically organized applications such threaded discussions on social media. In this study, we created distance matrices to represent twenty-four social media corpora and examined several statistical tests to compare pairs of corpora with respect to the topical homogeneity of documents within each corpus. Results from three studies suggested that a matrix of cosine distances calculated from vector summaries of short phrases contains useful information about how closely the documents within a corpus relate to one another. Both the tested summarization method and a non-parametric test for comparing cosine distance matrices appear to have utility for characterizing and comparing corpora containing brief messages.

Keywords:
social media; text mining; statistical analysis; word embedding; Monte Carlo simulation

Assessing Topical Homogeneity with Word Embedding and Cosine Distance Matrices

1. Introduction

The development of algorithmic methods to analyze natural language text has accelerated over recent decades (Dumais, 1994; Hofmann, 1999; Landauer et al., 1998; Papadimitriou et al., 2000). Techniques such Latent Dirichlet Analysis (Blei et al., 2003; Blei & Lafferty, 2007; Blei et al., 2010; Blei, 2012) and Latent Semantic Analysis (Evangelopoulos, et al., 2012) have provided methods for analyzing a corpus of textual data to reveal statistical regularities in word meaning and document content. Researchers have also developed other innovations such as structural topic modeling (Roberts et al., 2014), supervised topic modeling (McAuliffe & Blei, 2007), joint latent topic modeling (Nallapati, et al., 2008), topic model visualization (Sievert & Shirley, 2014), word embedding (Mikolov et al., 2013), sentence embedding (Reimers & Gureych, 2019), and other text analysis methods (e.g., Devlin et al., 2018; Peters et al., 2018). Many of these techniques share the goal of modeling regularities at the level of phrases, sentences, and/or paragraphs within a body of text.

In contrast to those methods, this paper focuses on a complete corpus as the unit of analysis and explores statistical methods for describing a corpus and making comparisons between pairs of corpora. One application of this capability lies in assessing topical homogeneity among documents in a corpus. Any application area that manages multiple corpora organized by topic – for example, a social media platform – could benefit from the capability of assessing topical homogeneity. By creating a statistical summary of a threaded discussion, one could illuminate aspects of user behavior, such as the formation of linguistic communities (e.g., Nguyen et al., 2017). For example, one might hypothesize that a corpus extracted from a threaded social media conversation about meal recipes would contain messages with linguistic commonalities about flavors, spices, and ingredients. In contrast, a different set of postings about food safety might contain a wider degree of linguistic variation over divergent topics such as recalls, food biology, hygiene techniques, shipping practices, and government regulations. Statistical analysis of topical homogeneity could document differences between these corpora indicative of the respective user communities that contributed to them.

As this example suggests, for this paper we define topical homogeneity as the extent to which documents within a corpus are semantically or linguistically close to one another. This idea is distinctive from topical coherence, a term that often refers to metrics assessing interpretability of a topic generated by a topic model (Röder et al., 2015). Given the focus on the relative topical homogeneity of corpora, we examined three research questions:

RQ1: What is a suitable method for creating numeric representations of topical homogeneity for a corpus of brief textual documents such as social media postings?

RQ2: Do numeric representations of topical homogeneity for commonly available corpora (e.g., threads from social media platforms) fit any theoretical statistical distributions?

RQ3: With knowledge of candidate distributions of topical homogeneity data, what test works best to detect differences in topical homogeneity between pairs of corpora?

Providing researchers with a statistical test to compare topical homogeneity across corpora could open exploration of new research questions. Given appropriate methods, researchers could also use topical homogeneity to compare a social media thread to itself at different points in time, to identify threads that contain outlier documents, and to examine whether communities of posters tend to keep their posts "on topic."

In this article, we evaluate a method for analyzing topical homogeneity in corpora of brief texts such as those that would be found in threaded social media conversations. After extraction of textual material from social media sources, this approach begins with word embedding to summarize the linguistic content in each posting. Next, we transform word vector representations into distance matrices that capture the similarity of each document to the other documents in the corpus. Finally, we examine comparative statistical tests for these distributions of distances. We report a Monte Carlo analysis that evaluated these statistical tests. The manuscript is organized as follows. In Section 2 we review the use of word embeddings for brief document summarization, similarity/distance measures for vector representations, and candidate distributions for distance values. In Section 3, we propose a method for measuring topical homogeneity and describe our methods. The results appear in Section 4 and we synthesize and interpret the results in Section 5.

2. Background

Over recent years statistical techniques for text analysis, such as topic modeling, have emerged as practical and important tools in social science, business, education, and many other fields (e.g., Lin & Wang, 2020). For example, in clinical psychology, researchers assessed whether depressed patients expressed common linguistic patterns that differed from non-depressed patients (Resnik et al., 2015). Green and Cross (2017) explored the evolution of political agendas of the European Parliament across multiple years. Shi et al. (2016) demonstrated a method of assessing “business proximity” – i.e., the extent to which two businesses perform similar functions in the marketplace. Statistical analysis in that study examined distances among documents using vector representations. These distances then became the basis for deciding which businesses were close and which were far from one another with respect to their marketplace functions. Relatedly, Lee et al. (2015) used statistical representations of topics to assess connections between corpora of teaching and research materials produced by faculty from 36 universities. Results showed similarity between teaching and research materials for introductory courses but not for advanced courses.

Results from both Lee et al. (2015) and Shi et al. (2016) presage new analytical possibilities once two or more naturally occurring sets of documents have statistical representations that enable comparisons between them. Massive amounts of online textual data are organized into topical structures – groups, threads, hashtags, etc. – that make such comparisons viable and potentially interesting. Given a suitable measure of distance between any pair of documents, one could represent a set of documents using a distance matrix. Under the assumption that selected documents comprise a subset of a larger ostensive set of documents that could appear within a topical thread, we might construe the observed values in the distance matrix as a sample from a theoretical population of distances for that thread.

2.1 Word Embedding for Brief Document Summarization

To create distributions of distance values between brief text postings we must use a summarization method that supports calculation of a measure of similarity or dissimilarity between a pair of documents. For this study we have chosen word embedding as a summarization method, but in principle any method supporting pairwise measurements of document similarity would work. Word embedding is an umbrella term for a variety of methods that represent terms as high dimensional numeric vectors. A common goal of these methods is to establish a kind of numerically encoded thesaurus, where words with similar vector representations share similar meanings (Conia & Navigli, 2020; Nguyen et al., 2019). The approach originates with ideas from linguists Harris (1968, p. 16), Firth (1957) and others that the contiguity of successive words informs their relationships to each other, also known as the distributional hypothesis. Practical approaches to word embedding have emerged as computational power and available digital corpora have increased. In 2013, Mikolov et al. publicized “word2vec,” a neural-network approach to computing vector representations. To train a word2vec model, one uses a large, varied corpus of text such as the total contents of articles from Wikipedia. Yamada et al. (2018) used text from Wikipedia to pretrain word2vec models in several languages. Once a model is trained, each word has a representation as a high dimensional vector of weights. Dictionaries of words with their vector representations are available with as few as 50 dimensions or as many as 1000 dimensions (e.g., Yamado et al., 2018; Sahlgren & Karlgren, 2005). Research suggests that more dimensions is not always preferable for every application (Das et al., 2019).

As a beneficial side effect of calculating vector representations in multidimensional space, one may compute a kind of semantic arithmetic. A commonly offered example indicates that "brother" – "man" + "woman" = "sister." Mikolov et al. (2013) conducted experiments with word embedding showing that this kind of semantic arithmetic provides usable results, particularly for short strings of words, such as those that might occur in a social media posting. Compositionality, a term borrowed from semantics, proposes that the meaning of a short phrase is closely related to the phrase's component words (Mikolov et al.; 2013). Experiments have supported compositionality (e.g., Seyeditabari & Zadrozny, 2017) and research on the mathematics of word vectors (Ethayarajh, Duvenaud, & Hirst, 2019) has suggested that improvements to the training processes used to create word and sentence embeddings will continue to enhance summarization of short phrases with numeric vectors. In particular, developments are underway for creating vector representations of complete sentences (e.g., Reimers & Gurevych, 2019) and for creating vector representations of word senses in addition to those for individual terms (Colla et al., 2020).

2.2 Similarity/Distance Measures for Vector Representations

One benefit of representing words and short phrases as vectors lies in the capability of measuring the proximity of a pair of vectors. A similarity/distance measure reflects the closeness or separation of two vectors by mapping the distance/similarity between the vectors into a single numeric value (Huang, 2008). The mapping depends both on the properties of the vectors and the measure itself (Huang, 2008). Euclidean distance, a commonly applied and well-known measure, calculates the length of a straight line between two points. Cosine similarity, a calculation of the angle in multidimensional space, has found common usage in text processing applications (Anderlucci et al., 2019). Researchers have also applied the Jaccard distance, the Pearson product-moment correlation, the Dice distance (Dice, 1945; Ali and Mahmood, 2020), and the Kullback-Leibler divergence to assess distance and/or similarity between vectors.

Euclidean distances have many applications in diverse fields such as data visualization, psychometrics, crystallography, machine learning, and signal processing (Dokmanic et al., 2015). Euclidean is the distance measure used in the ubiquitous K-means algorithm (Liao et al., 2013). When measuring distance between two text documents C_a, C_b represented by their word vectors $\overrightarrow{t_a}$, $\overrightarrow{t_b}$, the Euclidean distance of the two corpora can be presented as:

$$D_E(\overrightarrow{t_a},\ \overrightarrow{t_b}) = (\sqrt[2]{\sum_{t=1}^{n} |w_{t,a} - w_{t,b}|^2})$$

The word set is $T = \{t_{1,}\ t_{2,} \ldots, t_{n,}\}$. W represents the word weights. Huang (2008) used the $tfidf$ value as word weights to measure the distance between two documents, that is $w_{t,a} = tfidf(d_a, t)$.

The Jaccard coefficient, also referred as Jaccard similarity coefficient, measures similarity between the union of objects. For text documents, the Jaccard coefficient compares the total weight of shared words with the total weight of words that are present in either of the two documents but are not shared words. The formula of Jaccard coefficient is:

$$J(\overrightarrow{t_a}, \overrightarrow{t_a}) = \frac{|\overrightarrow{t_a} \cdot \overrightarrow{t_a}|}{|t_a|^2 + |t_b|^2 - \overrightarrow{t_a} \cdot \overrightarrow{t_a}}$$

The Pearson’s correlation coefficient can also measure how two vectors are related. The correlation coefficient indicates the ratio between the covariance and the standard deviations of the objects. When measuring similarity of two text documents by given the word set $T = \{t_{1,}\ t_{2,} \ldots, t_{n,}\}$, a commonly used mathematical form of the Pearson’s correlation coefficient is:

$$P(\overrightarrow{t_a}, \overrightarrow{t_a}) = \frac{n \sum_{t=1}^{n} w_{t,a} \cdot w_{t,b} - TW_a \cdot TW_b}{\sqrt{[m \sum_{t=1}^{n} w_{t,a}^2 - TW_a^2][m \sum_{t=1}^{n} w_{t,a}^2 - TW_b^2]}}$$

$TW_a = \sum_{t=1}^{n} w_{t,a}$ and $TW_b = \sum_{t=1}^{n} w_{t,b}$.

Finally, cosine similarity measures the cosine of the angle between vectors. Given two documents represented as word vectors, the cosine distance between them is:

$$C(\overrightarrow{t_a}, \overrightarrow{t_a}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_a}}{\overrightarrow{|t_a|} \times \overrightarrow{|t_a|}}$$

One important characteristic of cosine similarity is that it is independent of document length. Qian et al. (2004) compared the Euclidean distance to cosine similarity for nearest neighbor queries, Huang (2008) reviewed the effectiveness of various similarity measures with applications to text clustering, and Vijaymeena and Kavitha (2016) surveyed distance measures used in text mining. Results show performance variations, but no one measure appears to have a consistent advantage across all analytic situations. Experiments by Cha (2007) comparing distance measures suggested that normalized Euclidean distance, cosine distance, Jaccard distance, Minkowski distance (with p=3), and Dice distance (Dice, 1945) tended to perform similarly over many scenarios. Cosine similarity has proven effective in application to word embedding (Ji & Eisenstein, 2013; Kenter & de Rijke, 2015). Based on this previous research, we used cosine distance matrices for distance calculations presented in this article.

2.3 Statistical Properties of Distance Matrices

Generally, a distance matrix is a n-by-n square matrix, where n is the number of objects represented, the diagonal of the matrix is zero, and all other values are zero or positive real numbers (Gower, 1985). Statisticians have examined properties of distance matrices. For example, values from a Euclidean distance matrix (EDM) generally fit a chi distribution, a right tailed distribution whose shape is governed by a single parameter *k* (Liberti & Lavor, 2017, p. 88). For non-Euclidean distances, one might model distance values using an all-purpose distribution such as the generalized lambda distribution, which has four parameters (location, scale, skewness, and kurtosis), or a more specific distribution such as gamma, which has two (shape and scale).

Mulekar et al. (2011, p. 1040) commented that little is known about distributions of distance measures other than in EDMs. Matrices of cosine distances may violate the triangle inequality and the coincidence axiom (Gower, 1985) and thus probably have different properties than EDMs. An alternative approach of measuring overlap between distributions, such as Cliff's Delta (Cliff, 2014), could provide a measure of differences between two samples of distances independent of the underlying theoretical distribution. Either by using a distribution-free test statistic or by identifying an appropriate theoretical distribution, it should be possible to use a statistical test to compare two samples of distance values computed from corpora.

2.4 Candidate Distributions for Distance Values

An appropriate theoretical distribution for distance values could enable modeling the contents of a distance matrix as a sample from a larger universe of similarly generated distances. In this section we discuss candidate distributions to represent the contents of a distance matrix of terms calculated from word embedding vectors.

Theoretical guidance on probability distributions related to word embedding is limited (e.g., Mikolov, 2013). Early indications suggested that word embedding coefficients created by neural network methods exhibited Gaussian distributions (Li et al., 2015), with each dimension centered on or near zero. If we used

Euclidean methods to compute distances between the vectors for neighboring terms, the resulting distance values would be distributed as chi (Liberti & Lavor, 2017, p. 88). Both Euclidean distances and the chi-distribution are unbounded, however, with no theoretical maximum. Therefore, the chi distribution as a model for cosine distances may not be ideal, because cosine distances have an upper bound. Cosine *similarity* values are bounded between -1 and 1. A common calculation of cosine *distance* (1 – cosine similarity) gives values in the range of 0 to 2, while another cosine distance measure is bounded from 0 to 1 (Anderlucci et al., 2019).

Potentially suitable positive, right-tailed distributions include the exponential (El-Sayyad, 1967), inverse Gaussian (Folks & Chikkara, 1978), Gumbel (Landwehr, et al., 1979), gamma (with separate shape and rate parameters; Stacy & Mihram, 1965) and beta (with two shape parameters; Fielitz & Meyers, 1975; also see Kotz & van Dorp, 2004). Each has numerous applications in science and engineering. For example, the Gumbel distribution models collections of extreme values generated by periodic events. Applicability of these distributions to non-Euclidean distance matrices is poorly researched. Our second study evaluated several of these theoretical distributions as possible candidates for representing matrices of cosine distances.

In the methods described below, we computed a distance matrix from a set of word vector summaries, using a pretrained d=50 word embedding model. Given one document D, represented by its word vectors $\vec{t}$, the word vector summary was represented as: $D = \sum_{k=1}^{n} t_k$. Where $\vec{t}$ is a multidimensional vector over the word set T = $\{t_{1,}\ t_{2,}\ \ldots, t_{n,}\}$. Each document thus had a summary vector in d=50 space, and we used cosine distances to calculate the proximity of each pair of these documents, i.e., within the corpora, $C = \{d_1, d_{2,\ldots} d_n\}$. We ignored the diagonal because it is all zeroes and discarded the redundant top triangle, leaving ((n * n)/2) – n unique distance values to represent a corpus of n documents.

3. Overview of Research Methods

Our work explored using word vectors and matrices of cosine distances in these three studies:

1. Confirmatory analysis of summary compositionality using a database of synonymous phrases;
2. Evaluation of distributions of cosine distance matrices representing corpora of short social media posts; and,
3. Monte Carlo analysis evaluating statistical tests for comparing cosine distance values from corpora.

In the first study, our goal was simply to lend support to a method of composing word vectors suggested by Mikolov et al. (2013; also see Salehi et al., 2015). Given work on vector arithmetic for analogies, we expected that the sum or average of individual word vectors in a short phrase should produce a summary vector helpful for examining the proximity of the phrase to another phrase. To assess this idea, we used the paraphrase database (PPDB) developed by Pavlick (Pavlick et al., 2015). The syntactic version of the English PPDB contains more than 1.7 million short phrases paired with close synonyms. For example, one entry contains the brief phrase "an understanding of," while the synonym says, "awareness about." We hypothesized that, across a substantial sample of such phrases, the summary vector for a phrase should be significantly closer to the summary vector for its synonym than to that of another randomly chosen phrase.

The second study focused on assessing fit to theoretical distributions using samples of cosine distances computed from social media corpora. We extracted twenty-four topically grouped postings from various areas of Reddit, Twitter, and YouTube as source data. For each corpus of postings, we summarized each post as a vector and used these to calculate a cosine distance matrix. We assessed the fit of each cosine distance matrix to various theoretical distributions.

In the third study, we used a Monte Carlo simulation to compare methods of pairwise hypothesis testing on cosine distance matrices. Here the goal was to assess which tests could accurately assess whether the distances among documents in one corpus were credibly different than those in another corpus.

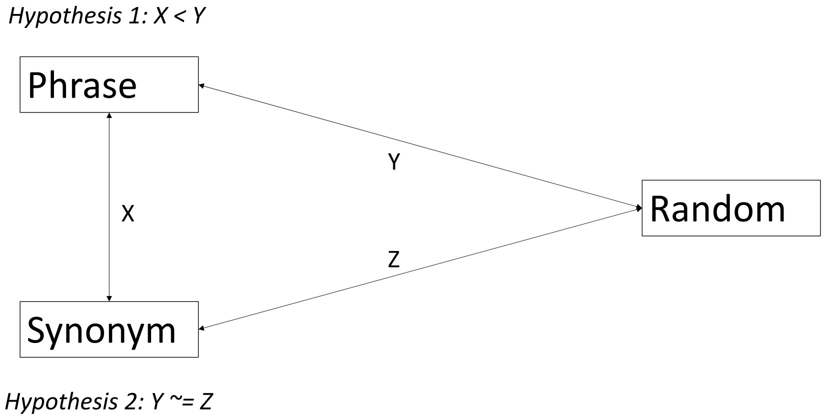
4. Results

4.1 Study 1: Compositionality of Word Vectors

We obtained the English syntactic paraphrase database (Pavlick et al., 2015) comprising 2.7 million phrases along with a close synonym for each entry. To save computing time, we worked with just the first 10,000 phrase pairs from this database. We tokenized words from each phrase and its synonym, while dropping punctuation and numbers and making tokens lowercase. We retained stop words. Individual word vectors were obtained by matching these tokens with the d=50 Wikipedia/Gigaword pre-trained GloVe model published by Pennington et al. (2014). The pre-trained model contains vectors for 400,000 terms including stop words. We summarized each PPDB phrase by combining vectors column-wise over d=50 columns.

For each randomly sampled entry from the PPDB we calculated three cosine distances: the first distance, X, was between a phrase and its synonym. The second distance, Y, was between the phrase and another randomly sampled phrase. The third distance, Z, was between the synonym and that other randomly sampled phrase. Figure 1 shows a schematic representation of these distances.

Figure 1: Schematic Showing Analysis of PPDB Synonyms



We conducted two non-parametric hypothesis tests on each set of phrases. For Hypothesis 1, X < Y, we expected a significant difference between X and Y, with the prediction that a phrase would be closer to its synonym than to some other randomly chosen phrase. For Hypothesis 2, Y ~= Z, we predicted that there should be trivial differences between these two distances, i.e., that the distances from a phrase to a random phrase should be about the same as the distance from the synonym to the same random phrase. A normality test on the data showed that these distance values were highly positively skewed, so we selected the Wilcoxon signed ranks test for paired data as a non-parametric alternative to the paired samples t-test. Weidermann & von Eye (2013) found that the power of the paired signed rank test was similar to and in some cases higher than the power of the paired samples t-test under most conditions. To detect a small effect at an alpha level of 0.05 and a power level of 0.80 with the t-test would only require n=138 paired observations (Cohen, 2013, p. 52), but we wanted to have additional statistical power to assess Hypothesis 2's assertion of no credible difference, so we opted for a larger sample of n=250 observations (Stanton, 2020).

Examples from Mikolov et al. (2013) called for summing vectors for each word to create a summary, whereas other researchers have suggested the arithmetic mean (Salehi et al., 2015). We expected that these two approaches would be statistically equivalent because the computations should not change the resulting direction of the vector in multidimensional space. We also tested a third strategy, to use the median value of individual word vector dimensions as the summary vector for the phrase. Table 1 documents the results for these three methods.

Table 1: Mean Synonym and Random Cosine Distances

	Method: Sum	**Method: Mean**	**Method: Median**
Mean X (phrase to synonym)	0.097	0.097	0.110
Mean Y (phrase to random)	0.252	0.252	0.260
Hypothesis 1: Y > X (Wilcoxon's V)	V = 885, p<.001	V = 885, p<.001	V = 1164, p<.001
X-Y: Cliff's Delta Effect Size	-0.70 (large)	-0.70 (large)	-0.65 (large)
Mean Z (synonym to random)	0.238	0.238	0.250
Hypothesis 2: Y~= Z (Wilcoxon's V)	V = 16986, p = 0.17	V = 16986, p = 0.17	V = 16264, p = 0.46
Y-Z: Cliff's Delta Effect Size	0.05 (negligible)	0.05 (negligible)	0.04 (negligible)

The first row in Table 1 (Mean X) shows the mean cosine distance between phrases and their synonyms, with separate cells for sum, mean, and median summarization methods. We used the cosine distance function from the text2vec R package, which calibrates cosine distances from 0 to 2, with values near 0 indicating that two vectors are highly similar while values near 2 indicate highly dissimilar vectors. The second row (Mean Y) shows the mean of the cosine distance between phrases and randomly selected alternative phrases. In accord with Hypothesis 1, Mean Y is notably larger than Mean X for all three methods. The Wilcoxon test confirmed Hypothesis 1: The cosine distance from a phrase to its synonym is much smaller than the distance from a phrase to another randomly selected phrase: Cliff's Delta effect size was -0.70 for the sum and mean methods and -0.65 for the median method (generally considered as large effect sizes; Cliff, 2014; Hess & Kromrey, 2004).

For Hypothesis 2, we sought to confirm that the cosine distance from a phrase to a randomly chosen phrase should not be credibly different from the cosine distance from the synonym to the *same* random phrase. Mean Y (the distance of a phrase to a random phrase) and Mean Z (the distance of the synonym to the same random phrase) were quite similar. The Wilcoxon test for Y-Z was not statistically significant for any of the summarization methods and Cliff's Delta effect sizes were negligible. The lower (-0.055) and upper (0.147) bounds of the confidence interval for

Cliff's Delta for the sum and mean methods were similar to those for the median method (-0.060; 0.142). Using procedural recommendations from Stanton (2020) as well as the confidence interval thresholds suggested by Romano et al. (2006), these would be considered trivial effects. Thus, as further support for compositionality, a phrase and its synonym do not have credibly different cosine distances to another random phrase.

Thus, results in Table 1 suggest that, when combining vectors of individual words within a short phrase, it does not matter whether one computes the sum or the mean of the component vectors: Resulting cosine distances are identical. The median summarization method, in which the median value for each dimension is selected for each position in the d=50 vector, showed similar results to the other two methods. Effect sizes for the median method were different than for the sum and mean methods, raising the possibility for future investigation that the median might work better for summarizing phrases where outliers existed among the component word vectors.

4.2 Study 2: Distributions of Distance Values

Study 1 suggested that combining individual word vectors for short phrases produces sensible results with respect to cosine distances among vectors representing those phrases. Therefore, a matrix of distances calculated from vector summaries may contain useful information about how closely the brief documents within a corpus relate to one another. All else equal, a distance matrix containing many values clustered near zero would suggest that the original phrases were topically homogeneous, whereas a matrix containing a wide range of values or many large values would suggest that the original phrases were topically heterogeneous. Before evaluating comparative tests, it would be helpful to have a distributional model for cosine distances. To this end, we extracted twenty-four corpora of social media texts.

One source of texts was the social media discussion platform Reddit. Reddit divides into smaller groupings called subreddits, where each subreddit covers a topical area such as a sport (e.g., baseball). For each of thirteen subreddits, we extracted comments by sampling the first ten conversation threads listed on the main page. A second source of comments was YouTube. Here we chose six videos, where each video had numerous top-level comments. Finally, we conducted hashtag searches on Twitter to find families of tweets ostensibly pertaining to the same topic. Each hashtag yielded at least 2000 tweets. Table 2 shows the corpora from thirteen subreddits, six YouTube videos, and five Twitter hashtag searches sorted by number of posts sampled (second column). The third column shows the number of distinct terms in the resulting term-document matrix.

Table 2: Corpora Sources, Overviews, and Results of Distributional Fit Tests

Corpus Source & Name	**Posts**	**Terms**	**Successful Fit Tests**
Reddit: mentalillness	40	896	beta
Reddit: Basketball	61	481	Gumbel, IG, beta
Reddit: HateCrimeHoaxes	173	1127	gamma, beta
Reddit: psychotherapy	206	2093	IG, beta
Reddit: Cricket	394	1860	Gumbel, gamma, IG, beta
Reddit: AgainstHateSubreddits	498	3578	IG, beta
Reddit: Anxiety	711	3987	IG, beta
Reddit: TopMindsOfReddit	744	4220	Gumbel, IG, beta
Reddit: IncelTears	879	3738	IG, beta
Reddit: depression	943	4884	Frechet, exp, beta
Reddit: sports	1076	3317	Gumbel, gamma, IG
Reddit: baseball	1366	3898	gamma, IG, beta
Reddit: insanepeoplefacebook	1505	5175	IG, beta
YouTube: Movie Review	1736	4069	Gumbel, gamma, IG, beta
YouTube: Cooking Video	2018	4191	Gumbel, gamma, IG, beta
YouTube: Tutorial Video	2038	4011	Gumbel, IG, log norm, beta
Twitter: NFL	2152	16335	Gumbel, IG, beta
YouTube: Song Video	2520	5053	Gumbel, gamma, IG, beta
Twitter: #iPhoneSE	2883	10572	Gumbel, gamma, IG, beta
Twitter: #Easter	3134	22874	Gumbel, IG, beta
Twitter: #COVID19	3273	22090	Gumbel, IG, beta
Twitter: #MeAt20	3830	12718	Gumbel, gamma, IG, beta
YouTube: Product Review	7391	8444	Gumbel, gamma, IG, beta
YouTube: Cat Video	8023	8661	gamma, IG, beta

Note: exp – exponential; IG – inverse Gaussian.

After tokenizing postings within each corpus, we created vector summaries (using the mean procedure documented in Study 1) with the d=50 Wikipedia/Gigaword GloVe model (Pennington et al., 2014) and then calculated a cosine distance matrix from vector summaries. We tested goodness of fit to twelve possible distributions for each sample of distances using the goft and goftest R packages: Cauchy, Gumbel, Frechet, generalized Pareto, exponential, gamma, inverse gaussian (Wald), Laplace, normal, log normal, Weibull, and beta. We downsampled to n=250 (for Gumbel and Frechet) or n=350 (all others) from each matrix because of sample size constraints on the estimators used in these procedures. Results showed that all but one corpus ("Reddit: sports") fit the beta distribution. Three other distributions worthy of additional exploration included Gumbel, gamma, and inverse Gaussian distributions.

Tables 3 and 4 show descriptive statistics. Table 3 shows four moments for the distance matrix of each corpus (mean, standard deviation, skewness, and kurtosis) as well as the Cliff's Delta effect size for each pairwise comparison. Table 3 shows comparisons among corpora *within* the same collection, whereas Table 4 shows pairwise values *between* collections. Recall that Cliff's Delta is a non-parametric measure of effect size (with a range of -1 to +1) for a comparison of two samples of data (Cliff, 2014; Hess & Kromrey, 2004). Cliff's Delta provides information on whether observations from one sample are generally larger than or smaller than observations from another without any assumptions about the underlying distributions. A positive Cliff's Delta value shows that distance values from the corpus shown in the row are generally larger than those from the column. A negative value shows the converse.

Table 3: Pairwise Comparisons Within Collections (Cliff's Delta)

Corpus Name	**Mean**	**SD**	**Skew.**	**Kurt.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1. r/IncelTears	.31	.24	1.69	6.38	–											
2. r/AgainstHate.	.32	.25	1.82	6.96	.05	–										
3. r/insanepeople.	.32	.24	1.44	5.08	.06	.01	–									
4. r/TopMindsOf.	.28	.20	1.71	6.45	-.05	-.10	-.11	–								
5. r/HateCrimeH	.29	.27	1.34	4.15	.23	.19	.18	.30	–							
6. r/depression	.26	.17	2.64	9.04	**-.51**	**-.56**	**-.55**	**-.50**	**-.68**	–						
7. r/Anxiety	.21	.19	2.62	11.38	-.32	-.38	-.38	-.30	**-.54**	.25	–					
8. r/mentalillness	.25	.22	1.35	3.58	-.20	-.26	-.26	-.18	-.43	.32	.10	–				
9. r/psychotherapy	.21	.18	3.02	14.51	-.30	-.36	-.37	-.28	**-.54**	.29	.04	-.06	–			
10. r/baseball	.36	.24	1.41	5.24	.16	.12	.10	.22	-.07	**.63**	**.48**	.36	.47	–		
11. r/Basketball	.31	.18	1.03	4.07	.10	.06	.04	.17	-.13	**.58**	.43	.30	.43	-.07	–	
12. r/Cricket	.37	.22	1.05	3.76	.21	.17	.15	.28	-.02	**.66**	**.52**	.40	**.52**	.05	.12	–
13. r/sports	.39	.24	1.27	4.65	.26	.23	.21	.33	.04	**.69**	**.57**	.44	**.57**	.11	.18	.06
1. Movie Review	.32	.22	1.49	5.47	–											
2. Product Review	.40	.26	1.18	4.23	.19	–										
3. Cat Video	.47	.28	1.00	3.41	.35	.16	–									
4. Cooking Video	.52	.29	0.69	2.69	.43	.26	.10	–								
5. Song Video	.31	.21	1.52	5.93	.00	-.19	-.36	-.44	–							
6. Tutorial Video	.33	.20	1.29	5.21	.07	-.13	-.31	-.39	.07	–						
1. #COVID19	.25	.15	1.94	8.55	–											
2. #Easter	.30	.17	1.47	6.08	.20	–										
3. #iPhoneSE	.30	.17	1.47	6.34	.19	-.01	–									
4. #MeAt20	.33	.22	1.44	5.35	.20	.02	.03	–								
5. #NFL	.27	.15	1.83	9.15	.08	-.13	-.11	-.14	–							

Notes: Cliff's Delta values are calculated as row corpus minus column corpus. Values of |d|<0.147 are negligible, |d|<0.33 are small, |d|<0.474 are medium, and |d|>=0.474 are considered large. Large values are bolded.

Table 4: Pairwise Comparisons Between Collections (Cliff's Delta)

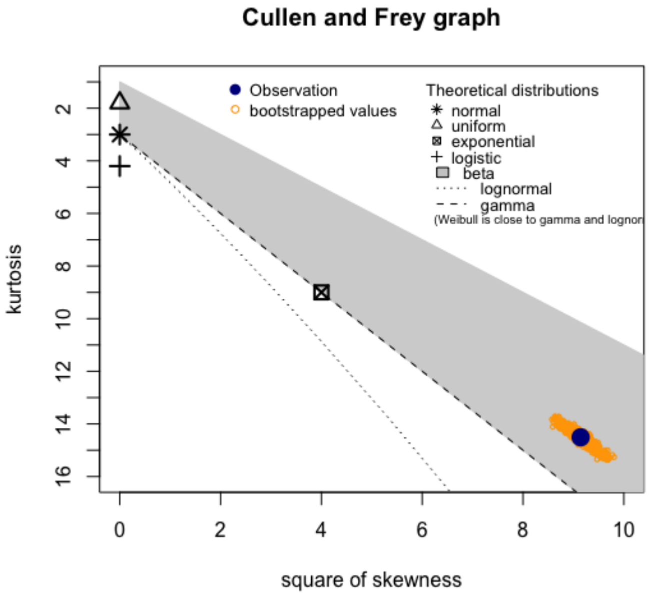
Corpus Name	**Movie**	**Product**	**Cat**	**Cooking**	**Song**	**Tutorial**	**#COVID19**	**#Easter**	**#iPhoneSE**	**#MeAt20**	**#NFL**
1. r/IncelTears	-.08	-.25	-.40	**-.47**	-.08	-.14	.06	-.10	-.09	-.11	.01
2. r/AgainstHate.	-.03	-.21	-.38	-.44	-.03	-.10	.12	-.05	-.04	-.06	.06
3. r/insanepeople.	-.02	-.19	-.35	-.42	-.02	-.08	.13	-.03	-.02	-.04	.07
4. r/TopMindsOf.	-.13	-.31	**-.47**	**-.54**	-.13	-.20	.01	-.16	-.15	-.17	-.05
5. r/HateCrimeH	.17	-.02	-.19	-.28	.17	.11	.34	.17	.17	.13	.28
6. r/depression	**-.58**	**-.67**	**-.76**	**-.79**	**-.58**	**-.63**	**-.52**	**-.62**	**-.61**	**-.60**	**-.57**
7. r/Anxiety	-.41	**-.55**	**-.67**	**-.70**	-.41	**-.47**	-.31	-.46	-.44	-.43	-.37
8. r/mentalillness	-.27	-.41	**-.54**	**-.58**	-.27	-.33	-.17	-.31	-.29	-.29	-.23
9. r/psychotherapy	-.40	**-.54**	**-.67**	**-.71**	-.40	-.47	-.29	**-.45**	-.44	-.43	-.36
10. r/baseball	.09	-.09	-.26	-.34	.09	.03	.26	.09	.10	.06	.20
11. r/Basketball	.04	-.16	-.33	-.42	.05	-.02	.23	.04	.05	.01	.16
12. r/Cricket	.15	-.04	-.22	-.31	.15	.09	.33	.15	.16	.12	.27
13. r/sports	.20	.01	-.16	-.25	.21	.14	.38	.21	.22	.17	.33
1. Movie Review	–	–	–	–	–	–	.17	-.01	.00	-.03	.10
2. Product Review	–	–	–	–	–	–	.35	.19	.22	.16	.30
3. Cat Video	–	–	–	–	–	–	**.53**	.37	.38	.32	**.48**
4. Cooking Video	–	–	–	–	–	–	**.59**	.46	.46	.40	**.55**
5. Song Video	–	–	–	–	–	–	.17	-.01	.00	-.03	.10
6. Tutorial Video	–	–	–	–	–	–	.25	.06	.07	.03	.18

Notes: Cliff's Delta values are calculated as row corpus minus column corpus. Values of |d|<0.147 are negligible, |d|<0.33 are small, |d|<0.474 are medium, and |d|>=0.474 are considered large. Large values are bolded.

Cliff's Delta values shown in Tables 3 and 4 indicate a broad range of pairwise differences among corpora. About 40 values fall into the range considered large (|d| >= 0.474) while many values fall into either small (0.147 <= |d| < 0.33) or medium (0.33 <= |d| < 0.474) ranges. The r/depression subreddit had more topical homogeneity than most other corpora both within and between collections. The r/anxiety subreddit also showed large differences with other corpora. The proportion of comparisons with negligible effect sizes (|d| < 0.147) varied substantially across collections. For example, among ten comparisons for the Twitter topics in Table 3, seven showed negligible differences. This suggests that the Twitter corpora are quite similar to one another with respect to topical homogeneity. Note that these effect size thresholds are simply rules of thumb (Romano et al., 2006) for providing an initial, exploratory view of how similar or different two corpora are with respect to homogeneity.

Skewness and kurtosis values shown in Table 3 play a role in identification of distributions from empirical data. Specifically, Cullen and Frey (1999, p. 126) proposed that plots contrasting skewness and kurtosis can guide the choice of a distributional model for a set of data. Figure 2 shows a Cullen and Frey plot using values obtained from the "Reddit: psychotherapy" corpus.

Figure 2: Cullen and Frey Graph Depicting r/psychotherapy subreddit



The large dot near the lower right corner of the graph represents observed kurtosis and squared skewness values for the distance matrix from the "Reddit: psychotherapy" corpus. Surrounding that dot with an irregular cloud, the plotting procedure used 1000 bootstrap samples to understand uncertainty surrounding the kurtosis and squared skewness values represented by the dot. Next, the shaded area, dotted lines, and special symbols represent kurtosis and squared skewness values expected under various probability distributions. For example, the normal distribution is represented by an asterisk near the upper left corner. That follows from the fact that a normal distribution has skewness of zero and kurtosis of three. The shaded region of Figure 1 represents the range of possible kurtosis and squared skewness values for the beta distribution, a family of probability distributions whose shape arises from two positive real valued parameters known as alpha and beta. Beta distributions fit in a bounded interval and model continuous data such as probability values. All corpora we studied, when subjected to the Cullen and Frey graph, had skewness and kurtosis values placing them in the grey region, providing further evidence in support of the fit results shown in Table 2. Most corpora had skewness and kurtosis values placing them near the dotted line, also supporting gamma as a candidate.

4.3 Study 3: Monte Carlo Simulations of Comparison Tests

Study 3 examined the performance of statistical tests for positional, shape, scale, and location differences between samples of distances obtained from the corpora (i.e., tests of two independent samples). We conducted two phases of Monte Carlo simulation analysis to examine performance following the recommendations of Carsey & Harden (2014). In the first phase, we checked the capability of each test to *avoid* detecting a difference when two samples of cosine distances were drawn from the same corpus. In the second phase, we examined each test’s ability to detect a difference between two cosine distance matrices when a difference was expected to be present.

Table 5 contains one row of simulation data for each test: Gumbel, beta, gamma, and inverse Gaussian, plus two non-parametric tests, the Mann-Whitney and the Kolmogorov-Smirnov. For completeness, we included the Student’s t-test though the right-skewed data were a poor fit to the normal distribution. Columns of Table 5 show different sample sizes drawn from the respective corpora from n=50 up to n=800.

Table 5: Test Performance for Samples of Cosine Distances Drawn from the Same Matrix

	n=50	**n=100**	**n=200**	**n=400**	**n=800**
t-test	0.955	0.953	0.953	0.952	0.954
Mann-Whitney	0.956	0.952	0.949	0.947	0.948
Kolmogorov-Smirnov	0.965	0.964	0.959	0.954	0.955

	n=50	**n=100**	**n=200**	**n=400**	**n=800**
Beta-Distribution A	0.532	0.407	0.411	0.467	0.570
Beta-Distribution B	0.419	0.295	0.283	0.338	0.448
Inverse Gaussian	0.970	0.970	0.973	0.972	0.975
Gumbel	0.977	0.979	0.976	0.982	0.986
Gamma-Distribution A	0.533	0.541	0.563	0.574	0.591
Gamma-Distribution B	0.210	0.281	0.365	0.429	0.502

In each trial, two samples of distances were drawn from the same corpus. Random sampling with replacement was used with the sample size shown to draw observations from the lower triangle of each cosine distance matrix listed in Table 2. Each cell represents the percentage of correct decisions (non-significant results) across 6000 simulation runs (24 corpora times 250 trials per corpus). Each test used a nominal p<0.05 decision-making criterion, so a successful test should have a correct detection rate of about 0.95, particularly at larger sample sizes. We were trying to detect true negatives correctly, so cells in Table 5 represent the *specificity* of each test at the respective sample size.

Table 5 shows that the Student's t-test had satisfactory performance at all sample sizes. Two non-parametric tests, the Mann-Whitney U test and the Kolmogorov-Smirnov test, also had good performance. A test of mean differences between two inverse Gaussian distributions, based on work by Folks and Chhikara (1978), performed slightly better than the t-test and the non-parametric tests. A test of differences in the location parameter of two Gumbel distributions (using estimated standard errors; Bury, 1999, p. 273) had the best performance. Tests of differences in shape/scale parameters for beta (using estimated standard errors) and gamma (using a Bayesian test) had poor performance at all sample sizes. Low success rates for the beta and gamma tests arose from incorrect detection of relatively minor differences in the two random samples drawn from each distance.

Next, Table 6 reports results for correctly detecting a significant difference between two *different* corpora when a difference in the two samples was expected to be present. We looked *within* each of the three collections (Reddit, Twitter, YouTube) to make two pairings. For the first pair of corpora, we selected two that seemed as *dissimilar* as possible. For the second pair we chose two that seemed as *similar* as possible. Because little is known about the nature of effect sizes and statistical power for these kinds of comparisons, the goal in choosing one similar and one dissimilar pair was to obtain simulation results averaged across smaller and larger observable differences between corpora. We made judgments about similarity and dissimilarity of corpora by examining the moments of each matrix (mean, variance, skewness, kurtosis) and the pairwise Cliff's delta values in Table 3. Within each cell of Table 6 each we report the correct decision rate averaged across 1500 trials: 250 sample draws from each of two pairs of corpora drawn from each of three collections. Because our goal here was to detect true positives correctly, these values represent the *sensitivity* of each test. Each value also represents the observed statistical power of the test at the given sample size. Conventionally, researchers often aim for power of at least 0.80 when designing studies, so we considered values of 0.80 or above as acceptable.

Table 6: Test Performance for Samples Drawn from Different Matrices

	50	**100**	**200**	**400**	**800**
t-test	0.457	0.548	0.637	0.711	0.777
Mann-Whitney	0.508	0.590	0.698	0.797	0.910
Kolmogorov-Smirnov	0.467	0.576	0.697	0.797	0.907
Beta-A	0.445	0.565	0.555	0.543	0.558
Beta-B	0.686	0.812	0.861	0.886	0.883
Inverse Gaussian	0.427	0.507	0.555	0.551	0.524
Gumbel	0.387	0.463	0.576	0.710	0.837
Gamma-A	0.669	0.718	0.775	0.816	0.878
Gamma-B	0.893	0.884	0.851	0.829	0.840

Table 6 shows poor results at smaller sample sizes: Samples of approximately n=400 distances are needed to achieve acceptable power across a majority of the tests. Note that a corpus containing about thirty comments produces a distance matrix whose lower triangle suffices to sample n=400 unique cosine distances, so this may represent a sensible lower limit for the size of a corpus that one could examine using one of these statistical tests. At n=400 and higher, the Mann-Whitney, Kolmogorov-Smirnov, beta-B, gamma-A, and gamma-B tests performed acceptably. The gamma-B test functioned well even with small sample sizes. The test of the Gumbel distribution had adequate performance, but only at n=800. The Student's t-test, beta-A, and inverse gaussian tests had poor performance even at n=800. This probably resulted from an inability to detect distinctions between corpora whose distance distributions were similar.

5. Discussion

The goal this study was to evaluate proposed methods for computing and analyzing topical homogeneity within corpora. In Study 1, we used a database of synonyms to demonstrate that summarization of brief phrases using word embedding values was workable. Relatedly, these findings confirmed that examining cosine distances among synonymous and non-synonymous phrases could provide useful analytical insights.

Given the success of Study 1, we calculated cosine distance matrices to represent topical distances among comments within each of twenty-four corpora extracted from three online social media sites. A review of the distributions of these distance matrices showed that they were all right-skewed and, as a result of the particular cosine distance calculation we used, contained values bounded between zero and two. In Study 2, we used fit statistics to assess distributional characteristics of cosine distance matrices for fit to theoretical distributions. Plausible candidates included beta, gamma, inverse Gaussian, and Gumbel distributions.

In Study 3, we used Monte Carlo simulations to examine the specificity and sensitivity of several statistical tests that compared positional values estimated from samples of cosine distance data. We surmised that these could represent the topical homogeneity of each corpus. We included tests with broad application in comparing pairs of samples: The Student’s t-test of two independent samples, the Mann-Whitney U-test of two independent samples, and the two-sample Kolmogorov–Smirnov test. The latter two are non-parametric, which seemed advantageous given results from Study 2 showing that distance distributions were skewed. We also included tests for beta, gamma, inverse Gaussian, and Gumbel distributions. Beta and gamma each have two parameters and we tested these separately.

Results showed that tests of positional values from the beta and gamma distributions had substantial power to detect differences but were also prone to false positives. The Student’s t-test and inverse Gaussian test were accurate in avoiding false positives but lacked statistical power to detect an effect when one was expected to be present. A test of the positional value of the Gumbel distribution had the best performance at avoiding false positives but needed a sample size of n=800 to achieve acceptable power in detecting real effects. The Kolmogorov–Smirnov test performed well, though it had slightly less power than the Mann-Whitney test to avoid false negatives at smaller sample sizes (n=50 and n=100). These results accord with Büning (2002), who concluded that the Kolmogorov-Smirnov two sample test was superior in robustness and power to other common two-sample tests when comparing right-skewed distributions. More generally, both the Mann-Whitney and Kolmogorov-Smirnov tests are regarded as effective, robust tests for skewed data (e.g., Özçomak et al., 2013) so we recommend either of these tests for comparing samples of cosine distance values.

Many research applications arise from these results. First, Study 1 adds to existing evidence that summarization using word embedding vectors can be useful. Many corpora exist with an abundance of short messages. Word embedding can be used to summarize any pair of short messages such that the cosine distance between the two messages represents linguistic similarity.

Next, results suggested that by summarizing each message in a group using word embedding vectors and computing a cosine distance matrix from these vectors, one develops a data structure representing topical homogeneity. The lower triangle of the resulting distance matrix contains a positive, bounded, right-skewed distribution summarizing linguistic similarity among a message set. A set of messages whose dissimilarity values cluster near zero share more linguistic similarities among them than a group of messages where some or many dissimilarity values are more distant from zero.

Study 3 showed that a non-parametric, two-sample test such as the Kolmogorov–Smirnov can assess which of two samples of cosine distances contains greater topical homogeneity. From a research perspective, this provides an analytical tool that allows investigators to understand if one set of messages contains more common terminology relative to some other set. With sufficient experience, researchers might be able to establish benchmarks for levels of similarity that represent various phenomena, such as discussions that drift “off topic” or discussions where disruptive users intentionally post material outside the topical scope of a thread.

These results also have practical applications. Given the volume of postings to social media services, content moderation has become more and more important. Automated tools to detect content categories and semi-automated tools to assist human moderators generally rely on statistical or machine-learning analysis. The capability of comparing topical homogeneity demonstrated in this paper provides a new tool to support this work. For example, a moderator could measure topical homogeneity to examine whether a thread contains a diversity of viewpoints by comparing the distribution of distances from the thread relative to another reference thread. By comparing snapshots of a thread at two different points in time, an algorithm could monitor whether postings containing novel terminology has emerged (Mei & Zhai, 2005). Previous studies have examined topical homogeneity within a single document (Gledson and Keane, 2008) or a corpus consisting of well-written short paragraphs such as abstracts of scientific papers (Sahlgren & Karlgren, 2005). However, as online communication proliferates, researchers often aggregate related colloquial texts such as tweets and then examine the aggregate structure to examine particular issues. For example, collections of tweets have been used for analyzing the climate of public opinion during elections (Skoric et al. 2012; Khatua et al., 2015). Two popular aggregation techniques involve hashtag search and keyword search. One assumption of aggregating texts by hashtag or keyword search is that texts using the same hashtags are actually discussing the same topic. However, this assumption may not always hold (Alvarez-Melis et al., 2016), so having a test to check the aggregation could be useful. Our findings suggest that tests of topical homogeneity could detect off-topic texts by monitoring distributional changes at different points in time. Finally, topical homogeneity could serve as a metadata element in information retrieval applications by summarizing the linguistic diversity among documents in a corpus. In these various uses, organizations may benefit from having this additional tool to support the work of content moderators.

The paper contains limitations that could spur further exploration. First, the word embedding vector representations used to create summaries were trained using one method and at a low dimensionality value of d=50. Although this seemed to work well in the investigation of synonyms, higher dimensional word embeddings used in other applications could yield different results (Sahlgren & Karlgren, 2005). Relatedly, newer methods of summarizing sentences have emerged (e.g., SentenceBERT). It is possible that a more sophisticated method of sentence summarization would improve the results reported in this paper.

Second, Study 2 used empirical methods of evaluating the characteristics of distance matrices generated from our 23 corpora. Fit tests comparing an empirical distribution to reference characteristics of a theoretical distribution can *disconfirm* a candidate distribution but cannot conclusively *confirm* one. Thus, when we conducted Study 2, we rejected several distributions (e.g., the normal distribution) as exhibiting poor fit to the data. Remaining candidates - beta, gamma, inverse Gaussian, and Gumbel – were plausible choices but could not be “proven” as such using fit tests. Ideally, a theoretical basis for understanding the mechanisms that generate these distributions would provide more robust support for the observed data.

Future research can address these limitations by extending the techniques demonstrated here. First, future work should examine the distance matrices generated by different word embedding and summarization techniques. Additionally, it would be helpful to explore whether newer summarization techniques such as sentence embedding could improve the analysis of topical homogeneity. Research can also explore the boundaries of the methods demonstrated in this paper, such as experimenting with corpora with only a few documents as well as with documents that contain lengthier texts. Similarly, studies could shed light on how assessments or comparisons of topical homogeneity will be most useful in practical areas such as content moderation. Convenient tools to extract two or more social media corpora, calculate their respective distance matrices, and apply the appropriate statistical tests would enable a range of applied experiments.

References

Ali, Z., & Mahmood, T. (2020). Complex neutrosophic generalised dice similarity measures and their application to decision making. *CAAI Transactions on Intelligence Technology*, *5*(2), 78-87.

Alvarez-Melis, D., & Saveski, M. (2016, March). Topic modeling in twitter: Aggregating tweets by conversations. In 10th international AAAI conference on web and social media.Anderlucci,

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, *27*(6), 55-65.

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

Büning, H. (2002). Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramer-von Mises two-sample tests. *Journal of Applied Statistics*, *29*(6), 907-924.

Carsey, T. M., & Harden, J. J. (2014). *Monte Carlo simulation and resampling methods for social science*. Thousand Oaks: Sage.

Cha, S.-H., (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 4* (1), 300-307.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5).

Cliff, N. (2014). *Ordinal methods for behavioral data analysis*. New York: Psychology Press.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. New York: Academic press.

Colla, D., Mensa, E., & Radicioni, D. P. (2020). Novel metrics for computing semantic similarity with sense embeddings. *Knowledge-Based Systems*, *206*, 106346.

Conia, S., & Navigli, R. (2020, December). Conception: Multilingually-enhanced, human-readable concept vector representations. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3268-3284).

Cullen, A.C. & Frey, H.C. (1999), *Probabilistic techniques in exposure assessment*. New York: Plenum Press.

Das, S., Ghosh, S., Bhattacharya, S., Varma, R., & Bhandari, D. (2019, March). Critical Dimension of Word2Vec. In *2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)* (pp. 202-206).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297-302.

Dokmanic, I., Parhizkar, R., Ranieri, J., & Vetterli, M. (2015). Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, *32*(6), 12-30.

Dumais, S. T. (1994). Latent semantic indexing (LSI) and TREC-2. *NIST Special Publication SP*, 105-219.

El-Sayyad, G. M. (1967). Estimation of the parameter of an exponential distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(3), 525-532.

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2018). Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70-86.

Fielitz, B. D., & Myers, B. L. (1975). Concepts, Theory, and Techniques: Estimation of Parameters in the Beta Distribution. *Decision Sciences*, *6*(1), 1-13.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis: 1–32. Reprinted in F. R. Palmer, ed. (1968). *Selected Papers of J. R. Firth 1952-1959*. London: Longman.

Folks, J. L., & Chhikara, R. S. (1978). The inverse Gaussian distribution and its statistical application—a review. *Journal of the Royal Statistical Society: Series B* (Methodological), 40(3), 263-275.

Gledson, A., & Keane, J. (2008, August). Measuring topic homogeneity and its application to dictionary-based word sense disambiguation. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) (pp. 273-280).

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, *67*, 81-97.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77-94.

Harris, Z. (1968). *Mathematical structures of language*. New York: John Wiley & Sons, New York.

Hess, M. R., & Kromrey, J. D. (2004, April). Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under non-normality and heterogeneous variances. In *annual meeting of the American Educational Research Association* (pp. 1-30).

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57).

Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).

Huang, A. (2008, April). Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).Ji, Y., & Eisenstein, J. (2013, October). Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 891-896).

Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411-1420).

Khatua, A., Khatua, A., Ghosh, K., & Chaki, N. (2015, January). Can# twitter_trends predict election results? Evidence from 2014 indian general election. In 2015 48th Hawaii international conference on system sciences (pp. 1676-1685). IEEE.

Kotz, S., & Van Dorp, J. R. (2004). *Beyond beta: other continuous families of distributions with bounded support and applications*. Singapore: World Scientific Publishing.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25* (2-3), 259-284.

Landwehr, J. M., Matalas, N. C., & Wallis, J. R. (1979). Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research*, *15* (5), 1055-1064.

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530-539).

Lee, H., Kwak, J., Song, M., & Kim, C. O. (2015). Coherence analysis of research and education using topic modeling. *Scientometrics*, *102*(2), 1119-1137.

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., & Chen, E. (2015, June). Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Liao, Q., Yang, F., & Zhao, J. (2013, November). An improved parallel K-means clustering algorithm with MapReduce. In 2013 15th IEEE International Conference on Communication Technology (pp. 764-768). IEEE.Liberti, L., & Lavor, C. (2017). *Euclidean distance geometry: an introduction*. Cham, Switzerland: Springer.

Lin, Y., & Wang, Y. M. (2020). Decision framework of group consensus with hesitant fuzzy linguistic preference relations. *CAAI Transactions on Intelligence Technology*, *5*(3), 157-164.

McAuliffe, J. D., & Blei, D. M. (2007). Supervised Topic Models. Proceedings of *Advances in Neural Information Processing Systems (NIPS 2007),* pp. 121-128.

Mei, Q., & Zhai, C. (2005, August). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198-207).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 262-272).

Montanari, A. L. & Viroli, C. (2019). The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015. *Statistical Science, 34*(2), 280-300.

Mulekar, M. S., Boone, J. S., & Aryal, S. (2011). Estimation of Sampling Distributions of the Overlapping Coefficient and Other Similarity Measures. In Karian, Z. A., & Dudewicz, E. J., *Handbook of Fitting Statistical Distributions with R. Chapman and Hall/CRC*, 1039-1090.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008, August). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 542-550).

Nguyen, T., O'Dea, B., Larsen, M., Phung, D., Venkatesh, S., & Christensen, H. (2017). Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia tools and applications, 76*(8), 10653-10676.

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, *182,* 104842.

Novák, M., Mírovský, J., Rysová, K., & Rysová, M. (2018, October). Topic–focus articulation: a third pillar of automatic evaluation of text coherence. In Mexican International Conference on Artificial Intelligence (pp. 96-108). Springer, Cham.

Özçomak, M. S., Kartal, M., Senger, Ö., & Çelik, A. K. (2013). Comparison of the Powers of the Kolmogorov-Smirnov Two-Sample Test and the Mann-Whitney Test for Different Kurtosis and Skewness Coefficients Using the Monte Carlo Simulation Method. *Journal of Statistical and Econometric Methods*, *2*(4), 81-98.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217-235.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015, July). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 425-430).

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 99-107).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064-1082.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006, February). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys. In *annual meeting of the Florida Association of Institutional Research* (pp. 1-33).

Sahlgren, M., & Karlgren, J. (2005, November). Counting lumps in word space: Density as a measure of corpus homogeneity. In International Symposium on String Processing and Information Retrieval (pp. 151-154). Springer, Berlin, Heidelberg.

Salehi, B., Cook, P., & Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 977-983).

Seyeditabari, A., & Zadrozny, W. (2017, May). Can word embeddings help find latent emotions in text? preliminary results. In proceedings of *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*.

Shumskaya, A. O. (2013). Comparing of Euclidean and Mahalanobis metrics while solving the problem of the text origin identification. American Journal of Control Systems an Information Technology, N2, 27.

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

Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., & Jiang, J. (2012, January). Tweets and votes: A study of the 2011 singapore general election. In 2012 45th hawaii international conference on system sciences (pp. 2583-2591). IEEE.

Stacy, E. W., & Mihram, G. A. (1965). Parameter estimation for a generalized gamma distribution. *Technometrics, 7*(3), 349-358.

Stanton, J. M. (2020). Evaluating equivalence and confirming the null in the organizational sciences. *Organizational Research Methods.*

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). 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* (pp. 952-961).

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*(2), 19-28.

Wiedermann, W., & von Eye, A. (2013). Robustness and power of the parametric t test and the nonparametric Wilcoxon test under non-independence of observations. *Psychological Test and Assessment Modeling*, *55*(1), 39-61.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Takefuji, Y. (2018). Wikipedia2vec: An optimized implementation for learning embeddings from wikipedia. *arXiv preprint arXiv:1812.06280*.