Semantic Topology: a New Perspective for Communication Style Characterization

Barbara Scalvini
University of the Faroe Islands
Leiden University
barbaras@setur.fo

Alireza Mashaghi
Leiden University
a.mashaghi.tabari@lacdr.leidenuniv.nl

Abstract

We introduce semantic topology, a novel framework for discourse analysis that leverages Circuit Topology to quantify the semantic arrangement of sentences in a text. By mapping recurring themes as series, parallel, or cross relationships, we identify statistical differences in communication patterns in long-form true and fake news. Our analysis of large-scale news datasets reveals that true news is more likely to exhibit more complex topological structures, with greater thematic interleaving and long-range coherence, whereas fake news favors simpler, more linear narratives. These findings suggest that topological features capture stylistic distinctions beyond traditional linguistic cues, offering new insights for discourse modeling.

1 Introduction

Recent years have witnessed increasing intersections between bioinformatics and natural language processing, spanning applications such as phylogenetic trees (Enright and Kondrak, 2011), text mining for automated information retrieval (Chen et al., 2013), network analysis (Mehler, 2006a), and bio-inspired computational models for language evolution (Araujo, 2007). However, many opportunities remain unexplored, particularly with the growing availability of data and machine learning techniques. One such opportunity involves leveraging *topology*, originally used to describe the three-dimensional structure of biological macromolecules, to analyze textual structure. Specifically, we propose applying the *Circuit Topology* (CT) framework (Mashaghi, 2021; Mashaghi et al., 2014; Golovnev and Mashaghi, 2020; Scalvini et al., 2020, 2023a) to quantify the arrangement of semantic "contacts" within a sequence of sentences. In biology, CT models how loops formed by intra-chain contacts influence macromolecular function; in language, these loops correspond to semantic recurrences, representing the re-introduction of a theme or topic.

Our work builds on the idea that repeated or closely-related concepts form the backbone of coherent discourse. By treating each sentence as a node and linking sentences with high semantic similarity, we can apply CT to capture how themes are repeated, interleaved, or nested (Figure 1). This yields topological descriptors—series, parallel, and cross—whose relative prevalence can characterize a text's rhetorical structure. Additionally, we introduce the *globularity score* as a means of weighting these topological features by the distance over which a semantic connection spans.

As a case study, we perform semantic topological analysis on two datasets, the Kaggle Fake News Dataset[1] and the Fake News Corpus (Pathak and Srihari, 2019). Fake news manipulates public opinion purposefully, often with profound political and social consequences. Although many linguistic features have been suggested for fake news extraction, it is still unclear which ones are the most meaningful (Choudhary and Arora, 2021; Verma et al., 2021). Here, we characterize these articles in topology space, and observe statistical differences in the semantic arrangement of true and fake news. We demonstrate how topological descriptors highlight characteristic stylistic patterns in long-form true and fake news (extended narratives composed by at least 15 sentences). All code used for the analysis outlined in this paper is freely available on Github [2].

2 Previous work

Discourse analysis has long sought to characterize how concepts and arguments are structured within a text. Early influential frameworks, such as *Rhetori-*

[1] https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets

[2] https://github.com/circuittopology/Semantic_Topology

Findings of the Association for Computational Linguistics: ACL 2025, pages 9223–9233

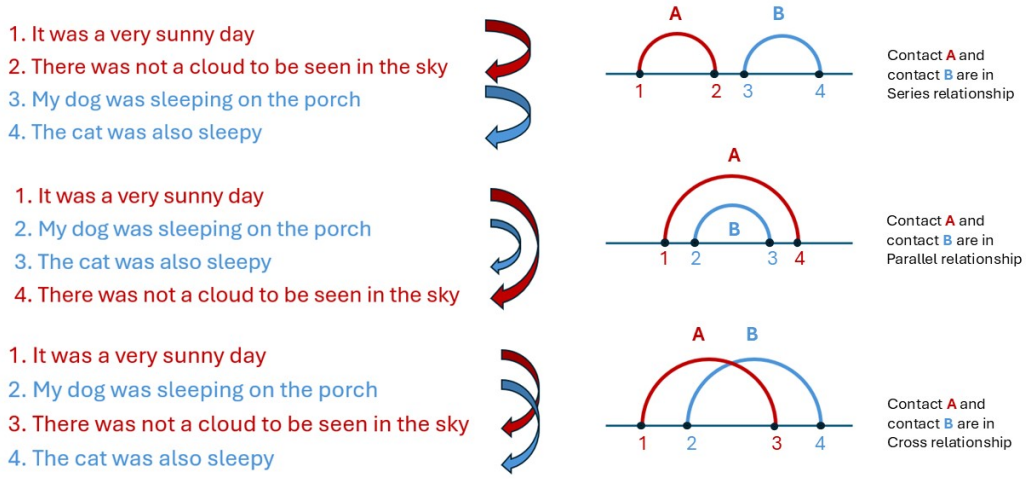

Figure 1: Graphical representation of the three fundamental topological relationships applied to a simple text example. The example illustrates two semantic contacts: Contact A - formed by the red-highlighted sentences related to the weather - and Contact B - formed by the blue-highlighted sentences about pets.

cal Structure Theory (RST) (Mann and Thompson, 1988), break down discourse into elementary discourse units and analyze how these units connect via rhetorical relations. Subsequent approaches, such as *Argumentation Mining* (AM) (Stede and Schneider, 2018) have focused on the detection of Argumentative Discourse Units (ADUs), such as claims, premises, and conclusion structures. In this framework, ADUs can have relationships of different kinds, e.g., support or attack. Similarly related work in *coherence modeling* (Hobbs, 1979; Barzilay and Lapata, 2008) has used features like entity grids or lexical chains to capture patterns of entity distribution in a text and coherence assessment. In parallel, the rise of *distributional* and *embedding-based* paradigms of semantics have offered new opportunities for automated discourse exploration. Topic modeling methods (Mersha et al., 2024; Gao et al., 2019; Das et al., 2015; Le and Mikolov, 2014) cluster semantically similar passages or documents, thereby allowing for the identification of high-level themes within texts. Topological frameworks have also been previously applied for the modeling of language structure, although these efforts remain sporadic. For instance, Mehler (2006b) represent text as a "small-world" network of textual units, revealing structural properties such as clustering or centrality. Moreover, related studies in *Topological Data Analysis* (TDA) (Zhu, 2013) have investigated persistent homology of text features, identifying semantic loops within documents. Misinformation detection represents a particularly appealing application for discourse analysis. Although many approaches have thus far prioritized lexical or stylistic cues (Rubin, 2017; Conroy et al., 2015), sentiment features (Bhutani et al., 2019), or source reliability, several recent studies have demonstrated the value of rhetorical structure for fake news detection. For instance, Kuzmin et al. (2020) incorporate RST-based features (bag-of-rst) to detect fake news in Russian, Karimi and Tang (2019) introduce a hierarchical discourse-level structure that outperforms prior baselines, and Atanasova et al. (2019) exploit both contextual and discourse-level information to improve claim classification and fact-checking. Work by Rubin and Lukoianova (2015) and Pisarevskaya (2017) likewise show that coherence relations and nuclearity can highlight differences between deceptive from truthful documents.

3 Methods

3.1 Circuit Topology

Circuit topology provides a general framework for the topological characterization of folded linear (molecular) chains. It involves two main steps: (1) identifying contacts that form loops within the chain, and (2) characterizing how these loops are arranged relative to one another.

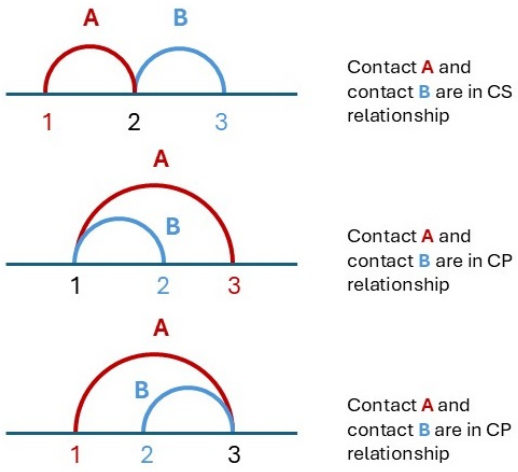

Figure 2: Graphical representation of the Concerted Series (CS) and Concerted Parallel (CP) relation. Here, one sentence in each diagram (sentence number in black) has a semantic contact with two other sentences.

This concept can be applied to text by treating it as a chain of sentences (or other elementary units, such as words). For simplicity, we will focus on sentences as the basic building blocks of the chain in this analysis. A contact in this textual chain occurs when two sentences are semantically similar.

Figure 1 provides an example of this idea with a text made of four sentences. Two sentences pertain to the weather (highlighted in red), and two pertain to pets (highlighted in blue). Each pair of semantically similar sentences forms a contact. For example:

- Contact A connects the weather-related sentences: "It was a very sunny day" and "There was not a cloud to be seen in the sky."
- Contact B connects the pet-related sentences: "My dog was sleeping on the porch" and "The cat was also sleepy."

If we decide that two sentences (S1,S2) are connected (i.e., they form a contact) when their similarity passes a certain threshold, this example has two contacts: A and B.

Figure 1 also shows three possible arrangements of these contacts:

- Example a: **Series relationship**. Each contact is separate. The weather sentences (Contact A) and pet sentences (Contact B) don't overlap. Themes are kept apart.
- Example b: **Parallel relationship**. Contact A (weather) wraps around Contact B (pets). The reader moves from one theme (weather) to the next (pets) and back to the first theme (weather).
- Example c: **Cross relationship**. The two contacts intersect. The reader alternates between themes: weather → pets → weather → pets.

In Circuit Topology, these arrangements are called Series (S), Parallel (P), and Cross (X), respectively. These relationships are defined between pairs of contacts. For example, we would say, "Contact A is in a Series relationship with Contact B."

We can give a formal definition of topological relations as follows. Let us call $C_{i,j}$ the contact connecting sentence i and j, and $C_{r,s}$ the contact connecting sentence r and s. Then we define the three topological relations as:

$$C_{i,j}\, S\, C_{r,s} \iff [i,j] \cap [r,s] = \emptyset$$

$$C_{i,j}\, P\, C_{r,s} \iff [i,j] \subset (r,s)$$

$$C_{i,j}\, X\, C_{r,s} \iff [i,j] \cap [r,s] \notin \begin{matrix}\{[i,j],[r,s]\} \cup \\ \mathcal{P}(\{i,j,r,s\})\end{matrix}$$

Here, $\mathcal{P}$ denotes the powerset, i.e., all subsets of a set including the null set ($\emptyset$). In simple terms, the series relations describes two contacts with no intersection; the parallel relation describes a situation where one contact is completely enveloped by the other; while the cross relation describes two intersecting contacts. Sentence indexes (i, j, r, s) are assigned by the order with which they appear in the text.

In practice, a sentence can present semantic contacts with multiple other sentences. In that case, multiple arcs will span from a sentence (Figure 2), and two sub-categories of topological relations are created - Concerted Series, CS (a Series relation between two contacts where one contact site is shared) and Concerted Parallel, CP (a Parallel relation where one contact site is shared). Formally, CS and CP relations are expressed as:

$$C_{i,j}\, \mathrm{CS}\, C_{r,s} \iff \begin{matrix}([i,j] \cap [r,s] = \{i\}) \vee \\ ([i,j] \cap [r,s] = \{j\})\end{matrix}$$

$$C_{i,j}\, \mathrm{CP}\, C_{r,s} \iff \begin{matrix}([i,j] \subset [r,s]) \wedge \\ (i = r \vee j = s).\end{matrix}$$

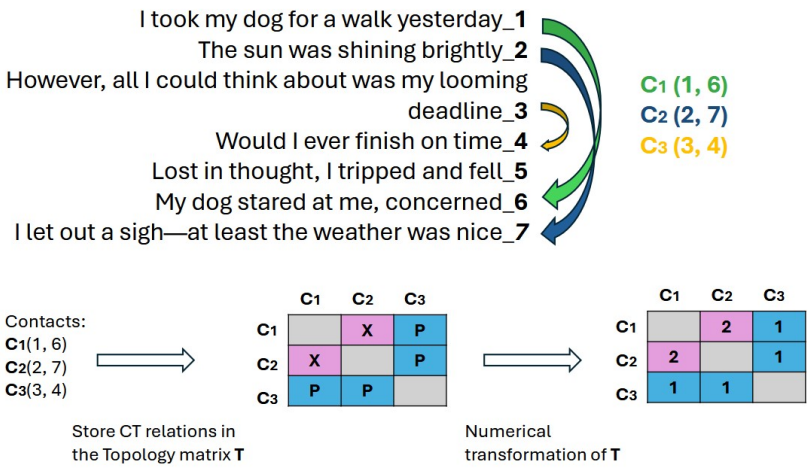

Figure 3: Example of an application of CT analysis to a short text. The threshold for semantic similarity was set to 0.4 for demonstrative purposes. Three semantic contacts emerge, connecting sentences 1 and 6 (C_1), 2 and 7 (C_2), and 3 and 4 (C_3). C_1 and C_2 intersect, placing them in a cross relation. Meanwhile, both C_1 and C_2 encompass C_3, meaning they are in a parallel relation with it. These relationships are then stored in the topology matrix, which can be further processed numerically to compute the Globularity Score. For this text, the topological fractions are as follows: X = 0.33 (one-third of the relations are cross), P = 0.66 (two-thirds of the relations are parallel), S = 0 (no series relations) and G = 0.583

For a first order analysis, these contacts are treated as regular series or parallel relations. This creates no issue in the pipeline, as contacts are inherently broken down into sentence pairs. Therefore, if a sentence is in contact with two other sentences, it will result in two separate contacts involving that sentence. In certain specialized applications CP and CS relations are treated separately (Scalvini et al., 2022). However, for most first-order applications, it is acceptable to incorporate them into series and parallel relations, as we do in this paper.

For more information about the formalism, we invite the reader to refer to Mashaghi (2021); Mashaghi et al. (2014). The topological relation between each pair of contacts in a text can be stored in a matrix for further analysis. Therefore, the *topology matrix* **T** will be a $N \times N$ matrix, where N is the total number of contacts in the text, and its possible elements are S, P and X (Figure 3).

3.2 Semantic Topology: the arrangement of semantic themes within a text

Originally developed for biopolymers, the Circuit Topology framework can also be applied to discourse analysis by examining S, P, and X relations in text. Series relations (S) reflect a linear, sequential structure where topics are introduced and concluded before moving on, in a list-like manner, resulting in minimal thematic interaction and lower complexity. In contrast, parallel (P) and cross (X) relations introduce structural intersections that shape how information is organized. Parallel relations (P) suggest a circular structure, where a topic is introduced, interrupted by other content, and revisited later. Cross relations (X) indicate an alternation between different semantic areas, with a back-and-forth discussion of topics. In sufficiently long texts, all three relations typically coexist in varying proportions, allowing each text to be mapped in a three-dimensional space based on its semantic topology.

3.3 Second order analysis: the Globularity Score

So far we have only considered the relative amount of S, P and X relations in a text. However, for a second order analysis, it is also possible to weigh these relations differently, depending on the range of the relation: that is, whether the two contacts forming the relation occur close by within the text, or not. This consideration is particularly relevant for P and X relations. The reason is intuitive: long-range alternation and reoccurrence of themes within the text might indeed indicate a cyclical structure of the text, where the author revisits certain semantic areas to draw their conclusions. On the other hand, short-range alternation of themes might just indi-

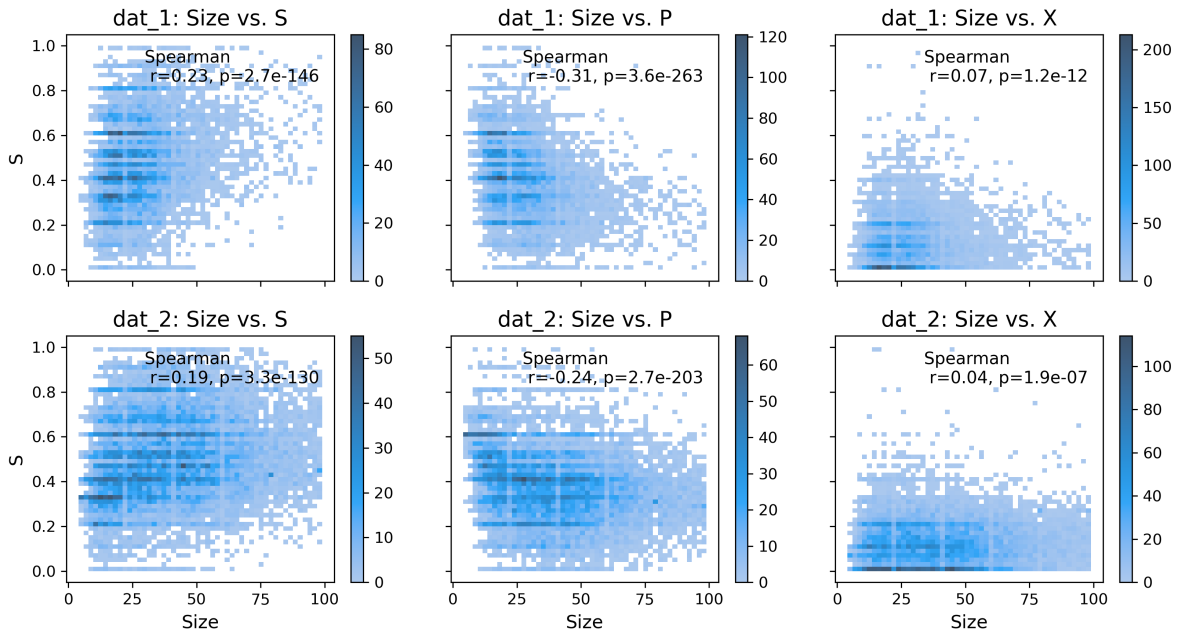

Figure 4: Scatterplot showing the relationship between the three topological relations (S, P, and X) and text size, measured as the number of sentences per text, along with their corresponding correlation values.

cate a very confused and non-linear information delivery. Under this assumption, we attempt to create a unified descriptor of topology, by weighing relations more when the contacts forming them are further away in the topology matrix. We call this descriptor the *globularity score* G, as it describes the tendency of the text to adopt a cyclical, globular topology.

In order to define the G score, we need to exploit the topology matrix $\mathbf{T}$ representation. We assign the three topological relations a numerical score, to allow for computation: S = 0, P = 1, and X = 2 (Figure 3). We choose this convention to emphasize the role of semantic alternation: S corresponds to no alternation, in P we alternate semantic areas once, and in X, twice. For a matrix of size $N \times N$, the score is defined as the weighted average of diagonal contributions, normalized by their theoretical maximum. The weighted contribution for diagonal d (off-diagonal entries $(i, i+d)$) is given by:

$$\text{score}_d = w_d \cdot \frac{1}{N-d} \sum_{i=0}^{N-d-1} \mathbf{T}[i, i+d]$$

$$\text{norm}_d = w_d \cdot \frac{1}{N-d} \sum_{i=0}^{N-d-1} \mathbf{N}[i, i+d]$$

where $w_d = \frac{d}{N-1}$ is the diagonal weight emphasizing distant relationships, and $\mathbf{N}$ is a normalization matrix with all off-diagonal values set to their theoretical maximum. The theoretical maximum is given by a matrix where all entries are X (X = 2). The total weighted score is:

$$\text{score\_tot} = \sum_{d=1}^{N-1} \text{score}_d$$

$$\text{norm\_score} = \sum_{d=1}^{N-1} \text{norm}_d$$

The final globularity score G is the ratio of the total weighted score to the total normalization:

$$G = \frac{\text{score\_tot}}{\text{norm\_score}}$$

ensuring $G \in [0, 1]$, where higher values indicate more compact and strongly connected topologies.

3.4 Hopping between semantic clusters: calculating the rate of semantic alternation

The concept of contacts based on semantic similarity is based on the notion that semantically similar elements will reside in spatial proximity in the multi-dimensional embedding space. As such, it is possible in principle to cluster all sentences that belong to the same semantic area together (Angelov, 2020; Grootendorst, 2022; Dieng et al., 2020). As

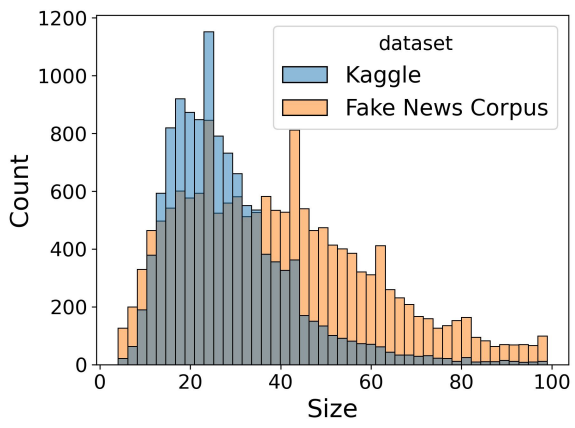

Figure 5: Distribution of article length (expressed in number of sentences) for the Kaggle dataset and the Fake News Corpus.

we have seen in the previous section, the text might hop between different semantic areas and create different semantic topologies. In the embedding space, this phenomenon translates into hopping between different semantic clusters. We can therefore calculate what is the rate associated to hopping, for each text. Here, the rate R is defined as:

$$R = \frac{1}{\langle N_s \rangle},$$

where $\langle N_s \rangle$ is the average number of consecutive sentences appearing in the text from the same cluster before hopping to a different cluster. N_s is therefore analogous to a characteristic lifetime of the cluster - in analogy with lifetime in biomolecular topology analysis (Scalvini et al., 2023b). In practice, we extract sentence embeddings and reduce their dimensionality to five components using UMAP (McInnes et al., 2018). Next, we apply HDBSCAN—a hierarchical extension of DBSCAN that explores multiple epsilon values and selects the most stable configuration—to cluster the reduced embeddings. Finally, we compute $\langle N_s \rangle$ and use it to determine the hopping rate R.

3.5 Experimental setup

In this study, we analyze articles sourced from the Kaggle Fake News Dataset, which contains 21,417 *True* and 23,481 *Fake* articles, as well as from the Fake News Corpus, from which we randomly select 30,000 "fake" and 30,000 "reliable" entries. We further filter these articles to include only those whose number of sentences, N_s, lies in the range $15 \leq N_s \leq 100$. This filtering yields about 43% of the original Fake News Corpus sample (25922 articles) and 56% of the Kaggle dataset (25060 articles). We do this filtering to ensure a sufficient number of sentences to observe a statistically significant number of semantic contacts. As part of this process, we also filtered out articles that were deemed unfit for analysis because they were in languages other than English or because they consisted solely of a title and advertisement banners. To calculate semantic similarity, we use the Sentence Transformer library (Reimers and Gurevych, 2019), employing the top-scoring STS model on the MTEB leaderboard (`bilingual-embedding-large`). We set the cosine similarity threshold to $t_s = 0.6$, ensuring that sentences in contact exhibit a moderate degree of thematic or contextual overlap. Lastly, we retain only those articles having more than five contacts ($N > 5$) for analysis of topological fractions (16298 articles for the Kaggle dataset and 13590 from the Fake News Corpus extract).

We found that choosing a higher semantic similarity threshold would significantly reduce the number of articles retained, raising questions about the method's validity. For instance, a threshold of 0.7 would result in only 3099 articles from the Fake News Corpus extract and 919 from the Kaggle dataset. These numbers are not unexpected, since articles do not typically display multiple sentences with such a high reciprocal semantic similarity. On the other hand, lowering the threshold to 0.5 or below would typically result in numerous contacts between unrelated sentences. Considering these factors, we considered a threshold of 0.6 to be a balanced and appropriate choice for this analysis.

The topological analysis consists of the following steps:

- Calculate the percentages of P, S, and X relations in each article, as well as the Globularity score G, and examine how these parameters vary with article size.
- Compare the statistical distributions of P, S, X, and G between true and fake news.
- Compute the hopping rate between semantic clusters within each article.
- Compare the statistical distributions of these hopping rates in true versus fake news.

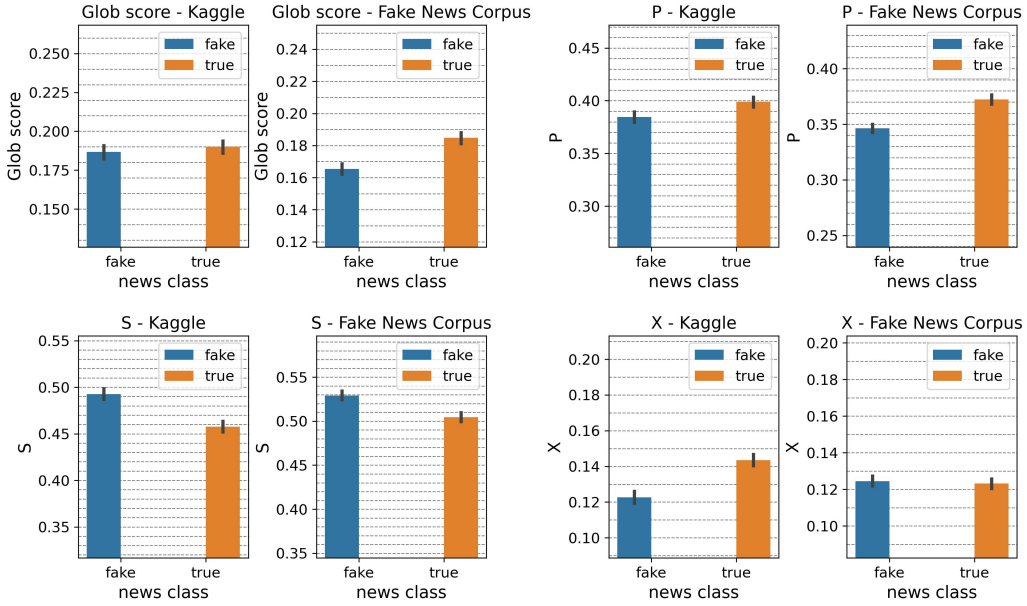

Figure 6: Bar plots reporting the average value of the three topological fractions (parallel, series, and cross) and the Globularity score for the two dataset extracts. Error bars represent 95% confidence intervals (computed via bootstrapping) around the mean.

4 Results

4.1 The ratio of semantic arrangements correlates with text size

To investigate whether the overall ratio of P, S, and X depends on the length of the text, we correlated them with the number of sentences in an article (Figure 4). Our findings show a weak but statistically significant relationship between text length and both the series and parallel ratios, with a small positive correlation for series and a small negative correlation for parallel. By contrast, the cross ratio exhibits an almost zero correlation. Importantly, these trends hold across both datasets.

These consistent correlations imply that there are fundamental principles governing how texts are constructed from a semantic topology perspective. In practical terms, texts are often divided into paragraphs, each focusing on a single topic, which naturally increases the number of series relationships (see example a, Figure 1). Consequently, longer texts have more paragraphs, more self-contained units, and therefore more series relations—resulting in a lower proportion of other topological relationships. Because these ratios depend on text size, comparing datasets with widely different length distributions (Figure 5) can be problematic. For this reason, we chose to analyze the two datasets separately.

4.2 Semantic topology identifies differential patterns in communication style between true and fake news.

Since our topological fractions (S, P, and X) depend nonlinearly on text size, it is essential to ensure that any observed differences between true and fake news are not merely artifacts of differing length distributions. To address this, we resample our data so that both classes (true vs. fake news) have an equivalent size distribution. This balancing sampling yields, for a similarity threshold of 0.6, a subset of 8708 articles from the Kaggle dataset, and 10452 articles from the Fake News Corpus extract. We calculated the semantic topology descriptors (S, P, X fraction and Globularity score) over these two subsets and compared the results for true and fake news. A consistent pattern seems to emerge (Figure 6). In both datasets, true news seem to display statistically higher Globularity score, lower series ratio, and higher parallel ratio. The cross fraction, on the other hand, seems to be higher for fake news in the Kaggle dataset, while results are inconclusive for the Fake News Corpus extract, where these two values seem equivalent. If we compare the distributions, we see that p values are consistently below the 0.05 threshold, except for cross

	True/Fake news – p value	
	N = 600	**N = 4000**
Kaggle - balanced		
Globularity score	0.0429	1.60e-08
series	0.0005	1.08e-11
parallel	0.0129	3.20e-06
cross	0.0003	1.05e-18
Hopping rates	0.0079*	0.0355
	N = 600	
FNC - balanced		
Globularity score	0.0261	1.58e-06
series	0.0040	1.21e-08
parallel	0.0002	5.60e-20
cross	0.4854	9.05e-05
Hopping rates	0.1830*	0.0004

Table 1: Comparison of True vs. Fake News (p-values) for Kaggle and FNC under balanced sampling. P-values were computed using the Kolmogorov–Smirnov test, except those marked with an asterisk (*), which were calculated using the Mann–Whitney U test because Levene's test indicated statistically equivalent variances (p > 0.01).

relation in the Fake News Corpus dataset, where p = 0.4854 for samples containing 600 articles. For all other topological descriptors, we observe statistically significant differences between the fake and true news classes, which become statistically more pronounced if we increase sample size (N = 4000 articles). While results for the cross relation remain ambiguous on the FNC dataset extract, we observe statistical difference for the Globularity score, which takes into account both parallel and cross relations, with cross relations weighing double in the calculation. This indicates that not only cross relation does play a role, but the range of the interaction matters as well. In simple terms, when a topic, concept, or semantic area is recalled in distant parts of the text, it plays an important role in shaping the overall semantic topology profile of the text, and possibly, its stylistic properties.

4.3 True news exhibit a slightly higher rate of hopping between semantic clusters

The results of the clustering procedure reveal that the two datasets differ, on average, in their number of semantic clusters, most likely due to differences in article length distributions (Figure 5). Specifically, the Fake News Corpus has an average of $N_{C,\text{FN}} = 4.95 \pm 0.02$ clusters per article, whereas the Kaggle dataset has $N_{C,\text{K}} = 3.66 \pm 0.01$. Moreover, we find that the rate of hopping between clusters correlates with the total number of clusters N_C, with a Spearman correlation coefficient of $r = 0.518$ (p-value < 0.01). Intuitively, this implies that the more semantic clusters an article possesses, the more frequently it alternates among them. In contrast, the correlation between hopping rates and article size (number of sentences) is weaker, at $r = 0.173$ (p-value < 0.01). Following the procedure in previous sections, we sampled the two datasets separately to create size-equivalent distributions of true and fake news. This yielded 19,888 articles from the Kaggle dataset and 20,498 from the Fake News Corpus. The resulting hopping rates R for true and fake news are depicted in Figure 7. In both datasets, true news exhibit slightly higher hopping rates than fake news. However, the difference is small, visible only in very large samples of articles (Table 1, $N = 4000$) and is much less pronounced than the difference in topological fractions. Finally, the gap between true and fake news appears to be overshadowed by the larger discrepancy between the two datasets themselves, suggesting that unknown stylistic factors—possibly linked to the selection process of the articles—may influence these observations.

5 Discussion

Here, we propose semantic topology as a candidate for discourse analysis. We found that semantic topology highlights statistical differences in long-form true and fake news, with complex topologies that entail semantic alternation (parallel and cross) being more prevalent in the former. These findings are in line with previous research indicating fake news as a simpler form of communication (Horne and Adali, 2017; Rashkin et al., 2017), although the topological complexity of semantic arrangement identifies a different level of analysis with respect to syntax and word choice. Hopping between different semantic clusters seems to happen less frequently in fake news articles. However, evidence for this particular effect is weak, and it requires further investigation of what other stylistic choices influence the rate of hopping between semantic clusters. Personal perception of semantic similarity is a key concept in the cognitive behavioral sciences (Goldstone, 1994), where it was proven to correlate with personality traits (Richie et al., 2020), and to participate in false memory generation (Buchanan et al., 1999). In consideration of

the findings presented above, there is reason to investigate whether the arrangement of semantically similar units (sentences) in a text might contribute to the psychological allure of fake news, as well as other cognitive phenomena involving effective communication.

The semantic topology framework, as introduced in this paper, represents a novel approach rooted in quantitative, data-driven modeling rather than the qualitative categories and relations typical of classical discourse theory. Establishing a direct correspondence between Circuit Topology (CT) relations and classical discourse frameworks is therefore not straightforward. However, certain patterns of semantic arrangements may be indirectly captured by semantic topology. For example, topic shifting in discourse might manifest as series (S) relations within our framework, where thematically distinct units follow one another without recurrence. Similarly, semantic merging could be represented by transitions from predominantly series topologies toward more parallel (P) and cross (X) structures, indicating increasing thematic interleaving. Moreover, classical discourse frameworks often aim to identify what we refer to as semantic contacts—recurrences of meaning across a text. Relations such as restatement, summary, and evaluation, as framed within Rhetorical Structure Theory (RST), might correspond to 'semantic contacts' in the semantic topology framework. However, our method focuses on capturing the pairwise arrangement of these contacts, offering a complementary perspective. While the model presented does not yet capture the full spectrum of semantic structuring described in discourse theory, it introduces a promising higher-order analytical dimension that warrants further exploration.

6 Conclusion

In this paper, we introduce semantic topology as a novel approach to discourse analysis, leveraging the Circuit Topology framework from bioinformatics. By mapping and quantifying sentence-level semantic relationships, we reveal significant differences in how true and fake news structure recurring themes. Our findings suggest that fake news follows simpler topological patterns, aligning with prior research on deceptive discourse. The Globularity Score and topological descriptors highlight the role of textual loops in coherence and complexity. Future work should explore interactions with other discourse frameworks, broader text types, and the cognitive significance of topological relationships. Additionally, this method offers potential insights into communication style variations across lifespan development and neurological conditions.

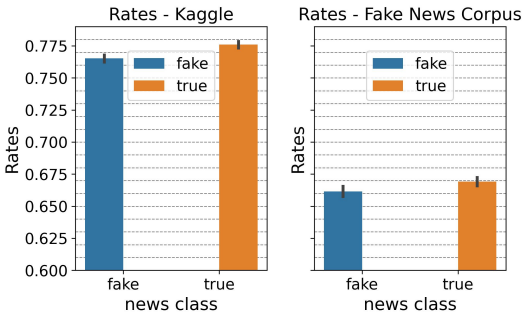

Figure 7: Bar plot showing the average Hopping Rate between semantic clusters for the two dataset extracts. Error bars represent the 95% confidence intervals (computed via bootstrapping) around the mean.

7 Limitations

We focused on long-form fake news articles since sentences serve as our basic unit of analysis and we needed to provide enough data to observe statistical patterns. While this choice may reduce the generalizability of our findings (given that much fake news appears in shorter formats like tweets), our framework is adaptable and could use alternative units (e.g., words or tokens) in future studies. Moreover, although our analysis reveals only very small statistical differences in communication styles, these findings hint at a characteristic communication style detectable by semantic topology. While we believe our conclusions to be significant from the point of view of qualitative analysis, further feature engineering work is needed to make the framework directly applicable for fake news detection.

References

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

Lourdes Araujo. 2007. How evolutionary algorithms are applied to statistical natural language processing. *Artificial Intelligence Review*, 28(4):275–303.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse

information. *Journal of Data and Information Quality*, 11:1–27.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5.

Lori Buchanan, Norman R. Brown, Roberto Cabeza, and Cameron Maitson. 1999. False memories and semantic lexicon arrangement. *Brain and Language*, 68(1-2):172–177.

Hongyu Chen, Bronwen Martin, Caitlin M. Daimon, and Stuart Maudsley. 2013. Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. *Frontiers in Physiology*, 4 JAN(January):1–6.

Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169(February 2020):114171.

Niall Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, Beijing, China. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Jessica Enright and Grzegorz Kondrak. 2011. The application of chordal graphs to inferring phylogenetic trees of languages. *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 545–552.

Wang Gao, Min Peng, Hua Wang, Yanchun Zhang, Qianqian Xie, and Gang Tian. 2019. Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, 61.

Robert Goldstone. 1994. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, Computers*, 26(4):381–386.

Anatoly Golovnev and Alireza Mashaghi. 2020. Generalized Circuit Topology of Folded Linear Chains. *iScience*, 23(9):101492.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):759–766.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov. 2020. Fake news detection for the Russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 45–57, Barcelona, Spain (Online). Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Alireza Mashaghi. 2021. Circuit Topology of Folded Chains. *Not. Am. Math. Soc.*, 68:420–423.

Alireza Mashaghi, Roeland J. van Wijk, and Sander J. Tans. 2014. Circuit Topology of Proteins and Nucleic Acids. *Structure*, 22(9):1227–1237.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Alexander Mehler. 2006a. In search of a bridge between network analysis in computational linguistics and computational biology–A conceptual note. *Proceedings of the 2006 International Conference on Bioinformatics Computational Biology (BIOCOMP'06), Las Vegas, Nevada*, pages 496–500.

Alexander Mehler. 2006b. Text linkage in the Wiki medium - a comparative study. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Melkamu Abay Mersha, Mesay Gemeda yigezu, and Jugal Kalita. 2024. Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. *Procedia Computer Science*, 244:121–132. 6th International Conference on AI in Computational Linguistics.

Archita Pathak and Rohini Srihari. 2019. BREAKING! presenting fake news corpus for automated fact checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362, Florence, Italy. Association for Computational Linguistics.

D. Pisarevskaya. 2017. Rhetorical structure theory as a feature for deception detection in news reports in the russian language. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*. ACM.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2931–2937.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

Russell Richie, Bryan White, Sudeep Bhatia, and Michael C. Hout. 2020. The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior Research Methods*, 52(5):1906–1928.

Victoria L. Rubin. 2017. Deception detection and rumor debunking for social media. In Luke Sloan and Anabel Quan-Haase, editors, *The SAGE Handbook of Social Media Research Methods*. SAGE, London.

Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917.

Barbara Scalvini, Helmut Schiessel, Anatoly Golovnev, and Alireza Mashaghi. 2022. Circuit topology analysis of cellular genome reveals signature motifs, conformational heterogeneity, and scaling. *iScience*, 25(3):103866.

Barbara Scalvini, Vahid Sheikhhassani, Nadine van de Brug, Laurens W. H. J. Heling, Jeremy D. Schmit, and Alireza Mashaghi. 2023a. Circuit topology approach for the comparative analysis of intrinsically disordered proteins. *Journal of Chemical Information and Modeling*, 63(8):2586–2602. PMID: 37026598.

Barbara Scalvini, Vahid Sheikhhassani, Nadine van de Brug, Laurens W. H. J. Heling, Jeremy D. Schmit, and Alireza Mashaghi. 2023b. Circuit topology approach for the comparative analysis of intrinsically disordered proteins. *Journal of Chemical Information and Modeling*, 63(8):2586–2602. PMID: 37026598.

Barbara Scalvini, Vahid Sheikhhassani, Jaie Woodard, Jana Aupič, Remus T. Dame, Roman Jerala, and Alireza Mashaghi. 2020. Topology of Folded Molecular Chains: From Single Biomolecules to Engineered Origami. *Trends in Chemistry*, 2(7):609–622.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11:1–191.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: Word Embedding over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Xiaojin Zhu. 2013. Persistent homology: An introduction and a new text representation for natural language processing. pages 1953–1959.