Topological decoding of biomolecular fold complexity

Scalvini, B.

CHAPTER 6:

BEYOND BIOPOLYMERS: THE CIRCUIT TOPOLOGY OF SEMANTIC SIMILARITY HIGHLIGHTS STATISTICAL DIFFERENCES IN THE COMMUNICATION STYLE OF TRUE AND FAKE NEWS

Fake news detection is one of the most urgent scientific and societal challenges of our time. Current methods involve neural networks analysis, but often rely on fact checking, because of the elusive nature of fake news. Here, we define a new linguistic feature, the topology of semantic arrangement: the specific order in which meaning is conveyed in a text. We characterize semantic arrangement over a fake/true news dataset, and demonstrate that the two groups show statistical topological differences. Semantic topology could therefore represent a new level of analysis, bridging psycho-linguistics and syntax, for communication style description. We suggest an interdisciplinary approach, encompassing the framework of Circuit Topology, originally created for the structural characterization of biopolymers, and pre-trained sentence embeddings (sBERT). Applications of the method span beyond fake news detection, from socio-linguistics to neuroscience, in view of the psycho-linguistics implications of semantic arrangement.

Publications associated to this chapter:

Barbara Scalvini, Alireza Mashaghi, *The circuit topology of semantic similarity highlights statistical differences in the communication style of true and fake news*, UNDER REVIEW

1. INTRODUCTION

One of the most pressing societal issues of our time concerns the proliferation of information availability, which often happens with no quality control. This leads to an abundance of often misleading information, with the potential of going viral, swaying election results, impacting vaccination rates and deeply affecting collective living. One of the main endeavors of natural language processing is to develop automated tools to provide insight in the reliability of available information [1]–[3]. Linguistic approaches mostly involve syntax, rhetoric and discourse analysis [1][4]–[7], but also social network behavior and propagation studies[8] are used to identify fake-news specific features. Recent years have seen a rise in the continuous exchange between bioinformatics and natural language processing approaches, from phylogenic trees [9] , text mining automated information retrieval methods [10], network analysis [11], bio-inspired computational models for natural language evolution [12] and so on. However, opportunities offered by intrinsic parallelisms between linguistic and biological systems have yet to be fully explored and exploited, especially in relation to the increasing availability of data and artificial intelligence tools and know-how.

Here, we suggest the topology of semantic similarity within a text as a new linguistic feature, capable of characterizing the communication style of the writer and ultimately help us discern between true and fake news. We base our analysis on the theoretical framework of Circuit Topology [13]–[16], which was originally created for the description and quantification of intra-chain contact arrangement in biopolymers. In a similar fashion, we can treat text as an ordered chain, where sentences represent the fundamental units - or words, depending on the desired level of resolution. In practice, we rely upon semantic similarity of sentences to identify semantic contacts within the text: we define a contact whenever two sentences score a higher similarity than a pre-defined threshold. In our proof-of-concept work, we analyze the Kaggle dataset for fake/true news to show how the circuit topology of such semantic-similarity defined contacts can be a promising descriptor of communication style, and therefore a linguistic feature for fake news detection.

In order to quantify similarity, we leverage on state-of-the art pre-trained embeddings. Word embeddings are mathematical representations of words, which often consist of vectors in a multidimensional space. Such representations are very powerful, as they capture the syntactic and semantic information conveyed by a word [17]. This encoding results in vectors corresponding to words with similar meanings being in close proximity in the embedding space [17]. As such, word embeddings have played in a fundamental role in the characterization of

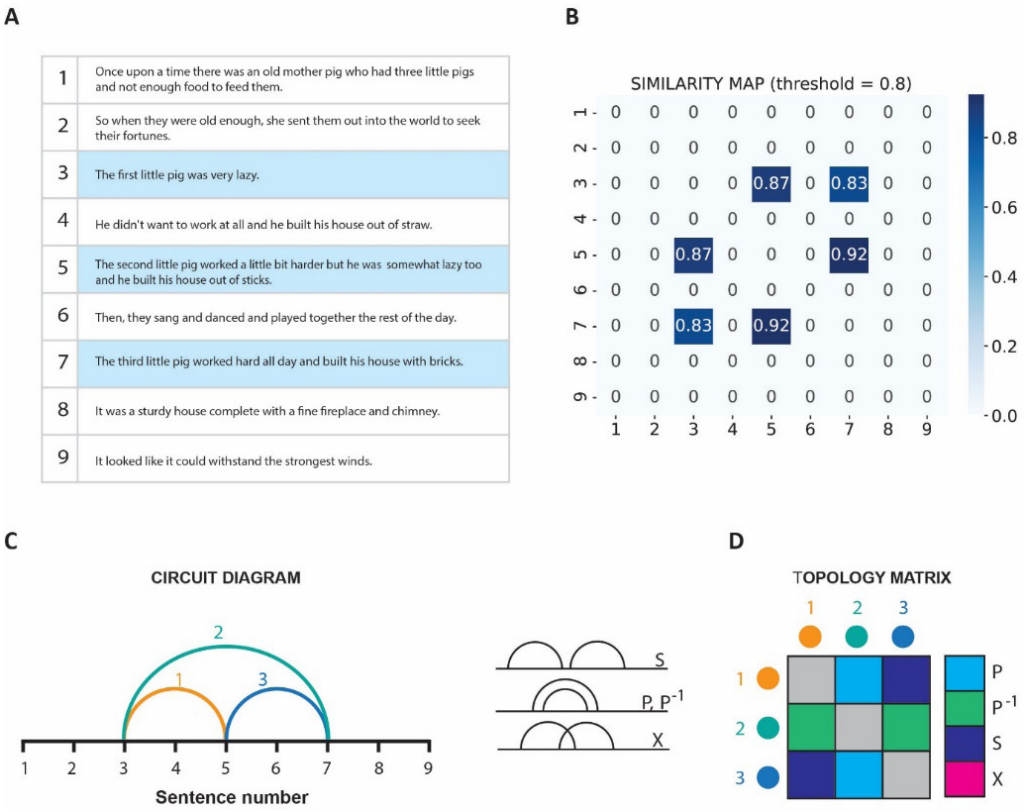

Figure 1. The arrangement of semantic contacts can be expressed in the formalism of Circuit Topology. Semantic contacts are created when two sentences score high in semantic similarity. ***A*** Sample of sentences taken from the traditional fairy tale Three little pigs. Highlighted in blue, sentences which score high in semantic similarity with each other. ***B*** Similarity map calculated from the sample sentences in A. Elements of the matrix scoring less than the threshold (0.8) were put to zero. Non-zero elements identify semantic contacts and contribute to the topological analysis. ***C*** Circuit diagram representation of the similarity map in B. Arcs indicate which sentences are in contact. With this representation, it is easy to identify the three core topological relations from the CT framework: series, parallel and cross. Contact 1 is in series with contact 3, and both contact 1 and 3 are in parallel with contact 2. ***D*** Topology matrix summarizing the topological relations between each pair of contacts in C.

biased communication [18] and fake news specifically [7][19]. Sentence embeddings, similarly, encode the meaning on the sentence level. We chose the sentence-BERT pre-trained embeddings for semantically meaningful similarity comparison, as they showed to outperform other state-of-the-art methods in common semantic textual similarity (STS) tasks.

2. RESULTS AND DISCUSSION

2.1. The topology of semantic contacts quantifies writing style.

True and fake news present statistically distinct semantic topologies, as defined by the reciprocal positioning of their *semantic contacts*. In order to identify such contacts, we use a similarity score based on the dot product of the two sentence embeddings. In Figure 1A and B we can see an example of such scores assigned to any two pairs of sentences in the text sample – the first 9 sentences in the traditional *Three little pigs* fairy tale. The scores are then arranged in a similarity matrix, in Figure 1B. We put to zero all elements that yield a score lower than the chosen threshold (0.8), as these do not represent a semantic contact. The pairs of sentences corresponding to non-zero elements of the similarity map identify contact sites within the text. Sentence 3 is in contact with Sentence 5 and 7, which are also in contact with each other. If we read these sentences carefully, we see they all present elements from semantic fields such as work (*hard work, lazy*) and the domain of building (*built, house, bricks*, etc), therefore justifying the detected similarity. The identified contacts can be represented in the form of a circuit diagram (Figure 1C), in a completely similar fashion to biopolymer circuit topology representations described previously[13][20]. In this representation, it is easy to recognize the fundamental topological relations formalized by circuit topology (Figure 1C, right panel). Contacts are then numbered by scanning the diagram left end to right end. Contacts 1 and 3 are enveloped by contact 2: they are therefore in parallel relation with contact 2, while they are in series with each other. All pairwise relations can then be stored in a topology matrix (Figure 1D), which encodes the overall semantic arrangement of the text.

For texts that are sufficiently long, a prevalence of series (S), parallel (P) or cross (X) semantic contacts yields a completely different structure in which meaning is conveyed. Series relations (S), as the name suggests, indicates a linear, serial delivery, where each topic is addressed and dealt with before moving on to the next item. This topology is list-like, with minimal interaction between different semantic areas. On the other hand, more complex relations such as cross and parallel require structural intersection between different semantic areas. An abundance of parallel relations (P, P^{-1}) might indicate a paragraph with a circular structure, where a topic is addressed at the beginning and at the end of the paragraph, enveloping the discussion of other topics in the middle part of the text. Cross (X), on the other hand, corresponds to an alternation of semantic groups. In this case, the text might discuss topic 1, move on to topic 2, go back to topic 1 and subsequently to topic 2. Ideally, a text with a sufficient number of sentences

will present all three topological relations in different quantities. We can therefore characterize each text by its percentage of series, parallel and cross relations.

2.2. True news present on average a higher percentage of complex topologies – cross, parallel – in their semantic arrangement than fake news.

We calculated the P, S and X percentages for texts contained in the fake news Kaggle dataset (21417 true/ 23481 false news articles). We only selected those texts that contained a number of sentences higher than 15, in order to avoid biases in topological arrangement generated by very short texts, where text length might strongly correlate with the topological fractions. We found that true news present statistically higher percentages of P and X with respect to their fake news counterparts. On the other hand, fake news have higher series fraction (Figure 2A). It is perhaps unsurprising that complex topologies are somewhat avoided in fake news. Previous studies maintained that fake news use simple language and syntax[21], therefore it is reasonable to believe that this simplicity is conserved at the level of semantic arrangement.

Textual forms of expression such as articles are not random collection of units such as words and sentences, but generally come with some degree of structure [22]. From the point of view of semantic arrangement, this structure is expressed in the form of a certain topological prescription (P, S and X percentage) which tends to be constant for longer texts (Figure 2B). Bigger variations of such prescriptions tend to happen for texts below 100 sentences. In view of this overarching topological structure, there is a non-zero correlation between the single topological relations and the size of the article: $r = 0.40$ for S, $r = -0.52$ for P and $r = 0.28$ for X. Since our dataset is not perfectly balanced in terms of article size, we selected 4 subsets of articles with similar sizes, 20-25, 25-30, 30-35 and 35-40 sentences, and performed statistical analysis separately for each subset. With this analysis design, we aim to uncouple the size effect from the stylistic one in topological arrangement, while at the same time obtaining datasets which are small enough to yield meaningful p-values from statistical tests ($N < 5000$). For each of these datasets, we compared the distribution of S, P and X for true and fake news. We also compared these values after normalization by the size of each text – in number of sentences: S/size, P/size, X/size (Figure 2C). For two-tailed p-values, a value of 0.05 is generally used as threshold for statistical significance: a p-value lower than that in this case would indicate the distributions of the specific topological relation for true and fake news to be statistically different. In Figure 2C we can see how this statistical significance is displayed for each relation and each subset, yielding the highest statistical signifi-

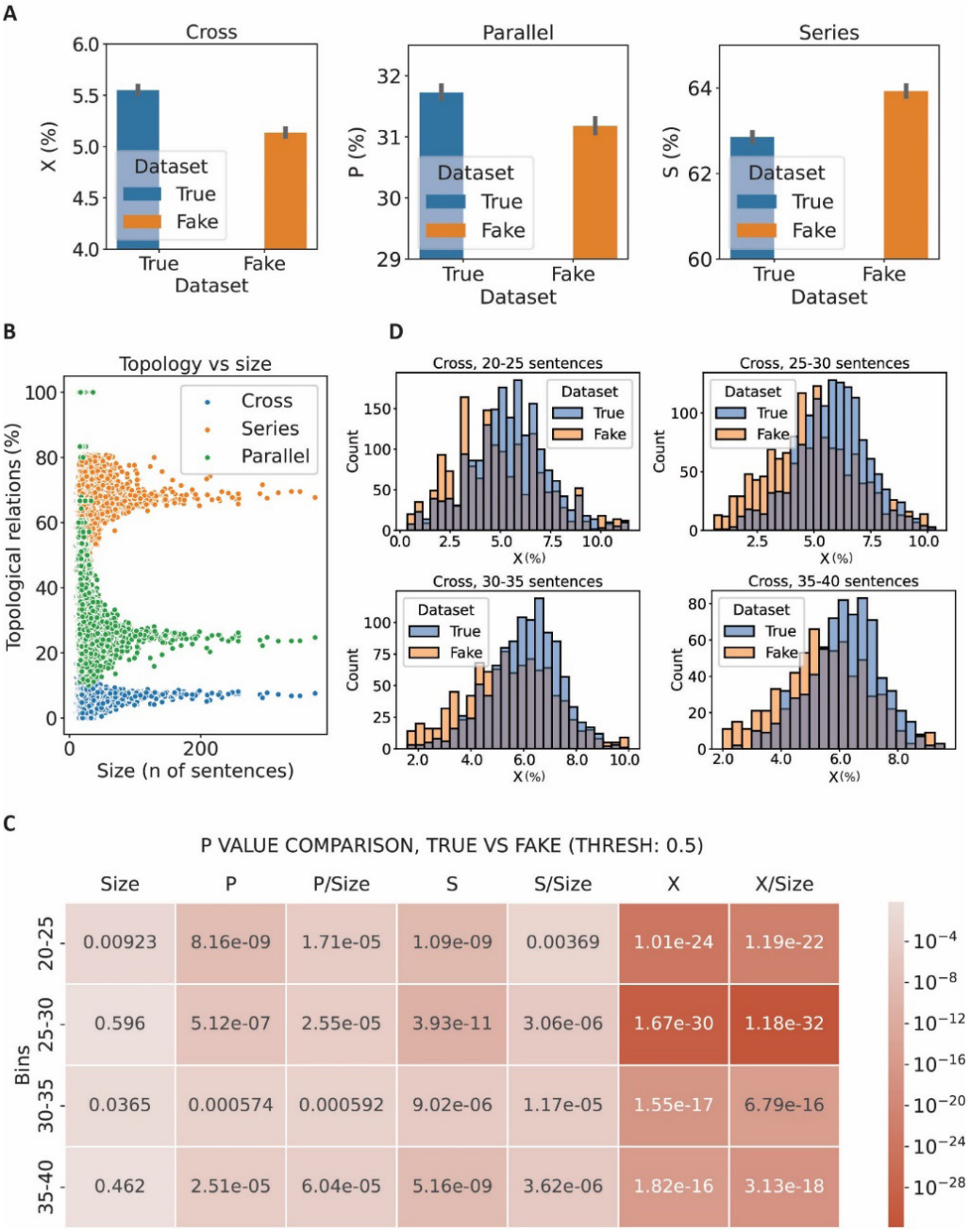

Figure 2. True and fake news present statistically distinct semantic contact topology. ***A*** Barplot of the percentages of parallel, series and cross relations in the true and fake news datasets. Error bars show a 95% CI for the mean. ***B*** Scatterplot of the three topological relations with respect to text size (expressed in number of sentences), for all texts included in the fake/true news dataset after length filtering. ***C*** P-values obtained by statistical comparison of the distributions of (normalized) topological parameters for true and fake news articles, subdivided

into subsets based on number of sentences. Normalization is done by dividing the topological parameters by text size (in number of sentences). P-values lower than 0.05 indicate a statistically significant separation between the two distributions. ***D*** Histogram of cross percentages for true and fake news articles in each subset.

cance for cross, where the p-value is lower by several orders of magnitude.

Being cross also the relation that least correlates with text length, we also see no remarkable difference between X and X/size in statistical significance. In Figure 2D we can see the distributions for cross relations in true and fake news, for the 4 subsets. The displacement towards higher values displayed by the true news dataset indicates that alternation of semantic groups is a much more common phenomenon in legitimate news sources. Therefore, a lack of such semantic pattern could be a linguistic feature to look for in automatic fake news detection.

2.3. True news hop between semantic clusters more frequently than fake news

The concept of contacts based on semantic similarity is based on the notion that semantically similar elements – words, sentences – 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, in a procedure similar to topic identification in documents [23]. As we have seen in the previous section, the text might hop between different semantic areas and create different semantic topologies. We have identified alternation between different semantic groups as a key feature distinguishing true from fake news. In the word embedding space, this phenomenon translates into hopping between different semantic clusters. In order to verify our findings, we can calculate what is the typical rate associated to hopping, for true and fake news. Here, the rate is defined as $1/\langle S_N \rangle$, where $\langle S_N \rangle$ is the average number of sentences appearing subsequently in the text from the same cluster before hopping to a different cluster; SN is therefore analogous to a characteristic *lifetime* of the cluster (in analogy with lifetime in biomolecular topology analysis). For sentence-transformers, the embedding space is a 768-dimensional dense vector space. Therefore, we first perform dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) method [24]. Subsequently, we performed density-based clustering (HDBSCAN) [25] to identify semantic groups. In Figure 3A an example of this procedure is displayed, with a 2-dimensional projection of a true news article with 2 semantic clusters. In Figure 3B, we can couple cluster information to sentence number, in order to observe hopping between the two groups

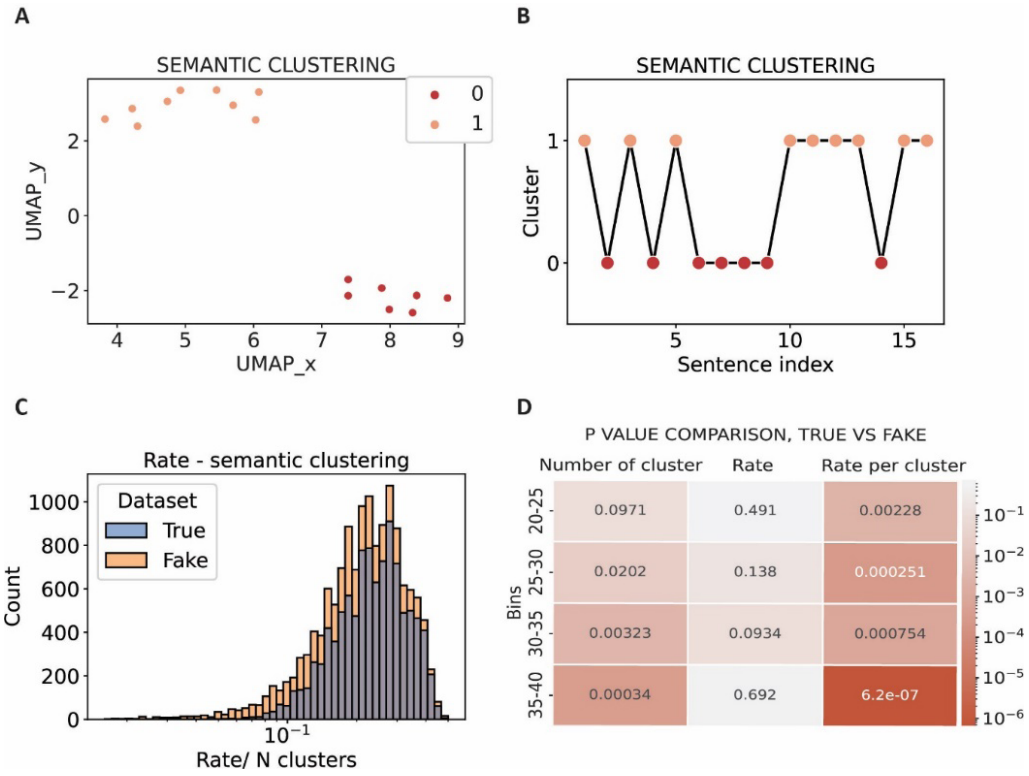

Figure 3. Alternation between different semantic clusters occurs with slower rates in fake news articles. ***A*** Two-dimensional representation of semantic clusters in a sample article from the true news dataset. ***B*** Cluster label (as in A) with respect to sentence number. Cluster hopping is represented as diagonal black lines. ***C*** Average rate of cluster hopping (normalized by number of clusters) for the true and fake news datasets. ***D*** P-values obtained by statistical comparison of the distributions of (normalized) rates for true and fake news articles, subdivided into subsets based on number of sentences. P-values lower than 0.05 indicate a statistically significant separation between the two distributions.

as the text progresses. The characteristic hopping rates (normalized by the number of clusters in the article) is displayed in Figure 3C, where we can observe a longer tail towards low rates for the fake news dataset. This finding confirms our previous topological observation. Alternation between topics and concepts is not only less present in fake news, but also slower. In order to assess statistical significance of our results, we once again split the data into subsets based on text length (20-25, 25-30, 30-35, 35-40). Results can be seen in Figure 3D. In all groups, the fake and true news distributions for normalized rates is statistically significant. A smaller statistical difference is visible also in number of clusters, where fake news have on average 4.52 ± 0.03 and true news 3.82 ± 0.02 clusters. This difference is however probably due at least in part to size effect, as longer texts are overrepresented in our fake news dataset. In conclusion, these rate differences

clearly relate to the topological differences we observe in semantic similarity. A higher rate in topic switching would naturally translate in an abundance of cross relations (topic-alternation) in the topological profile. However, the reasons for this discrepancy between datasets can be multiple. It might be due to a higher stylistic complexity, to an enrichment in topics and nuances in true news, to a higher repetitiveness in fake news content, or most likely to a cumulative effect of all these factors.

3. CONCLUSIONS AND FUTURE WORK

Fake news manipulate public opinion purposefully, often with profound political and social consequences. Many websites exist to check information quality, but we are still lacking spontaneous and wide-spread flagging of fake-news. Although many linguistic features have been suggested for fake news extraction, it is still unclear which ones are the most meaningful [5][7]. Here, we propose semantic topology as a candidate for the task, for its ability to bridge psycho-linguistics, syntax and writing style. Thanks to the formalism of Circuit Topology (CT), combined with pre-trained sentence embeddings, we were able to identify semantic similarity contacts within the text and their arrangement. We found that semantic topology is statistically different in true and fake news, with complex topologies such as parallel and cross being underrepresented in fake news. In particular, switching between different semantic groups seems to happen less and with a slower rate in fake news articles. These findings are in line with previous research indicating fake news as a simpler, more vague form of communication [21][26], although the topological complexity of semantic arrangement identifies a different level of analysis with respect to syntax and word choice. Personal perception of semantic similarity is a key concept in the cognitive behavioral sciences [27], where it was proven to correlate with personality traits [28], and to participate in false memory generation [29]. In consideration of the initial 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. In view of the link between semantic arrangement and cognition, a possible application of quantification techniques in this respect could be applied to speech analysis for diagnosis of dementia and other neurodegenerative diseases [30], which might disrupt language patterns in patients. Generalizations of our findings requires further analysis over different datasets, in order to minimize potential dataset-specific biases, and possibly cancel out size effect contributions. Moreover, the efficiency of semantic similarity contact identification would most likely improve by fine-tuning the sentence-transformers model over our dataset, in order to capture the meaning of domain specific

lexicon more adequately. For example, if we look at the example in Figure 1A – 1B, the model fails to identify Sentence 4 as a possible contact for Sentence 3, 5 and 7, in spite of high content similarity. This loss of context understanding will be improved in future work by fine-tuning the model over the extended dataset, possibly with unsupervised training methods such as TSDAE (Transformer-based Sequential Denoising Auto-Encoder) [31]. Overall, our early findings promise to add a new layer in language complexity characterization, with possible implications not only for fake news detection but also for any type of cognitive or socio-linguistic textual analysis.

4. METHODS

4.1. Data pre-processing

The articles were parsed into sentences using the SpaCy open-source software library [32], with an available trained pipeline for English (*en_core_web_lg*). The first and last sentences of the text was processed to remove references to pictures, twitter handles, videos associated to the article, date and location, in order to remove possible biases due to the recurring editorial structure of the two datasets. Only texts containing more than 15 sentences were retained. The total size of the dataset after filtering was 10244 true and 14283 fake news. Sentence embeddings were then extracted with the python framework SentenceTransformers [33]. We chose the pretrained *all-mpnet-base-v2* model, as it was reported to have the best all-round quality for many use cases on the SBERT.net website [34].

4.2. Circuit Topology analysis

Threshold for semantic similarity for results reported in the paper was set to 0.5. Thresholds equal to 0.6, 0.7 and 0.8 were also tested. Overall result trends do not change with choice in threshold, although 0.5 was found to be the optimal threshold to ensure fake/true news separation while also being computationally convenient. Topological relations were assigned based on sentence indexes, following the algebraic relations defined by Mashaghi et al. [14]. In this particular instance, no distinction was made between concerted (CS, CP) and non-concerted relations. Concerted relations are a subset of topological relations where one contact site is shared. We found that in linguistic systems, with our definition of semantic contacts and contact sites (sentences), concerted parallel are preponderant and do not constitute an exception in terms of semantic arrangement. For

this reason CP and P, as well as CS and S relations were grouped together.

4.3. Semantic clustering

Sentence embeddings were reduced in dimensionality using the Uniform Manifold Approximation and Projection (UMAP) method. Parameters used for dimensionality reduction were n_neighbors= 4, n_components= 5, min_dist=0.0 and metric='cosine'. Data displayed in this chapter were obtained for random_state = 42. After the dimensionality reduction, the HDBSCAN open source software library was used for clustering [25]. Parameter inputs for the hdbscan. HDBSCAN instance were min_cluster_size=2, metric='euclidean', cluster_selection_method='eom' and alpha = 1.3.

4.4. Statistical analysis

Separation in subsets based on text length was made by selecting texts containing between 20-25, 25-30, 30-35 and 35-40 sentences. These intervals were chosen because most articles in our filtered dataset fall in the 20 – 40 range, therefore offering enough statistics for analysis. An equal number of sentences from fake and true news datasets was included in the subsets. The number was chosen by taking the minimum number m between $N_{f,i}$ and $N_{t,i}$, where $N_{f,i}$ and $N_{t,i}$ are the number of articles falling in subset i from the fake and true news datasets respectively. Whenever the number of articles in interval i exceeded m, the articles were picked randomically (random_state = 42). The obtained distributions of topological parameters were then tested for normality (Shapiro-Wilk test) and for equal variance (Levene test) in distribution comparison. P values in Figure 2C and 3D were then obtained by the following rules:

- Whenever the two distributions resulted to be normal and with equal variances, the Student's T-test was used for comparison.
- If the two distributions had unequal variance, the Welch's T-test was used.
- If one or both distributions failed the normality requirement a non-parametric test was preferred (Mann-Whitney U Test in case of equal variance, Kolmogorov-Smirnov otherwise).

All tests were two-tailed. All p-values below the 0.05 threshold were considered significant. All correlations were quantified by Spearman correlation coefficient.

5. REFERENCES

[1] P. Nordberg, J. Kävrestad, and M. Nohlberg, "Automatic detection of fake news," CEUR Workshop Proceedings, vol. 2789, pp. 168–179, 2020.

[2] J. Zhang, B. Dong, and P. S. Yu, "FakeDetector: Effective fake news detection with deep diffusive neural network," Proceedings - International Conference on Data Engineering, vol. 2020-April, pp. 1826–1829, 2020, doi: 10.1109/ICDE48307.2020.00180.

[3] J. Shaikh and R. Patil, "Fake news detection using machine learning," Proceedings - 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, iSSSC 2020, vol. 2020, 2020, doi: 10.1109/iSSSC50941.2020.9358890.

[4] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake News Early Detection: A Theory-driven Model," Digital Threats: Research and Practice, vol. 1, no. 2, pp. 1–25, 2020, doi: 10.1145/3377478.

[5] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," Expert Systems with Applications, vol. 169, no. February 2020, p. 114171, 2021, doi: 10.1016/j.eswa.2020.114171.

[6] S. Ghosh and C. Shah, "Towards automatic fake news classification," Proceedings of the Association for Information Science and Technology, vol. 55, no. 1, pp. 805–807, Jan. 2018, doi: 10.1002/pra2.2018.14505501125.

[7] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881–893, 2021, doi: 10.1109/TCSS.2021.3068519.

[8] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, no. Icwsm, pp. 626–637, 2020.

[9] J. Enright and G. Kondrak, "The application of chordal graphs to inferring phylogenetic trees of languages," Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 545–552, 2011.

[10] H. Chen, B. Martin, C. M. Daimon, and S. Maudsley, "Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications," Frontiers in Physiology, vol. 4 JAN, no. January, pp. 1–6, 2013, doi: 10.3389/fphys.2013.00008.

[11] A. Mehler, "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, pp. 496–500, 2006.

[12] L. Araujo, "How evolutionary algorithms are applied to statistical natural language processing," Artificial Intelligence Review, vol. 28, no. 4, pp. 275–303, 2007, doi: 10.1007/s10462-009-9104-y.

[13] A. Mashaghi, "Circuit Topology of Folded Chains," Not. Am. Math. Soc., vol. 68, pp. 420–423, 2021, doi: 10.1090/noti2241.

[14] A. Mashaghi, R. J. Van Wijk, and S. J. Tans, "Circuit topology of proteins and nucleic acids," Structure, vol. 22, no. 9, pp. 1227–1237, 2014, doi: 10.1016/j.str.2014.06.015.

[15] A. Golovnev and A. Mashaghi, "Generalized Circuit Topology of Folded Linear Chains," iScience, vol. 23, no. 9, p. 101492, 2020, doi: 10.1016/j.isci.2020.101492.

[16] B. Scalvini et al., "Topology of Folded Molecular Chains: From Single Biomolecules to Engineered Origami," Trends Chem, vol. 2, no. 7, pp. 609–622, 2020, doi: 10.1016/j.trechm.2020.04.009.

[17] F. Almeida and G. Xexéo, "Word Embeddings: A Survey," no. 1991, 2019.

[18] M. Knoche, F. Lemmerich, R. Popović, and M. Strohmaier, "Identifying biases in politically biased wikis through word embeddings," HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 253–257, 2019, doi: 10.1145/3342220.3343658.

[19] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," Expert Systems with Applications, vol. 128, pp. 201–213, 2019, doi: 10.1016/j.eswa.2019.03.036.

[20] A. Golovnev and A. Mashaghi, "Topological Analysis of Folded Linear Molecular Chains," in Topological Polymer Chemistry, Singapore: Springer Singapore, 2022, pp. 105–114. doi: 10.1007/978-981-16-6807-4_7.

[21] B. D. Horne and S. Adali, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," 2017.

[22] W. Kintsch and J. C. Yarbrough, "Role of rhetorical structure in text comprehension.," Journal of Educational Psychology, vol. 74, no. 6, pp. 828–834, Dec. 1982, doi: 10.1037/0022-0663.74.6.828.

[23] D. Angelov, "Top2Vec: Distributed Representations of Topics," pp. 1–25, 2020.

[24] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018.

[25] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172. doi: 10.1007/978-3-642-37456-2_14.

[26] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, pp. 2931–2937, 2017, doi: 10.18653/v1/d17-1317.

[27] R. Goldstone, "An efficient method for obtaining similarity data," Behavior Research Methods, Instruments, & Computers, vol. 26, no. 4, pp. 381–386, Dec. 1994, doi: 10.3758/BF03204653.

[28] R. Richie, B. White, S. Bhatia, and M. C. Hout, "The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures," Behavior Research Methods, vol. 52, no. 5, pp. 1906–1928, 2020, doi: 10.3758/s13428-020-01362-y.

[29] L. Buchanan, N. R. Brown, R. Cabeza, and C. Maitson, "False memories and semantic lexicon arrangement," Brain and Language, vol. 68, no. 1–2, pp. 172–177, 1999, doi: 10.1006/brln.1999.2072.

[30] W. Jarrold et al., "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," pp. 27–37, 2015, doi: 10.3115/v1/w14-3204.

[31] K. Wang, N. Reimers, and I. Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning," Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, pp. 671–688, 2021, doi: 10.18653/v1/2021.findings-emnlp.59.

[32] I. Honnibal, Matthew Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." 2017.

[33] N. Reimers and I. Gurevych, "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, pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[34] "SBERT.net." https://www.sbert.net/docs/pretrained_models.html