6-5-2023

Spectral analysis perspective of why misinformation containment is still an unsolved problem

Vishnu S. Pendyala
San Jose State University, vishnu.pendyala@sjsu.edu

Foroozan Sadat Akhavan Tabatabaii
San Jose State University

Recommended Citation

Vishnu S. Pendyala and Foroozan Sadat Akhavan Tabatabaii. "Spectral analysis perspective of why misinformation containment is still an unsolved problem" *2023 IEEE Conference on Artificial Intelligence (CAI)* (2023). https://doi.org/10.1109/CAI54212.2023.00099

Spectral analysis perspective of why misinformation containment is still an unsolved problem

Vishnu S. Pendyala
Department of Applied Data Science
San Jose State University
San Jose, CA 95192-0250
Email: vishnu.pendyala@sjsu.edu

Foroozan Sadat Akhavan Tabatabaii
Department of Applied Data Science
San Jose State University
San Jose, CA 95192-0250
Email: foroozan.akhavan@sjsu.edu

***Abstract*—Misinformation is still a major societal problem. The arrival of ChatGPT only added to the problem. This paper analyzes misinformation in the form of text from a spectral analysis perspective to find the answer to why the problem is still unsolved despite multiple years of research and a plethora of solutions in the literature. A variety of embedding techniques are used to represent information for the purpose. The diverse spectral methods used on these embeddings include t-distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA). The analysis shows that misinformation is quite closely intertwined with genuine information and the machine learning algorithms are not as effective in separating the two despite the claims in the literature.**

I. Introduction

Misinformation, fake news, and lies have been adversely impacting the society in a number of significant ways. The advent of generative conversational AI applications such as ChatGPT has only exasperated the problem [1]. Misinformation can be multi-modal. Generative AI is capable of producing misinformation in multiple modalities as well such as by generating images from text [2]. However, while the difficulties in detecting fake images are well documented in the literature [3], the same is not true for misinformation in the form of text. On the other hand, some studies have reported 100% accuracy with detecting AI generated text using simple language models such as BOW [4]. Only time can confirm if AI generated text can indeed be easily detected. There is also abundant literature on solving the misinformation containment problem with human generated text but it is a well known fact that the problem is still largely unsolved [5]. For this study, we do not distinguish between human and AI generated text.

Machine learning, although sometimes impacted by fairness issues [6], has been used to address a wide variety of problems in the quest of societal progress [7], including misinformation containment [8]. Given the current gap in the literature in sufficiently identifying the reasons why misinformation containment is still an unsolved problem, there is a need to focus on why even the now ubiquitous machine learning is unable to solve the problem. This work is an attempt to determine what makes it so difficult to identify misinformation and the limitations of the current Natural Language Processing (NLP) and Machine Learning (ML) techniques in doing so using spectral analysis.

II. Related Work

Misinformation containment is proven in the literature to be NP-hard [9]. Misinformation detection can be addressed using diverse approaches, including algorithms such as the Kalman Filter [10], statistical techniques, and first order logic [11]. However, it is established in the literature that machine learning is a good alternative to heuristic algorithms to solve NP-hard problems [12]. Another view of misinformation is that it is a major contributor to uncertainty in the world, which in turn significantly contributes to the carbon footprint. Machine learning helps in reducing the carbon footprint [13], including that from the uncertainty caused by misinformation. A further literature survey naturally shows a comprehensive use of machine learning and deep learning in conjunction with NLP techniques to address the problem.

In previous work, the first author of this paper surveyed the various approaches to securing trust specifically in online social networks [14], proposed some ideas [15] and explored some of the reasons why the problem of misinformation containment is unsolved despite the exhaustive research and a number of solutions [5]. The contribution of this work is in uniquely explaining the reasons based on a spectral analysis of the genuine and misinformation in the dataset. A thorough quantitative analysis of the determinants found that homogeneity of the communities in terms of their information consumption pattern is the primary driver for misinformation spread [16]. Other than that, to the best of our knowledge, this is the first work that attempts to explain using spectral analysis, why misinformation containment is mostly an unsolved problem.

III. Dataset and Preprocessing

The LIAR dataset [17] contains around 12,800 short statements collected from various sources such as political debates, Facebook posts, news releases, and tweets that are labeled manually. There are six fine-grained labels for each of the 12,800 statements: true, mostly-true, half-true, barely-true, false, and pants-fire, indicating the degrees of truthfulness. For the purposes of this project, the labels were encoded into three categories, the first class consists of all true labels, the second class contains false, and the third class contains pants-fire statements. After encoding the labels, the dataset

has 1047 'pants-fire' statements, 2507 'false' statements, and 9237 'true' statements. After splitting the data into train, test and validation sets, the true and false statements in the train set downsampled to 839 statements in each of the three classes in order to get a balanced dataset. Since the language models such as BERT are pre-trained, not having a huge dataset is not an issue. Usual preprocessing of the dataset was done to make it ready for the classification task. For instance, to rule out any non-English occurrences in the statements, all statements containing non-ASCII characters were removed.

IV. Methodology

The statements in the dataset are converted into vectors using embeddings techniques based on BERT [18], S-BERT [19], which build upon the transformer architecture [20] and also on the Doc2VecC framework [21]. The embeddings project the statements into the latent feature vector space. A spectral analysis of the statement vectors is then performed using two substantially different techniques - the t-SNE [22] and PCA [23] algorithms depicting the actual labels. PCA is a linear approach, while t-SNE is a non-linear visualization technique. Hence the two are fundamentally different in their approach. The diversity in the representation learning, classification algorithms, and spectral techniques is to ensure that as much as possible, there is no bias in our conclusions that is attributable to the technique.

The dimensionality is reduced to two, so that the latent feature space can be visualized better. The analysis shows that the data is highly non-linear. The statements are then classified using two machine learning algorithms, Support Vector Machines (SVM) using the Radial Basis Function (RBF) kernel and K-Nearest Neighbors (KNN), which are known to perform well on non-linear data. These classification algorithms work in fundamentally different ways as well. The hyperparameters such as the kernel function used for SVM and the value of k for KNN are determined through the process of validation, trying out several alternatives and choosing the best ones. PCA and t-SNE are run again on the data, but this time, depicting the predicted labels. A number of useful conclusions can be drawn from the analysis as detailed in the following sections.

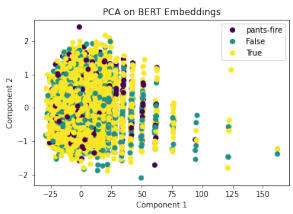
V. Experiments

A. Statement Embeddings using BERT

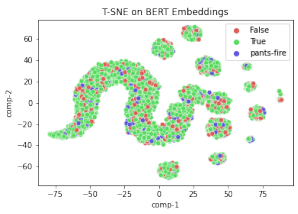
The pre-trained model called bert-base-uncased from the Huggingface open-source library is used to vectorize the statements. The dataset is downsampled to make it balanced. The final generated vectors are present in a 2D array with 2517 rows and 768 embedding dimensions. The results of PCA and t-SNE are visualized in Fig. 1a and Fig. 1b.

B. Statement Embeddings Using Doc2VecC

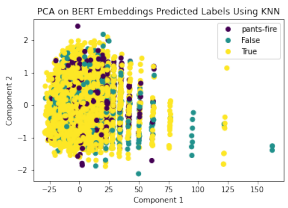
The Doc2VecC framework [21] represents a document as an average of the word embeddings of randomly sampled words in that document. Word embeddings generated by the Doc2vecC with respect to the context are reported to be significantly better than those generated by Word2Vec [21]. The Doc2vecC framework represents a document as an average of the word embeddings of randomly sampled words in that document. An additional corruption model is included in the algorithm that gives more importance to informative words and suppresses common words by using data-dependent regularization. The original Doc2VecC algorithm produces a vector consisting of 100 dimensions. For this project, it is changed to 256 dimensions to capture more features from each statement. The visualizations using PCA and t-SNE are shown in Fig. 2a and Fig. 2b.



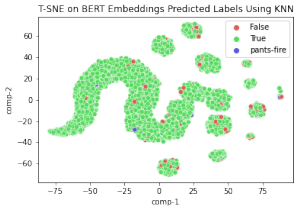
(a) Applying PCA on BERT Embeddings



(b) Applying t-SNE on BERT Embeddings

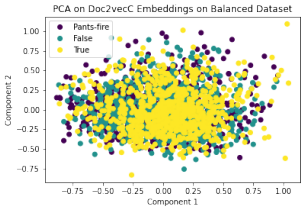


(c) Applying PCA on BERT Embeddings Predicted Labels Using KNN

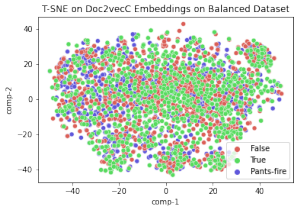


(d) Applying t-SNE on BERT Embeddings Predicted Labels Using KNN

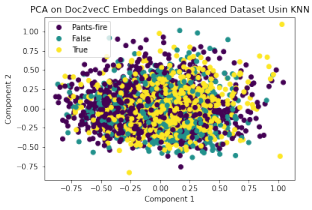
Fig. 1: Applying PCA and t-SNE on BERT Embeddings Using Mean of the Word Tokens



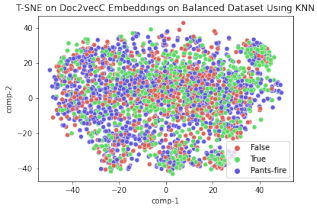
(a) Applying PCA on Doc2vecC Embeddings



(b) Applying t-SNE on Doc2vecC Embeddings

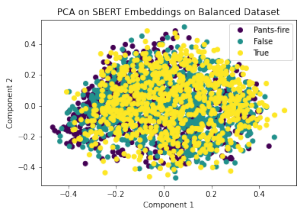


(c) Applying PCA on Doc2vecC Embeddings Predicted Labels Using KNN

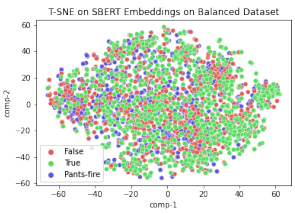


(d) Applying t-SNE on Doc2vecC Embeddings Predicted Labels Using KNN

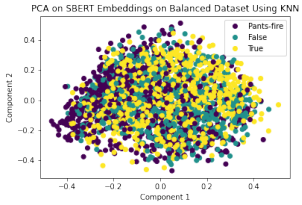
Fig. 2: Applying PCA and t-SNE on Doc2vecC Embeddings



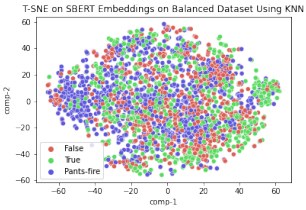
(a) Applying PCA on S-BERT Embeddings



(b) Applying t-SNE on S-BERT Embeddings

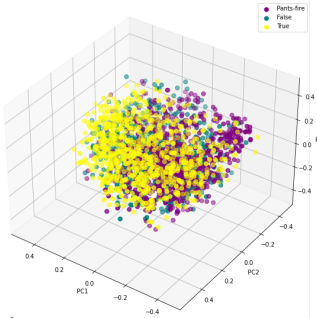


(c) Applying PCA on S-BERT Embeddings Predicted Labels Using KNN

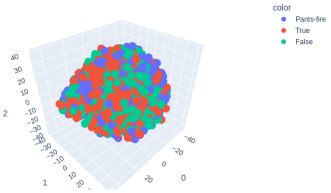


(d) Applying t-SNE on S-BERT Embeddings Predicted Labels Using KNN

Fig. 3: Applying PCA and t-SNE on S-BERT Embeddings



(a) Applying Three Dimensional PCA on S-BERT Embeddings predicted Labels Using SVM



(b) Applying Three Dimensional t-SNE on S-BERT Embeddings predicted Labels Using SVM

Fig. 4: Applying PCA and t-SNE on S-BERT Embeddings

C. Statement Embeddings Using Sentence-BERT

Sentence-BERT (SBERT) [19] is a modification of the pre-trained BERT algorithm. SBERT embeds sentences faster and more accurately in comparison to BERT and its optimized variant, RoBERTa. This model maps sentences and paragraphs to a 384-dimensional vector and can be used for tasks like clustering or semantic search. To vectorize the statements of the LIAR dataset, a pre-trained model "all-MiniLM-L6-v2" was applied. The result of the spectral analysis are shown in Fig. 3a and Fig. 3b.

D. Classification of the statements

Since this is a multi-class problem, the decision function for the SVM algorithm has been chosen to be "ovo," which stands for "one-vs-one". Based on our experiments during validation, for the KNN classifier, six nearest neighbors were considered (k=6), using Minkowski distance metric, uniform weights, and kd-tree structure.

VI. Results

The performance results of classifications using SVM and KNN models for BERT, Doc2VecC and SBERT embeddings are shown in Table I. In addition to the usual metrics such as accuracy, precision, recall, and F1-score, since the dataset was originally not balanced, we also computed the Cohen's Kappa score. For brevity, experiments on the imbalanced dataset are not included in this paper. The ROC curves are also not shown for the same brevity reasons. Instead the area under the ROC curves is tabulated in Table I.

VII. Discussion

As shown in the spectral analysis with dimensionality reduced to the two most important latent features in Figures 1(c), 1(d), 2(c) 2(d), and 3(c), 3(d) and three latent features in Fig. 4a and Fig. 4b, there are no natural discernible clusters of the various classes. Each class of statements is spread across the feature space and the classes are not significantly distinguishable. None of the diverse embedding schemes used place the data points in natural clusters. Instead, they place the statement vectors from different classes next to or even coinciding with each other so much so that even classifiers like SVM with RBF kernel and K-NN that can classify highly non-linear data also cannot perform well. The numbers in Table I are indicative of this fact.

It must be noted that the visualizations are highly scaled down versions of the original feature space but are quite representative of the relative positions of the feature vectors. Because of the compacted feature space, many feature vectors overlap with each other making it appear as if there are clusters in some combinations of the embedding scheme and spectral method. Adding a dimension and visualizing the embeddings in 3D as in Fig. 4a and Fig. 4b does not help much. However, correlating such figures with the other figures, the dataset characteristics, and also the accuracy metrics in Table I, it can be concluded that in terms of the embeddings computed, there is not a substantial difference in the embeddings for the true and other classes of textual information.

This is true across the diverse embedding schemes that we used. The plots, which are of different shapes and distributions, indicate that the embeddings generated by the three frameworks are substantially different. Nevertheless, irrespective of the embedding scheme, the spectral analysis using two diverse techniques confirms our conclusion that the current representation learning approaches are unable to capture the varying degrees of truthfulness in textual information. Machine learning is primarily learning by similarity [24] [25]. Learning happens by discovering similarities and dissimilarities among data. In the case of misinformation detection, the representation learning is unable to discern among similarities and dissimilarities of the varying degrees of truth in the textual information.

VIII. Conclusion

The complexity of classifying truth from lies is explained in this work in an attempt to answer why the problem of mis-

TABLE I: Results of Applying Different Models on Balanced Dataset with Multiclass Classification

models	Accuracy	Training Accuracy	Kappa Score	F-1	Precision	Recall	ROC Accuracy
BERT+SVM	36.87%	47.82%	05.31%	51.96%	60.26%	47.82%	56.24%
BERT+KNN	32.56%	55.71%	33.57%	37.65%	60.15%	32.56%	53.32%
Doc2VecC+SVM	30.27%	68.10%	52.15%	33.17%	56.40%	30.27%	45.38%
Doc2VecC+KNN	44.66%	55.47%	33.21%	48.37%	53.31%	44.66%	44.63%
SBERT+SVM	45.61%	89.40%	84.10%	51.17%	65.87%	45.61%	63.43%
SBERT+KNN	39.13%	60.65%	40.97%	45.01%	63.87%	39.13%	59.70%

information containment is unsolved. The work used multiple schemes of embeddings, spectral analysis, multiple supervised learning algorithms, and evaluation metrics to explain the reasons. Current representation learning does not significantly distinguish between true, false, and the "pants on fire" kind of information with varying degrees of truthfulness. In future, we plan to analyze the reasons from various perspectives and propose better solutions at performing the classification. We also plan to use more evaluation metrics, more spectral analysis methods, and do similar analysis of the text produced by generative AI as well. Another possible research direction is to invent a new embedding scheme for text that is capable of capturing the veracity of its information content.

REFERENCES

[1] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1567, 2023.

[2] Vishnu S Pendyala and VigneshKumar Thangarajan. Reconnoitering generative deep learning through image generation from text. In *Deep Learning Research Applications for Natural Language Processing*, pages 113–131. IGI Global, 2023.

[3] Lilei Zheng, Ying Zhang, and Vrizlynn LL Thing. A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58:380–399, 2019.

[4] Ilker Cingillioglu. Detecting ai-generated essays: the chatgpt challenge. *The International Journal of Information and Learning Technology*, 2023.

[5] Vishnu S Pendyala. Misinformation containment using nlp and machine learning: Why the problem is still unsolved. In *Deep Learning Research Applications for Natural Language Processing*, pages 41–56. IGI Global, 2023.

[6] Vishnu S Pendyala and HyungKyun Kim. Analyzing and addressing data-driven fairness issues in machine learning models used for societal problems. In *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–7. IEEE, 2023.

[7] Vishnu S Pendyala. *Machine Learning for Societal Improvement, Modernization, and Progress*. IGI Global, 2022.

[8] Vishnu S Pendyala and Silvia Figueira. Towards a truthful world wide web from a humanitarian perspective. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 137–143. IEEE, 2015.

[9] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674, 2011.

[10] Vishnu S Pendyala, Yuhong Liu, and Silvia M Figueira. A framework for detecting injected influence attacks on microblog websites using change detection techniques. *Development Engineering*, 3:218–233, 2018.

[11] Vishnu Pendyala. Veracity of big data. *Machine Learning and Other Approaches to Verifying Truthfulness*, 2018.

[12] Guoliang Li, Xuanhe Zhou, and Lei Cao. Machine learning for databases. In *The First International Conference on AI-ML-Systems*, pages 1–2, 2021.

[13] Vishnu S Pendyala and Saritha Podali. Machine learning for ecological sustainability: An overview of carbon footprint mitigation strategies. *Machine Learning for Societal Improvement, Modernization, and Progress*, pages 1–26, 2022.

[14] Vishnu S Pendyala. Securing trust in online social networks. In *Secure Knowledge Management In Artificial Intelligence Era: 8th International Conference, SKM 2019, Goa, India, December 21–22, 2019, Proceedings 8*, pages 194–201. Springer, 2020.

[15] Vishnu S Pendyala. Evolving a truthful humanitarian world wide web. 2018.

[16] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[17] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] Minmin Chen. Efficient vector representation for documents through corruption. In *International Conference on Learning Representations*, 2017.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[23] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[24] Vishnu S Pendyala. Relating machine learning to the real-world: analogies to enhance learning comprehension. In *Soft Computing and its Engineering Applications: Third International Conference, icSoftComp 2021, Changa, Anand, India, December 10–11, 2021, Revised Selected Papers*, pages 127–139. Springer, 2022.

[25] Vishnu Pendyala and Rakesh Amireddy. Enhancing the cognition and efficacy of machine learning through similarity. *SN Computer Science*, 3(6):442, 2022.