Does the Geometry of Word Embeddings Help Document Classification? A Case Study on Persistent Homology Based Representations

Paul Michel* Carnegie Mellon University pmichel1@cs.cmu.edu

Abhilasha Ravichander* Carnegie Mellon University aravicha@cs.cmu.edu

Shruti Rijhwani* Carnegie Mellon University srijhwan@cs.cmu.edu

Abstract

We investigate the pertinence of methods from algebraic topology for text data analysis. These methods enable the development of mathematically-principled isometric-invariant mappings from a set of vectors to a document embedding, which is stable with respect to the geometry of the document in the selected metric space. In this work, we evaluate the utility of these topology-based document representations in traditional NLP tasks, specifically document clustering and sentiment classification. We find that the embeddings do not benefit text analysis. In fact, performance is worse than simple techniques like *tf-idf*, indicating that the geometry of the document does not provide enough variability for classification on the basis of topic or sentiment in the chosen datasets.

1 Introduction

Given a embedding model mapping words to n dimensional vectors, every document can be represented as a finite subset of $\mathbb{R}^n$. Comparing documents then amounts to comparing such subsets. While previous work shows that the Earth Mover's Distance (Kusner et al., 2015) or distance between the weighted average of word vectors (Arora et al., 2017) provides information that is useful for classification tasks, we wish to go a step further and investigate whether useful information can also be found in the 'shape' of a document in word embedding space.

Persistent homology is a tool from algebraic topology used to compute topological signatures (called *persistence diagrams*) on compact metric spaces. These have the property of being stable with respect to the Gromov-Haussdorff distance (Gromov et al., 1981). In other words, compact metric spaces that are close, up to an isometry, will have similar embeddings. In this work, we examine the utility of such embeddings in text classification tasks. To the best of our knowledge, no previous work has been performed on using topological representations for traditional NLP tasks, nor has any comparison been made with state-of-the-art approaches.

We begin by considering a document as the set of its word vectors, generated with a pretrained word embedding model. These form the metric space on which we build persistence diagrams, using Euclidean distance as the distance measure. The diagrams are a representation of the document's geometry in the metric space. We then perform clustering on the Twenty Newsgroups dataset with the features extracted from the persistence diagram. We also evaluate the method on sentiment classification tasks, using the Cornell Sentence Polarity (CSP) (Pang and Lee, 2005) and IMDb movie review datasets (Maas et al., 2011).

As suggested by Zhu (2013), we posit that the information about the intrinsic geometry of documents, found in the persistence diagrams, might yield information that our classifier can leverage, either on its own or in combination with other representations. The primary objective of our work is to empirically evaluate these representations in the case of sentiment and topic classification, and assess their usefulness for real-world tasks.

2 Method

2.1 Word embeddings

As a first step we compute word vectors for each document in our corpus using a word2vec (Mikolov et al., 2013) model trained on the Google

*The indicated authors contributed equally to this work.

Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 235–240,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

News dataset[1]. In addition to being a widely used word embedding technique, word2vec has been known to exhibit interesting linear properties with respect to analogies (Mikolov et al., 2013), which hints at rich semantic structure.

2.2 Gromov-Haussdorff Distance

Given a dictionary of word vectors of dimension n, we can represent any document as a finite subset of $\mathbb{R}^n$. The *Haussdorff distance* gives us a way to evaluate the distance between two such sets. More precisely, the Haussdorff distance d_H between two finite subsets A, B of $\mathbb{R}^n$ is defined as:

$$d_H(A, B) = \max(\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A))$$

where $d(x, Y) = \inf_{y \in Y} \|x - y\|_2$ is the distance of point x from set Y.

However, this distance is sensitive to translations and other isometric[2] transformations. Hence, a more natural metric is the **Gromov-Haussdorff distance** (Gromov et al., 1981), simply defined as

$$d_{GH}(A, B) = \inf_{f \in E_n} d_H(A, f(B))$$

where E_n is the set of all isometries of $\mathbb{R}^n$.

Figure 1 provides an example of practical Gromov-Haussdorff (GH) distance computation between two sets of three points each. Both sets are embedded in $\mathbb{R}^2$ (middle panel) using isometries i.e the distance between points in each set is conserved. The Haussdorff distance between the two embedded sets corresponds to the length of the black segment. The GH distance is the minimum Haussdorff distance under all possible isometric embeddings.

We want to compare documents based on their intrinsic geometric properties. Intuitively, the GH distance measures how far two sets are from being isometric. This allows us to define the geometry of a document more precisely:

Definition 1 (Document Geometry) *We say that two documents A, B have the same* geometry *if* $d_{GH}(A, B) = 0$*, ie if they are the same up to an isometry.*

Mathematically speaking, this amounts to defining the geometry of a document as its equivalence class under the equivalence relation induced by the GH distance on the set of all documents.

[1] https://code.google.com/archive/p/word2vec/

[2] $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is *isometric* if it is distance preserving, ie $\forall x, y \in \mathbb{R}^n, \|f(x) - f(y)\|_2 = \|x - y\|_2$. Rotations, translations and reflections are examples of (linear) isometries.

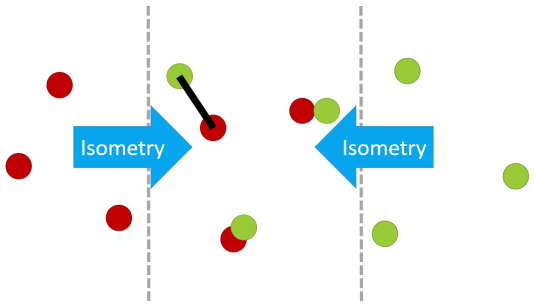

Figure 1: Gromov-Haussdorff distance between two sets (red, green). The black bar represents the actual distance (given that the isometric embedding is optimal).

Comparison to the Earth Mover Distance : Kusner et al. (2015) proposed a new method for computing a distance between documents based on an instance of the Earth Mover Distance (Rubner et al., 1998) called Word Mover Distance (WMD). While WMD quantifies the total cost of matching all words of one document to another, the GH distance is the cost, up to an isometry, of the worst-case matching.

2.3 Persistence diagrams

Efficiently computing the GH distance is still an open problem despite a lot of recent work in this area (Mémoli and Sapiro, 2005; Bronstein et al., 2006; Mémoli, 2007; Agarwal et al., 2015).

Fortunately, Carrière et al. (2015) provides us with a way to derive a signature which is stable with respect to the GH distance. More specifically, given a finite point cloud $A \subset \mathbb{R}^n$, the persistence diagram of the Vietori-Rips filtration on A, $Dg(A)$, can be computed. This approach is inspired by persistent homology, a subfield of algebraic topology.

The rigorous definition of these notions is not the crux of this paper and we will only present them informally. The curious reader is invited to refer to Zhu (2013) for a short introduction. More details are in Delfinado and Edelsbrunner (1995); Edelsbrunner et al. (2002); Robins (1999).

A persistence diagram is a scatter plot of 2-D points representing the appearance and disappearance of geometric features[3] under varying resolutions. This can be imagined as replacing each point by a sphere of increasing radius.

We use the procedure described in Carrière et al.

[3] such as connected components, holes or empty hulls

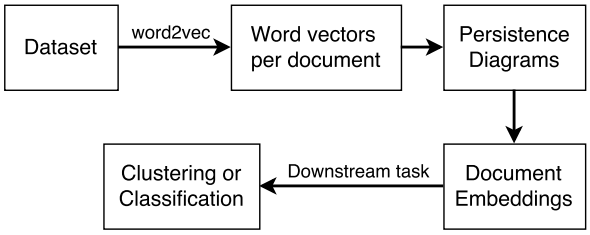

Figure 2: Method Pipeline

(2015) to derive fixed-sized vectors from persistence diagrams. These vectors have the following property: if A and B are two finite subsets of $\mathbb{R}^n$, $Dg(A)$ and $Dg(B)$ are their persistence diagrams, $N = \max(|Dg(A)|, |Dg(B)|)$ and $V_A, V_B \in \mathbb{R}^{\frac{N(N-1)}{2}}$, then

$$\|V_A - V_B\|_2 \leqslant \sqrt{2N(N-1)} d_{GH}(A, B)$$

In other words, the resulting signatures V_A and V_B are stable with respect to the GH distance. The size of the vectors are dependent on the underlying sets A and B. However, as is argued in Carrière et al. (2015), we can truncate the vectors to a dimension fixed across our dataset while preserving the stability property (albeit losing some of the representative ability of the signatures).

3 Experiments

3.1 Experiments

The pipeline for our experiments is shown in Figure 2. In order to build a persistence diagram, we convert each document to the set of its word vectors. We then use Dionysus (Morozov, 2008–2016), a C++ library for computing persistence diagrams, and form the signatures described in 2.3. We will subsequently refer to these diagrams as Persistent Homology (PH) embeddings. Once we have the embeddings for each document, they can be used as input to standard clustering or classification algorithms.

As a baseline document representation, we use the average of the word vectors for that document (subsequently called AW2V embeddings).

For clustering, we experiment with K-means and Gaussian Mixture Models (GMM) on a subset[4] of the Twenty Newsgroups dataset. The subset was selected to ensure that most documents are from related topics, making clustering non-trivial, and the documents are of reasonable length to compute the representation.

[4]alt.atheism, sci.space and talk.religion.misc categories

For classification, we perform both sentence-level and document-level binary sentiment classification using logistic regression on the CSP and IMDb corpora respectively.

4 Results

4.1 Hyper-parameters

Our method depends on very few hyper-parameters. Our main choices are listed below.

Choice of distance We experimented with both euclidean distance and cosine similarity (angular distance). After preliminary experiments, we determined that both performed equally and hence, we only report results with the euclidean distance.

Persistence diagram computation The hyper-parameters of the diagram computation are monotonic and mostly control the degree of approximation. We set them to the highest values that allowed our experiment to run in reasonable time[5].

4.2 Document Clustering

We perform clustering experiments with the baseline document features (AW2V), *tf-idf* and our PH signatures. Figure 3 shows the B-Cubed precision, recall and F1-Score of each method (metrics as defined in Amigó et al. (2009)). To further assess the utility of PH embeddings, we concatenate them with AW2V to obtain a third representation, AW2V+PH.

With GMM and AW2V+PH, the F1-Score of clustering is 0.499. In terms of F1 and precision, we see that *tf-idf* representations perform better than PH, for reasons that we will discuss in later sections. In terms of recall, PH as well as AW2V perform fairly well. Importantly, we see that all the metrics for PH are significantly above the random baseline, indicating that some valuable information is contained in them.

4.3 Sentiment Classification

4.3.1 Sentence-Level Sentiment Analysis

We evaluate our method on the CSP dataset[6]. The results are presented in Table 1. For comparison, we provide results for one of the state of the art models, a CNN-based sentence classifier (Kim,

[5]Selected such that the computation of the diagram of the longest file in the training data took less than 10 minutes.

[6]For lack of a canonical split, we use a random 10% of the dataset as a test set

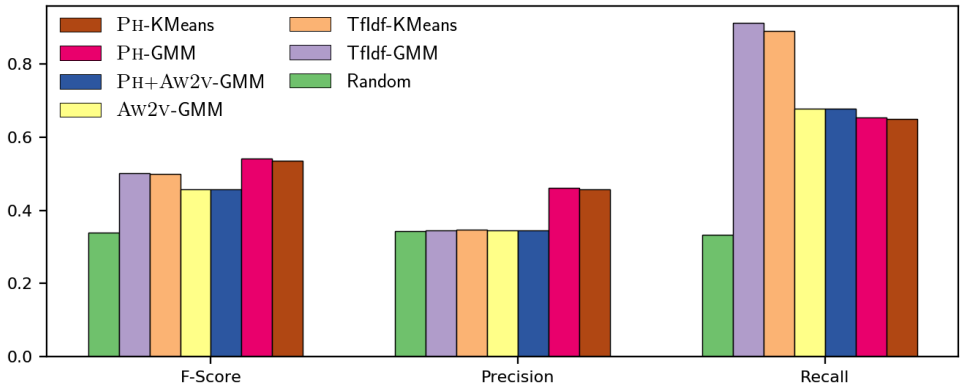

Figure 3: Results for clustering on 3 subclasses of the Twenty Newsgroups dataset

Model	Accuracy
CNN Non-Static	81.5%
PH + LogReg	53.19%
AW2V + LogReg	77.13%
AW2V + PH + LogReg	77.13%

Table 1: Performance on the CSP dataset

Model	Accuracy
Paragraph Vector	92.58%
PH + LogReg	53.16%
AW2V + LogReg	82.94%
AW2V + PH + LogReg	83.08%

Table 2: Performance on the IMDb dataset

2014). We observe that by themselves, PH embeddings are not useful at predicting the sentiment of each sentence. AW2V gives reasonable performance in this task, but combining the two representations does not impact the accuracy at all.

4.3.2 Document-Level Sentiment Analysis

We perform document-level binary sentiment classification on the IMDb Movie Reviews Dataset (Maas et al., 2011). We use sentence vectors in this experiment, each of which is the average of the word vectors in that sentence. The results are presented in Table 2. We compare our results with the paragraph-vector approach (Le and Mikolov, 2014). We observe that PH embeddings perform poorly on this dataset. Similar to the CSP dataset, AW2V embeddings give acceptable results. The combined representation performs slightly better, but not by a margin of significance.

5 Discussion and Analysis

As seen in Figure 3, the PH representation does not outperform *tf-idf* or AW2V, and in fact often doesn't perform much better than chance.

One possible reason is linked to the nature of our datasets: the computation of the persistence diagram is very sensitive to the size of the documents. The geometry of small documents, where the number of words is negligible with respect to the dimensionality of the word vectors, is not very rich. The resulting topological signatures are very sparse, which is a problem for CSP as well as documents in IMDb and Twenty Newsgroups that contain only one line. On the opposite side of the spectrum, persistence diagrams are intractable to compute without down-sampling for very long documents (which in turn negatively impacts the representation of smaller documents).

We performed an additional experiment on a subset of the IMDb corpus that only contained documents of reasonable length, but obtained similar results. This indicates that the poor performance of PH representations, even when combined with other features (AW2V), cannot be explained only by limitations of the data.

These observations lead to the conclusion that, for these datasets, the intrinsic geometry of documents in the word2vec semantic space does not help text classification tasks.

6 Related Work

Learning distributed representations of sentences or documents for downstream classification and information retrieval tasks has received recent attention owing to their utility in several applications, be it representations trained on the sen-

tence/paragraph level Le and Mikolov (2014); Kiros et al. (2015) or purely word vector based methods Arora et al. (2017).

Document classification and clustering (Willett, 1988; Hotho et al., 2005; Steinbach et al., 2000; Huang, 2008; Xu and Gong, 2004; Kuang et al., 2015; Miller et al., 2016) and sentiment classification (Nakagawa et al., 2010; Kim, 2014; Wang and Manning, 2012) are relatively well studied.

Topological data analysis has been used for various tasks such as 3D shapes classification (Chazal et al., 2009) or protein structure analysis (Xia and Wei, 2014). However, such techniques have not been used in NLP, primarily because the theory is inaccessible and suitable applications are scarce. Zhu (2013) offers an introduction to using persistent homology in NLP, by creating representations of nursery-rhymes and novels, as well as highlights structural differences between child and adolescent writing. However, these techniques have not been applied to core NLP tasks.

7 Conclusion

Based on our experiments, using persistence diagrams for text representation does not seem to positively contribute to document clustering and sentiment classification tasks. There are certainly merits to the method, specifically its strong mathematical foundation and its domain-independent, unsupervised nature. Theoretically, algebraic topology has the ability to capture structural context, and this could potentially benefit syntax-based NLP tasks such as parsing. We plan to investigate this connection in the future.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O) under the Low Resource Languages for Emergent Incidents (LORELEI) program issued by DARPA/I2O under Contract No. HR0011-15-C-0114. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

We are grateful to Matt Gormley, Hyun Ah Song, Shivani Poddar and Hai Pham for their suggestions on the writing of this paper as well as to Steve Oudot for pointing us to helpful references. We would also like to thank the anonymous ACL reviewers for their valuable suggestions.

References

Pankaj K Agarwal, Kyle Fox, Abhinandan Nath, Anastasios Sidiropoulos, and Yusu Wang. 2015. Computing the gromov-hausdorff distance for metric trees. In *International Symposium on Algorithms and Computation*. Springer, pages 529–540.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval* 12(4):461–486.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations. To Appear*.

Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. 2006. Efficient computation of isometry-invariant distances between surfaces. *SIAM Journal on Scientific Computing* 28(5):1812–1836.

Mathieu Carrière, Steve Y Oudot, and Maks Ovsjanikov. 2015. Stable topological signatures for points on 3d shapes. In *Computer Graphics Forum*. Wiley Online Library, volume 34, pages 1–12.

Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. 2009. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*. Wiley Online Library, volume 28, pages 1393–1403.

Cecil Jose A Delfinado and Herbert Edelsbrunner. 1995. An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere. *Computer Aided Geometric Design* 12(7):771–784.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. 2002. Topological persistence and simplification. *Discrete and Computational Geometry* 28(4):511–533.

Mikhael Gromov, Jacques Lafontaine, and Pierre Pansu. 1981. Structures métriques pour les variétés riemanniennes .

Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. 2005. A brief survey of text mining. In *Ldv Forum*.

Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. pages 49–56.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *In EMNLP*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. pages 3294–3302.

Da Kuang, Jaegul Choo, and Haesun Park. 2015. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, Springer, pages 215–243.

Matt J Kusner, Yu Sun, Nicholas I Kolkin, Kilian Q Weinberger, et al. 2015. From word embeddings to document distances. In *ICML*. volume 15, pages 957–966.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. pages 1188–1196.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 142–150. http://www.aclweb.org/anthology/P11-1015.

Facundo Mémoli. 2007. On the use of gromov-hausdorff distances for shape comparison .

Facundo Mémoli and Guillermo Sapiro. 2005. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics* 5(3):313–347.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Timothy A Miller, Dmitriy Dligach, and Guergana K Savova. 2016. Unsupervised document classification with informed topic models. *ACL* .

Dmitriy Morozov. 2008–2016. Dyonisus : a c++ library for computing persistent homology. `http://mrzv.org/software/dionysus/`.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '10, pages 786–794. http://dl.acm.org/citation.cfm?id=1857999.1858119.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Vanessa Robins. 1999. Towards computing homology from finite approximations. In *Topology proceedings*. volume 24, pages 503–532.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*. IEEE, pages 59–66.

Michael Steinbach, George Karypis, Vipin Kumar, et al. 2000. A comparison of document clustering techniques. In *KDD workshop on text mining*.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 90–94. http://dl.acm.org/citation.cfm?id=2390665.2390688.

Peter Willett. 1988. Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management* 24(5):577–597.

Kelin Xia and Guo-Wei Wei. 2014. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering* 30(8):814–844.

Wei Xu and Yihong Gong. 2004. Document clustering by concept factorization. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '04, pages 202–209. https://doi.org/10.1145/1008992.1009029.

Xiaojin Zhu. 2013. Persistent homology: An introduction and a new text representation for natural language processing. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.