Topological Data Analysis In Text Classification Based On Word Embedding And TF-IDF

Xiaoyang Wen
The Experimental High School Attached To Beijing Normal University, Beijing, China
Corresponding author’s address: wenxiaoyang2003@sina.cn

Abstract. As a fresh and rapidly-developing method in data science, topological data analysis (TDA) offers a new set of ways to look at data and derive features out of high-dimensional models with topological and geometric tools. In this paper, the author briefly introduces the topological concepts that are involved several researches, then compares and examines different methods of extraction of topological features from the texts. The result shows that these topological tools provide some additional features of the document that are not detected by using the original methods. In the experiment, adding these topological features to the usual text mining tools results in improvement of prediction accuracy (as much as 5%). However, as expected, these topological features alone are not sufficient to classify text documents. Future experiments and discussions need to be conducted to determine whether these methods could be combined to make better classifications.

1. Introduction
The primary goal in topological data analysis (TDA) is to analyze the shapes in data. While TDA has got significant attention in data mining for numeric data[1], its application in natural language processing still appears to be challenging. Defining shapes in the text seems much more challenging even though vector spaces are used as standard tools to define geometries in text mining and information retrieval, and conceptual spaces are relevant for cognitive modeling and semantics of natural language.

In this paper, the author mainly studies the research method in the following papers: Topological Data Analysis In Text Classification: Extracting Features With Additive Information[2] and An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis[3]. And the focus is on introducing and examining two methods to apply TDA to text classification. Term frequency (or TF-IDF) and word embeddings are the most frequently used methods to translate the text into numerical data. Therefore, they deserve to be examined, as a priority, for potential to reveal their hidden dimensions by applying topological methods.

2. Overview

2.1. Simplicial Complexes
Simplicial complexes can be seen as higher dimensional generalization of graphs. It is a topological object that is bonded by a set of simplexes such as points, line segments, triangles, and higher dimensional counterparts. Given a set X = {x0, ⋯, xk} ⊂ Rd of k+1 independent points, the k-dimensional simplex σ = [x0, ⋯, xk] spanned by X is the convex hull of X. The points of X are called
the vertices of \( \sigma \) and the simplices spanned by the subsets of \( X \) are called the faces of \( \sigma \). A geometric simplicial complex \( K \) in \( \mathbb{R}^d \) is a collection of simplices such that:

i) any face of a simplex of \( K \) is a simplex of \( K \),

ii) the intersection of any two simplices of \( K \) is either empty or a common face of both[4].

Figure 1. Examples of simplexes

2.2. Betti Numbers and Homology

In mathematics, homology is a general way of associating a sequence of algebraic objects such as abelian groups or modules to other mathematical objects such as topological spaces. Homology groups were originally defined in algebraic topology. There are different types of homology in topology such as simplicial homology, Group homology, Singular homology.

Definition (Betti numbers): Given \( n \in \mathbb{Z}^+ \), the \( n \)-th Betti number of a simplicial complex \( K \) is denoted by \( \beta_n(K) \), and is defined as \( \beta_n(K) = \dim (H_n(K, F)) \).[4]

2.3. Persistent Homology

Persistent homology is an efficient tool which is widely used on studying and analyzing data sets. It tries to find and track homological groups and holes with the help of filtration. A visual representation of persistent homology is the persistent diagram.

When dealing with a large number of discrete data points with multiple features in a high dimensional space, one can set a radius for each point so that the points within each others’ radius would connect together to form geometric shapes. Then one computes the number of holes or loops in the resulting simplicial complex. If the radius of the ball is increased, of which the points are at the centre, more points would be connected, with holes forming and then disappearing. By using a persistent diagram, one could record the changes in the topological features, namely components, loops, and holes, as represented by different Betti numbers. Here, the author summarized some of the common objects and their Betti numbers, which represents their topological features.

Table 1. Betti numbers for some topological shapes

| Order | Type      | A Point | Circle | Sphere | Torus |
|-------|-----------|---------|--------|--------|-------|
| \( \beta_0 \) | components | 1       | 1      | 1      | 1     |
| \( \beta_1 \) | loops     | 0       | 1      | 0      | 2     |
| \( \beta_2 \) | voids     | 0       | 0      | 1      | 1     |
| \( \beta_3 \) | 3D holes  | 0       | 0      | 0      | 0     |

3. Methodology

3.1. Topology Features in Word Embeddings

The author first analyzes the topological features in word embeddings. Assume that a \( T \)-token documents is represented in a space of \( D \) dimensions, the matrix for the word embedding is \( \Psi \in \mathbb{R}^{T \times D} \), which can be treated as a \( D \)-dimensional time series, which has a length of \( T \). Firstly, researchers smooth the time series dimension by applying the following equation to each column:

\[
\Psi'_1 = 1/8\Psi_{t3} + 1/4\Psi_{t2} + 1/2\Psi_{t1} + \Psi_{t} + 1/2\Psi_{t+1} + 1/4\Psi_{t+2} + 1/8\Psi_{t+3}
\]  

Next, the researchers calculated the distance between the smoothed version of different embedding dimensions by using the following equation:
Thus, researchers create a matrix $\Theta_{D \times D} = [\theta(\Psi'(i), \Psi'(j))]$ which can be interpreted as an adjacency matrix of a graph. They then apply persistent homology on the matrix to get the Betti numbers $\beta_0$ and $\beta_1$, which corresponds to the components and loops of a series of high dimensional data. Then for each embedding dimension, they exclude the corresponding vertex of the graph and measure the change in persistence diagrams. And by using the Wasserstein distance to measure the minimum cost of moving the dots in the first diagram to convert it into the second diagram, they effectively represent the importance of each particular dimension in a document.

### 3.2. Topological Features From Term Frequency Space

To apply persistent homology on TF-IDF space, researchers divide the textual document to a fixed number of blocks and then search for repetitive patterns in the text. There is no required number of blocks to split a given document into, but it is noted that generally using a large number could make TF-IDF vectors too sparse, and then comparing them would not be useful. An example of which would be splitting a word document of 100 tokens into 50 blocks, which makes each block containing only 2 elements, and most of the blocks would have zero similarity.

In the experiment, researchers work on 10 vertices, which are represented by their TF-IDF vectors, as shown in the figure above. The figure shows the application of persistent homology. As the size of the radius of the vertices increase, the number edges which connect the vertices will increase. The distance of the vertices is given by the cosine similarity of the vectors associated with each vertex. Researchers get 9 diameters of birth and 9 diameters of death with 10 vertices. Because all the diameters of birth equal to zero for the 0th topological dimension, researchers only retrieve 9 death diameters. For topological dimension 1 (loops), different documents result in different loops. Thus, retrieving all of birth and death diameters will get different numbers of features for different textual documents. Therefore, researchers summarize the information into five features: loop numbers, average birth diameters, average duration diameter, standard deviation of birth diameters, and standard deviation of duration diameters. The resulting 14 features (9 from dimension 0 plus 5 from dimension 1) represent patterns in the text.
3.3. TDA-Based Mapper Analysis Online Algorithm
Mapper is an effective method for constructing useful combinatorial representations of geometric information about high dimensional point cloud data[5]. In the paper “An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis,” researchers adopted a different ideology compared to the methods discussed above. Though also utilizing TF-IDF vectors, researchers then applied “t-SNE” as a filter function. Next researchers chose resolution and overlap. There are many ways to choose resolution and overlapping percentage. Figure 3 illustrates different resolution and overlapping amounts. The more resolution researchers have, the better data will be partitioned and classified and higher the overlapping percentage is, the more compact our resulting graph would be[5].

![Figure 3. Comparison of different resolutions and overlapping percentage: As is evident from the above table, the more resolution we have the better our resulting graph will be classified and the more the overlapping percentage is, the more compact the resulting graph would become.](image)

4. Result and Discussion

4.1. Performance with word embedding and TF-IDF
Researchers in the first experiment utilized Wikipedia Movie Plot from Kaggle to test their text classification method. They selected the movie plots of 4 major genres: Drama, Comedy, Action, and Romance. In order to make sure the classification is solely based on plot texts, researchers focused only on the documents containing more than 200 tokens(words).

Researchers used FastText pre-trained on Wikipedia 2017 with 300d vectors vocabulary size. The selection of the model is based on the result in an initial experiment where FastText performed better than Google word2vec, GloVe, and other methods. They harnessed Ripser package to apply persistent homology and extract topological features. The TF-IDF vectors were extracted with text2vec package.

Based on the measures described in the methodology section, researchers computed two series of topological features based on word embedding and TF-IDF space vectors, namely TF1 and TF2. Firstly, they filled the XGBoost classifier with TP1 and set the parameters (max_depth = 2, eta = 1, and 25 iterations). Researchers then tried adding TP2 features to the same classifier to improve the model. They also tried a bidirectional LSTM to classify the documents without making use of topological features, whose results contain 64 dimensions of output.

Though bidirectional LSTM indicated better performance compared to the XGBoost model with topological features, researchers believed that there are certain information that is exclusively represented by topological features and are not detected by LSTM. Thus, they managed to combine the LSTM and the XGBoost with topological features. Researchers utilized the probabilities resulted from
the two models, LSTM and XGBoost, to a logistic regression model. As indicated in the table below, the model that combines the two original models most effectively outperforms the LSTM accuracy by 1.6%, as well as a 5.1% increase in the F1-score.

Note that the topological features that are extracted from the word embedding space (TP1 and TP2) can classify the records alone with an accuracy, comparable but not equal to the LSTM. Besides, the topological features extracted from TF-IDF space are primarily used to reflect some repetitive patterns in the text. However, using the topological feature sets can boost the accuracy of classification in the ensemble model.

### Table 2. Precisions and F1-scores of different models

| Classifier          | Prec. | Rec. | F1   | Acc. |
|---------------------|-------|------|------|------|
| 1 BiLSTM            | 68.0  | 59.7 | 0.608| 76.2 |
| 2 XGBoost on TP1    | 59.6  | 53.2 | 0.560| 71.1 |
| 3 XGBoost on TP1 & TP2 | 59.9  | 53.7 | 0.564| 71.4 |
| 4 BiLSTM + XGBoost on TP1 | 67.8  | 64.8 | 0.656| 77.3 |
| 5 BiLSTM + XGBoost on TP1 & TP2 | 68.5  | 64.6 | **0.659** | **77.8** |

4.2. Performance with mapper

In the second experiment, researchers used the textual data (poems) of two Iranian poets Hafez and Ferdowsi. The data set gathered from "Shahnameh" (An epic book from Ferdowsi) and "Ghazaliat-e-Hafez", that includes about different 9000 hemistich (ranging from epic wars to love and romance) from both books.

First, researchers partitioned the whole graph into 3 clusters ("Hafez", "Ferdowsi", "Both"), which in "Hafez" cluster researchers had the nodes which include the high percent of Hafezian poems, similarly in "Ferdowsi" cluster researchers had the nodes which include the high percent of poems of Ferdowsi and in the "Both" cluster researchers had about the same amount of both poems. They examined what percent of poems in "Hafez" cluster really belong to Hafezian poems and the same method for other clusters. To do this they simply divided the number of Hafezian poems in each node in the "Hafez" cluster by the number of all poems in each node in the same cluster, and they did the same test to other clusters as well.

Percentage of accuracy = (Number of poems of "Hafez" in each node of the cluster) / (Number of all poems in the whole cluster)

So if researchers have the accuracy percentage of a for a cluster it means that a percent of the poems in that cluster has been labelled correctly. After examining the first test on each cluster, researchers got the following results: for "Hafez" cluster percentage of accuracy was 80 percent, for "Ferdowsi" cluster percentage of accuracy was about 94 percent and for "Both" cluster percentage of accuracy was 40 percent for Hafez poems and 60 percent for Ferdowsi poems. So for the "Hafez" cluster it can be said that 80 percent of the poems in the cluster has the right label and so on for other clusters.

5. Conclusion

Through the analysis, the researcher utilized two different methods in the first paper, namely word embedding and TF-IDF vectors, to represent the topological features of text documents to classify different documents. The first method converts the word embedding into a high-dimensional time series and applied persistent homology as a measure of topological data analysis. The TF-IDF method augments the classical TF-IDF representation of the text with topological features. These are both methods that combine original text analysis algorithms with topological features. Compared to this, the second paper had a better understanding of the evaluation process of TDA. Using the mapper algorithms, researchers were more successful at classifying the documents into different categories. Comparing the two methods, researchers from the first paper were better at harnessing existing methods while paper two’s authors utilized more geometric features of the documents. Future studies should be designed and carried out to accomplish harder tasks that could combine the advantages of both paper. In other words,
researchers should utilize mapper to discover more features of the word documents and then using them as inputs to BiLSTM, XGBoost, or even other natural language processing algorithms to see the performance and form new ideas.

Acknowledgment
First and foremost, I would like to show my deepest gratitude to my teachers, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank my friends and parents for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

References
[1] Cássio MM Pereira and Rodrigo F de Mello 2015 Persistent homology for time series and spatial data clustering Expert Systems with Applications vol 42 Issues 15-16 September 2015 pp 6026-6038
[2] Shafie Gholizadeh, Ketki Savle, Armin Seyeditabari and Wlodek Zadrozny 2020 Topological Data Analysis In Text Classification: Extracting Features With Additive Information [Online] Available, https://arxiv.org/abs/2003.13138, Accessed on April 10, 2020.
[3] Naiereh Elyasi and Mehdi Hosseini Moghadam 2019 An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis. [Online] Available, https://arxiv.org/abs/1906.01726, Accessed on April 10, 2020.
[4] Liang C 2013 The Application of Topological Data Analysis in Practice and Its Effectiveness
[5] Chazal F and Michel B 2017 An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists [Online] Available, https://arxiv.org/abs/1710.04019, Accessed on April 15, 2020.