A modularized framework for explaining hierarchical attention networks on text classifiers

Mahtab Sarvmaili[†,*], Amilcar Soares[‡], Riccardo Guidotti[◇], Anna Monreale[◇], Fosca Giannotti[◇], Dino Pedreschi[◇] and Stan Matwin[†,§]
[†] Dalhousie University, Halifax, Canada
[‡] Memorial University of Newfoundland, St. John's, Canada
[◇] ISTI-CNR, Pisa, Italy
[§] Institute of Computer Sciences, Polish Academy of Sciences, Warsaw, Poland

Abstract

The last decade has witnessed the rise of a black box society where classification models that hide the logic of their internal decision processes are widely adopted due to their high accuracy. In this paper, we propose FEHAN, a modularized Framework for Explaining HiErarchical Attention Network trained to classify text data. Given a document, FEHAN extracts sentences most relevant to the assigned class. It then generates a set of similar sentences using a Markov chain text generator, and it replaces the salient sentences with the synthetic ones, resulting in a new set of semantically similar documents in the vicinity of a given instance. The generated documents are used to train an interpretable decision tree that identifies words explaining the reason for the classification outcome. A quick inspection of these synthetic documents and their salient words helps explain why the black-box has assigned a given class to a document. We performed a qualitative and quantitative evaluation of FEHAN and a baseline on four different datasets to show the effectiveness of our proposal.

Keywords: Explainability, Text Classifiers, Hierarchical Attention Model

1. Introduction

Many automated decision systems rely on highly accurate classifiers, such as deep neural networks. Due to their hidden, difficult to comprehend internal structure, as well as to their sheer size, they are often referred to as *"black box"* models [1]. At the same time, making critical decisions concerning humans without understanding the justification of such a decision is unacceptable both ethically and legally [2]. The widespread adoption of machine learning algorithms has increased the necessity to *trust* to these models to employ them for decision-making [3] in critical situations. Therefore, there is an increasing interest in the Machine Learning community in deriving explanations able to describe a black box's behavior. Various types of black-box explanation algorithms exist, but from a top-level perspective, they are categorized as model-specific versus model-agnostic, and local versus global [4].

Backpropagating the importance signal from the output neuron to the input for each one of the given instances is an explanation approach illustrated by [5]. DeepLift and IntegratedGradient [6] are examples of this method that use a reference example for computing the feature importance. Although these methods can be applied to parameterized functions such as deep neural networks, these methods are strictly dependant on the choice of baseline example. They require a vigorous search to find the best example. ABELE [7] is a model-agnostic explainer for image classifiers that exploit adversarial autoencoders [8] to generate local neighborhood. Although this model had remarkable results, it is only appropriate for image classifiers. SHAP [9] is a model explainer that uses the shapely values to assign the importance values to the input features that represent each one's impact on the probability

[*] corresponding_author@example.ca

distribution of class labels. LIME [10] attempts to explain a black-box model's behavior for a given instance by generating a local neighborhood. This model is considerably simpler and faster than other models, but this algorithm's neighborhood generation can produce examples that are not faithful to the original document. Additionally, based on the argument provided by [11] the attention and attribution mechanism may not be sufficient to explain the classifier's decision for two reasons: (i) the attention mechanism assigns a real value from interval [0,1] to all of the document's sentences. The more critical the sentence is, the higher the assigned value, and it will be close to zero for less significant sentences, but not zero. (ii) the assigned score doesn't distinguish the importance of each sentence for each class label. Since the attention assigns a real value to all sentences, the significance of each sentence on the class labels distribution is not clear. Therefore, the interpretation attention-based models based on the importance value is ambiguous.

In this paper, we propose FEHAN, a modularized Framework for Explaining Hierarchical Attention Network. FEHAN attempts to locally explain the behavior of Hierarchical Attention Network (HAN) [12]. This modular framework is instantiated with HAN that is an attention-based recurrent neural network for classifying a document. The attention layer distinguishes the Informative Sentences (IS) in a given document. Accordingly, they have more impact on assigning a class label for a given instance, so by replacing them with artificial sentences, the data for a document's vicinity can be created. For a given instance classified by HAN, FEHAN generates a set of semantically similar documents. This new set of synthetic documents is exploited to train an interpretable model - a decision tree - from which the important words can be extracted to construct a saliency map explaining the class label for a given document. Learning the interpretable classifier (i.e., decision tree) on the neighborhood documents derives the *important words* representing the features that locally explain the focus of the black-box classifier while classifying a document. In this manner, the original essence of a given document is well preserved, and it will be enriched with semantically similar examples.

We have conducted the experiments on four sentiment analysis benchmark datasets: IMDB, Amazon, Yelp, and U.S. Airline Tweets. The obtained results show that the neighborhood generated by our approach is more diverse for training the interpretable model rather than LIME's. Also, FEHAN preserves the original document's essence in the neighborhood data while it enriches the generated documents by adding semantically similar sentences to the given instance. In this manner, the generated neighborhood data remain faithful to the original document when it has better coverage over the vicinity of the given instance. We then evaluate the FEHAN approach's fidelity against LIME in datasets with text data, demonstrating that we significantly outperform LIME.

The rest of the paper is organized as follows. Section 2 presents related works and recalls notions and procedures composing the explanation method, which is described in Section 3. Section 4 presents qualitative and quantitative experiments. Finally, Section 5 summarizes our contribution, its limitations, and future research directions.

2. Background

The widespread adoption of machine learning algorithms has increased the necessity to *trust* to these models to employ them for decision-making [3] in critical situations. The need to understand the decisions of black-box models has resulted in the growth of research on research on the explainability of these models [1, 13–15].

Various types of black-box explanation algorithms exist, but from a top-level perspective, they are categorized as model-specific versus model-agnostic and local versus global [4]. Recently, early inroads into the explainability of deep learning models for text data have been made [16, 17]. Frequently these systems rely on special features of the classified

documents. The Generative Explainable Framework [16], e.g., assumes the availability of short, aspect-focused summaries of the documents classified. In contrast, [17] relies on product reviews and ties the explanation concept to this domain. In this paper we propose a modular framework for the local explanation of an attention-based deep text classifier by investigating the local neighborhood of a given document. The basic components used in our framework are described in the rest of this section.

2.1. **Hierarchical Attention Network**

One of the document classification algorithms that takes advantage of Recurrent Neural Network (RNN) along with attention mechanism is *Hierarchical Attention Network* (HAN) [12]. HAN attempts to construct the latent representation of documents from the aggregated latent representations of sentences in that document. This model also exploits two levels of attention mechanisms [18] to increase/decrease the value of individual words/sentences while classifying documents. HAN is composed of a word sequence encoder, word-level attention, sentence encoder, and sentence-level attention. The structure of this model is illustrated in Figure 1. A characteristic of this model is the extraction of importance coefficients at word-level and sentence-level, which gives us a better understanding of HAN's decision procedure. This property is a crucial factor in our neighborhood generator module.

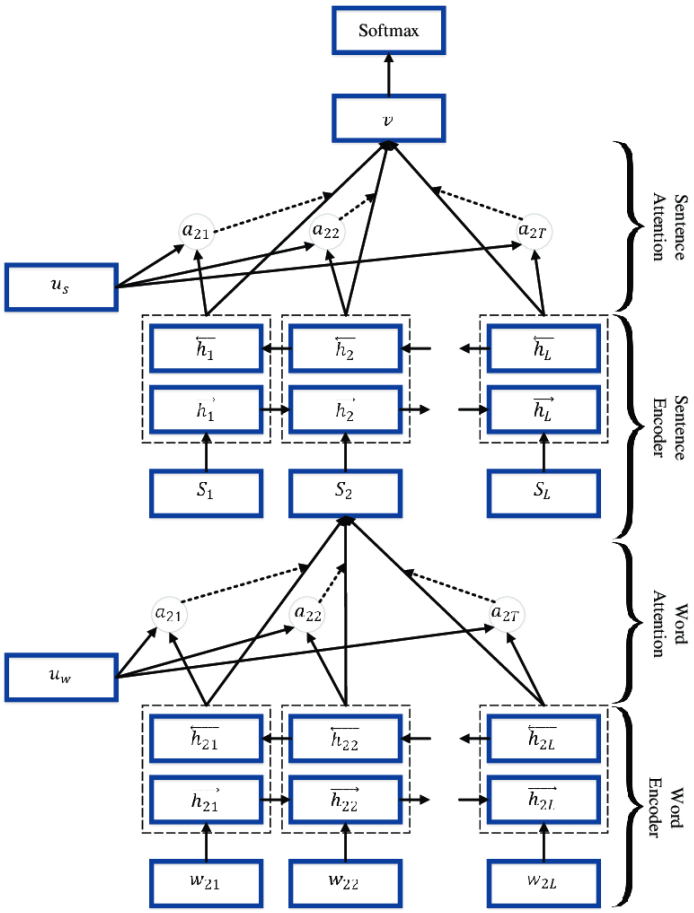

Figure 1. An overview of HAN.

2.2. **Markov Chain Text Generator**

In our proposal, we employ a Markov Chain Text Generator to generate the synthetic neighborhood, representing the key factor in understanding a black-box model's local behavior. Markov chain is a mathematical stochastic model that describes a sequence of events. In this regard the probability of observing each event s_i such as only depends on its preceding event s_{i-1} which is formulated as $P(s_i\,|s_{i-1}\,,\cdot,s_1) = P(s_i\,|s_{i-1})$. The probability of transition from one state to another is called transition probability and can be stored as a transition matrix. The number of rows and columns in this matrix is equal to the number of states n; hence it would be a square $n \times n$ matrix. The simplicity and transparency of this method encouraged us to employ it for neighborhood generation. Additionally, Markov chain text generator doesn't impose any restriction on the order of elements and it allow us to reconstruct the underlying text, unlike n-gram models. Markov Chain's use is essential, as it preserves the distribution of features (in our case, words and their frequencies) of its inputs. In our Markov model, each token of the corpus is a *state*, and we learn the probability of transitions between these *states*. To generate samples from this model, we can start either start from a random token or a selected token and then move forward.

3. **Modularized Framework for Explaining Hierarchical Attention Network**

Explaining the decision of a black box model b on a given instance d, i.e., $b(d) = y$, means presenting an explanation e, that belongs to a human-understandable domain E ($e \in E$). This work focuses on explaining HAN black box decisions, an attention-based recurrent neural network for text documents. Thus, the proposed framework FEHAN is model-specific [19] because tailored to HAN's interpretation. FEHAN describes HAN's local behavior for a specific data point d by inspecting its vicinity. The idea is based on the intuition that, although the decision boundary for the black box can be arbitrarily complex over the whole data space, in the local neighborhood of a data point, there is a high chance to learn an interpretable model able to capture it [4, 10]. Hence, creating a set of semantically similar instances in the vicinity of a given document and exploring the prediction of text classifiers on them. The explanation e produced by FEHAN is a saliency map highlighting the crucial words of the document d contributing to the black-box model's decision.

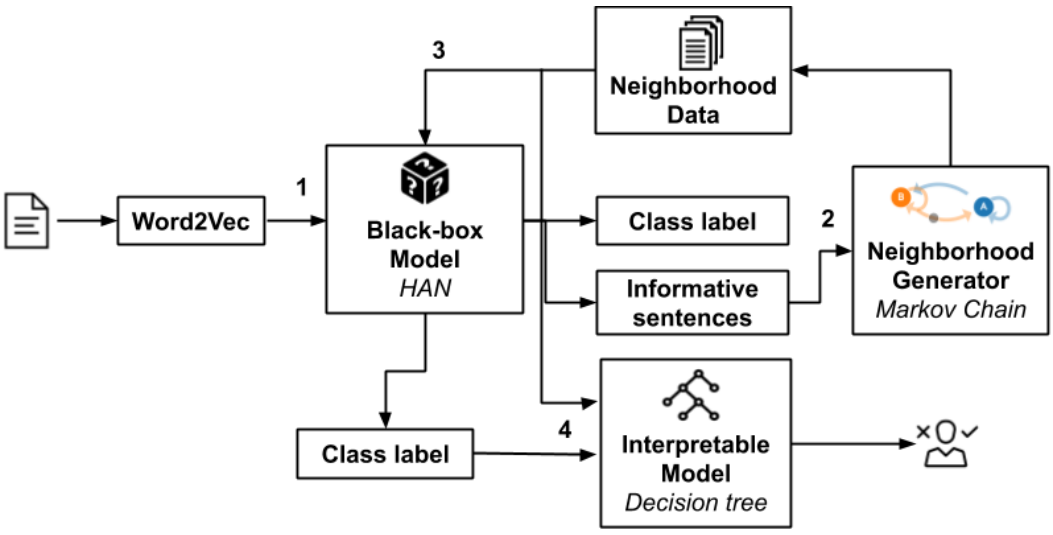

Figure 2. An overview of the FEHAN framework.

An overview of the general structure of FEHAN is given in Figure 2, and this process takes place at the inference phase. The main elements of FEHAN are: *(i)* the informative sentences extraction, *(ii)* the neighborhood generator module, and *(iii)* the interpretable model. The explanation procedure is detailed as follows.

Informative Sentences Extraction. In the first step, HAN receives as input a selected document d, processed with the word2vec embedding model [20], and returns the predicted class label y and the most Informative Sentence (IS) identified in d. The sentence attention layer in the structure of HAN returns a score for each one of the sentences that measure the importance of that sentence for the classification of the document. After extracting the importance signal from the Sentence Attention layer, the IS and the original document are passed to the neighborhood generator module (step 2). The number of IS is a data-dependent hyper-parameter that varies from a dataset to another. For example if the average number of sentences in each document is around 10 sentences, we would select the top three sentences as IS. In the few cases that the length of document is less than the predetermined length of document, the attention layer assigns high scores to the sentences that are empty.

Our explanation method takes advantage of the attention score to obtain the IS; however, if the attention scores are not available, other methods such as backward elimination [21] can be applied to extract the IS.

Neighborhood Generator. This module first receives the selected document along with the index of IS. To generate the synthetic sentences S, it looks at the first element of IS. If the first element is a word, the Markov chain chooses it as the initial state; otherwise (the index of IS refers to the end of the document), the Markov chain starts from a random word. The general structure of Markov Chain text generator is similar to the transition matrix, but for the implementation purpose we have employed dictionary of dictionaries to preserve all possible states (words) and all possibilities for the next item in the chain. Then it creates the neighborhood data H by replacing the IS of the document to be explained with the ones in S (step 3). Note that, given a document d with m IS, the module generates $m \times |S|$ synthetic documents, exchanging time by time one of the m sentences with one of the sentences in S. Finally, each synthetic document in H is labeled by using the HAN. For the generation of the synthetic sentences S, we proposed the Markov Chain. This model is based on a sound mathematical foundation, and it tries to model the probability of observing series of events. For text generation, each one of these events is a token of the corpus. In that manner, the synthetic sentences will be semantically similar to the original sentence, according to the distributed semantic principle that "a word is characterized by the company it keeps" [22].

Interpretable Classifier & Explanation.

FEHAN builds an interpretable decision tree c trained on the locally generated documents (step 4). To this end, first, each document in the neighborhood is transformed into a frequency vector representation by using the bag of unigrams. Then, an interpretable classifier is trained on this data representation. Suppose the number of instances of different class labels in the neighborhood is not balanced. In that case, we employ a *heuristic* proposed in [23], that puts higher weight on the minority class and lower weight on the majority class. This model is finally used for extracting the important features (words) for any class label, useful for producing an explanation. The explanation of our method is the saliency map of important words identified by the interpretable model. We have chosen the decision tree as the interpretable model because the graphical presentation of a decision tree allows the reader to overview a complex model easily. Also, the most influential features on the prediction of class label are structured in a top-down format that means the level of each feature in the tree shows the relative importance of features in the prediction [24].

An example of a returned saliency map based on the outputs of FEHAN and LIME is presented in Figure 3, where the words relevant to the identified class label are colored in

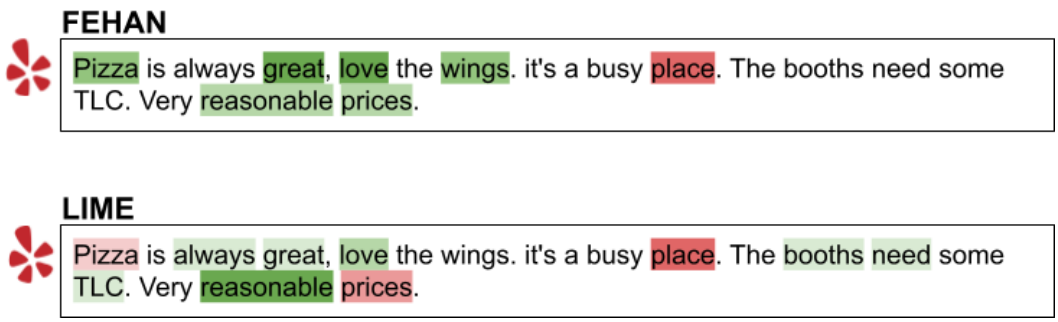

Figure 3. An original document of Yelp data classified by HAN as a *"5 star"* place (the scale is 1 - negative - to 5 - as positive-). The green shades represent the important features for assigning the class label 5 star to the data and red shades for assigning the 1 star. Stronger colors, highlight more important features for classification. The important words are extracted from the decision tree.

green (the shades of green shows the importance of words that agree with the current class), and the ones that are relevant for the opposite class (in this example, negative class) are colored in red. The intensity of colors shows the importance of that word for a class label. In Figure 3, words in green are essential words that support the assignment of the *positive* label to the document, while words in red support its assignment to the negative class. The interpretable text classifier is trained on neighborhood documents generated with artificial sentences created by a Markov Chain text generator.

We highlight that FEHAN[1] is presented as a modularized framework for understanding the behavior of attention-based document classifiers. Our proposed framework's modularity facilitates its development for similar scenarios or usage of other components with the same characteristics.

4. Experiments

FEHAN has been developed in Python, using, the `keras`[2] `Tensorflow`[3] libraries for the HAN, and `scikit-learn`[4] for the decision tree. We experimented FEHAN on four different free available text datasets with different characteristics and features (see Table 1): (i) `IMDB` data, containing highly polarized opinions reviews on movies [25], (ii) `Yelp` data recording reviews for businesses [26], (iii) `U.S. Airline` twitter data, containing anonymous tweets related to the U.S Airlines [27], and (iv) `Amazon` dataset of product reviews [28]. We compare FEHAN's results against the well-known LIME (Local Interpretable Model-agnostic Explanations) [10].

To evaluate the behavior of HAN for document classification, we split the data into three subsets of training *(80%)*, testing *(10%)*, and validation *(10%)* based on the suggested splitting proportions showed in [29]. After training the HAN and the Markov model, we placed them in our framework for the explanation process. For each given instance, we created a set of 300 neighborhood examples, and then we feed them back to HAN for classification.

We compared the outcomes of FEHAN against the LIME text explainer regarding its fidelity to the black-box explanation. An interpretable model's fidelity indicates its faithfulness to imitating the behavior of black-box's behavior in the neighborhood of a particular

[1] https://github.com/MahtabSarvmaili/FEHAN
[2] https://keras.io/
[3] https://tensorflow.org/
[4] https://scikit-learn.org/

Table 1. Summary of the statistics of the four datasets. the number of sentences is shown by #s (average and maximum per document). The average and maximum number of words is shown by #w.

	Yelp	**Amazon**	**Airline tweets**	**IMDB**
Documents	700000	278677	14640	50000
Categories	5	5	3	2
Average #w	9	8	7	13
Max #w	438	169	20	384
Average #s	8	4	2	10
Max #s	150	122	9	117

data point. This is important because the meaningfulness of an explanation should be at least *locally faithful*. The *local faithfulness* of the model relates to the model behavior in the original instance's surroundings being predicted. Our decision tree's fidelity is calculated as the accuracy of the interpretable model's prediction w.r.t the HAN's prediction.

To measure the fidelity, we have tested the model with 6 uniformly sampled sets of the test data. The number of test instances in those experiments is 50, 100, 150, 200, 250, and 300. We reported the observed fidelity for them in Figure 4. The results show that on all datasets, FEHAN outperforms LIME in imitating the black-box behavior. Moreover, FEHAN has remarkably less variance than LIME. Based on the obtained results, FEHAN presents more faithfulness to black-box behavior rather than the local interpretable classifier of LIME.

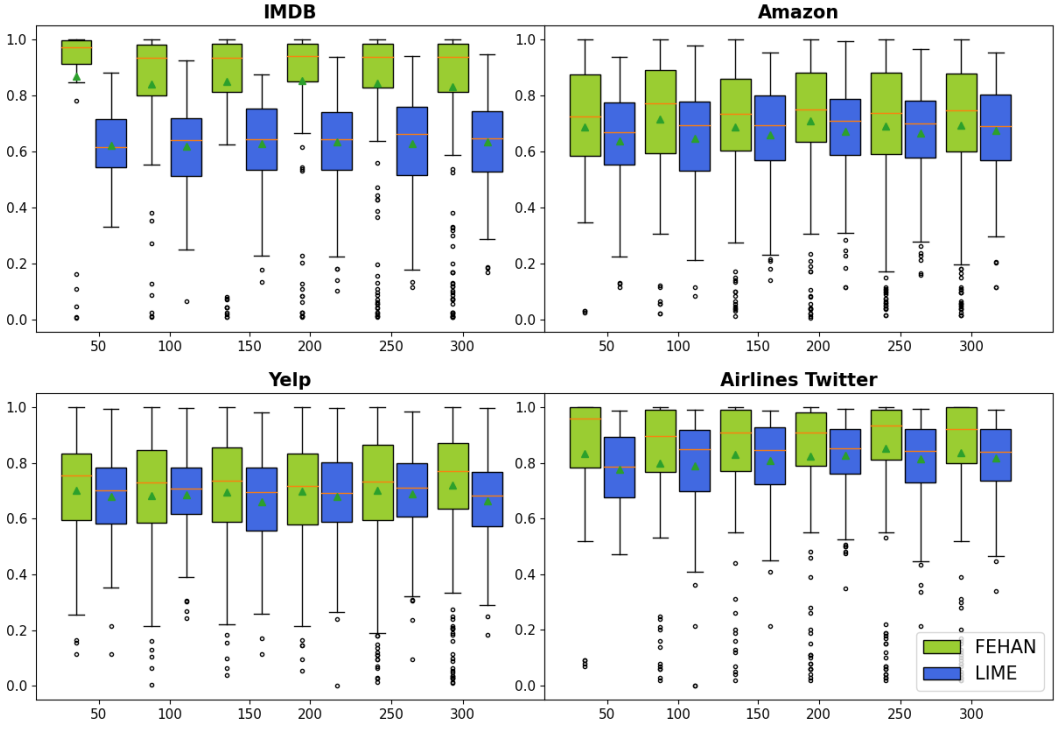

Figure 4. Box plots of $fidelity$ for the four datasets.

Moreover, we conducted the statistical tests on our obtained results to ensure that our results are meaningful in all datasets. To validate our results, we employed the Wilcoxon

test to analyze the fidelity of these models. For this purpose, we reported the obtained p-values of the Wilcoxon test for all four datasets in Table 2. These very low p-values manifest that the differences between FEHAN and LIME are significant; hence FEHAN statistically outperforms LIME regarding fidelity in all benchmark datasets.

Table 2. The $p-values$ of Wilcoxon test for black-box text classifiers

	IMDB	**Amazon**	**Yelp**	**Airline tweets**
50	$1.94 \times e^{-05}$	$2.81 \times e^{-01}$	$4.6 \times e^{-01}$	$1.51 \times e^{-02}$
100	$3.53 \times e^{-11}$	$1.59 \times e^{-02}$	$5.2 \times e^{-01}$	$6.53 \times e^{-02}$
150	$1.69 \times e^{-15}$	$2.14 \times e^{-01}$	$1.89 \times e^{-01}$	$2.4 \times e^{-02}$
200	$5.23 \times e^{-19}$	$1.7 \times e^{-02}$	$2.57 \times e^{-01}$	$2.38 \times e^{-02}$
250	$3.07 \times e^{-22}$	$8.8 \times e^{-02}$	$1.82 \times e^{-01}$	$5.27 \times e^{-06}$
300	$1.24e-22 \times e^{-22}$	$4.1 \times e^{-02}$	$1.99 \times e^{-06}$	$1.13 \times e^{-05}$

Afterward, we numerically evaluated the density and cohesion of neighborhood data generated by the two explanation approaches [30]. To quantitatively show that FEHAN produces higher quality neighborhood text data in comparison to LIME, we explored two approaches: *(i)* measuring the cosine distance between the original document and neighborhood (i.e., the cohesion of the neighborhood), and *(ii)* measuring the Local Outlier Factor (LOF) (i.e., the density and compactness of neighborhood).

The average cosine distance value between the original document and the neighborhood data examples shows the degree of similarity between the neighborhood data and the original document. The results in Table 3 show that the average cosine distance of FEHAN neighborhood is greater than LIME. The higher value of FEHAN's cosine distance on four datasets is related to the vocabulary's diversity in the generated documents.

Table 3. The average cosine distance between the original document and neighborhood data generated by FEHAN and LIME

	IMDB		Amazon		Yelp		Airline Twitter	
	FEHAN	LIME	FEHAN	LIME	FEHAN	LIME	FEHAN	LIME
50	0.674	0.333	0.719	0.320	0.700	0.319	0.789	0.335
100	0.677	0.324	0.750	0.328	0.685	0.321	0.800	0.330
150	0.713	0.322	0.765	0.330	0.682	0.322	0.796	0.333
200	0.689	0.322	0.739	0.329	0.697	0.322	0.794	0.330
250	0.702	0.323	0.739	0.330	0.705	0.322	0.788	0.334
300	0.706	0.322	0.748	0.330	0.702	0.321	0.792	0.329

The Local Outlier Factor (LOF) is a metric for anomaly detection proposed by [31]. In LOF, the data's local density is compared against its neighbors' local densities using a k-nearest neighborhood to identify similar density regions. The points with a considerably lower density to their neighbors are considered outliers using this strategy. In our paper, we have employed the LOF to evaluate the neighborhood goodness, i.e., denser neighborhoods with a lower presence of outliers.

Table 4 reports the average LOF for FEHAN and LIME on the four datasets. We observe that FEHAN has lower LOF for most datasets than LIME, which means that the FEHAN's neighborhood generation based on the Markov Chain leads to a much denser neighborhood with a smaller presence of outliers with respect to LIME.

Comparing the cosine similarity and LOF of generated data shows that FEHAN's has a higher diversity in the number of words. These words are also more semantically similar to the original document rather than LIME.

Table 4. The average LOF between the original document and neighborhood data generated by FEHAN and LIME

	IMDB		Amazon		Yelp		Airline Twitter	
	FEHAN	LIME	FEHAN	LIME	FEHAN	LIME	FEHAN	LIME
50	1.42 * e 07	2.22	1.028	1.78	1.057	7.55 * e 07	1.042	2.11 * e 07
100	7.80 * e 07	2.35	1.027	2.55 * e 07	1.071	4.12 * e 07	1.037	2.10 * e 07
150	1.15 * e 07	2.36	1.028	5.88 * e 07	1.076	2.52 * e 07	1.96 * e 07	4.88 * e 07
200	1.94 * e 07	2.29	1.034	1.45 * e 07	1.067	3.43 * e 07	1.042	2.11 * e 07
250	1.13 * e 07	9.61	1.035	1.53 * e 07	1.063	3.54* e 07	1.89 * e 07	5.03 * e 07
300	6.19 * e 07	2.30	1.79 * e 07	4.26 * e 07	1.069	2.50 * e 07	1.038	3.84 * e 07

Finally, to evaluate the neighborhood data qualitatively, we provide an example of neighbors generated by FEHAN and LIME and the predicted class label assigned by their interpretable models. Figure 5 depicts examples of positive, neutral, and negative neighborhood documents generated and classified by FEHAN's decision tree and LIME's text explainer for the Amazon[5]. The green, blue, and red color shows the positive, neutral, and negative classes. The original and pre-processed document is given in the first two rows of this figure. Since the training process uses the data after pre-processing, the generated instances follow the same style. We observe that FEHAN generated more similar documents to the original one when compared with LIME. The integrity of the data is well preserved, with semantically equivalent examples sampled from the original dataset. FEHAN keeps the central concept of a given instance and non-important parts, and it tries to generate examples that have the same context of the original example but in the vicinity of the data. On the other hand, LIME loses the information in the process of neighborhood generation. This is because LIME suppresses words in the document, while FEHAN only replaces informative sentences with synthetic sentences. Moreover, the process of neighborhood generation by LIME can result in invalid examples due to eliminating all input features. This condition gets worsened for datasets that have generally shorter documents.

5. **Conclusions**

In this paper, we have presented FEHAN, A modularized Framework for Explaining Hierarchical Attention Network. We argue that the common model-agnostic explanation approach (LIME) based on an interpretable classifier's properties in the synthetic, local neighborhood of the instance explained is sub-optimal for text data. We show with FEHAN how neighborhood data can assist in the opening of the black box. Our experimental results show that FEHAN's explanations are better than those provided by LIME's implementation for text data quantitatively and qualitatively. The neighborhood's evaluation indicates that the FEHAN not only preserves the essence of the original document but also enriches the generated synthetic data with semantically similar sentences. This feature of our model is significantly strengthened for smaller datasets when randomly eliminating words can result in invalid examples. Our results also indicate that FEHAN's predictions remain faithful to the HAN's behavior rather than LIME in all of the datasets.

Although the proposed framework is modular in terms of limitations, we observe that it cannot be viewed as a classifier-agnostic approach, as the use of HAN or any other black box providing important sentences is essential. Moreover, the results also depend on the quality of the used embedding models.

As future work, we intend to study the use of other back box models alternative to HAN. One possibility is using methods based on extracting text summarizing (e.g., [32]), which

[5]Although the Amazon dataset has five categories, we have selected the most frequent ones, which are the negative, positive, and neutral generated instances from FEHAN and LIME neighborhood

Original Document	They are just a basic cotton blend short. They have nice length and fit well. But, don't expect to be getting a Nike, Adidas, or Under armor quality product.
Cleaned document	basic cotton blend short . nice length fit well . expect getting nike adidas armor quality product .
FEHAN-Positive	basic cotton blend short . nice length fit well . expect getting nike adidas armor quality product . great support protection outside
FEHAN-Negative	basic cotton blend short . nice bag received extremely disappointed purchase . expect getting nike adidas armor quality product
FEHAN-Neutral	basic cotton blend short . nice sole allows walking extended periods time stability running shoe . expect getting nike adidas armor quality product
LIME- Positive	basic cotton short . length. adidas product.
LIME- Negative	blend short. expect nike armor.
LIME- Neutral	basic cotton blend short. nice length fit well. expect getting nike adidas armor quality product.

Figure 5. A neighborhood example generated by FEHAN and LIME for an instance of Amazon dataset. The green, blue and red depicts the most positive, neutral and the most negative classes in this dataset. FEHAN's decision tree and LIME's text explainer determines the positive and negative label for the generated neighborhood data.

extracts from the document sentences salient for its content. Another essential factor to investigate is the influence of the quality of the text embeddings on the results. One could experiment with more powerful embeddings like BERT. Additionally, we can employ a robust lexical database such as Wordnet for the neighborhood generation. An interesting question to evaluate would be whether using such stronger embeddings translates into significantly improve explanations or whether the word2vector approach suffices. While decision trees are generally accepted as white-box classifiers, alternatives exist (e.g., decision lists), and they could be tried in FEHAN. Finally, further quantitative evaluation of FEHAN should look at the quality of the neighborhood, e.g., by using them in a k-NN classifier and comparing the quality of such classifier to the one obtained from other synthetically generated neighborhoods.

Acknowledgements

The authors would like to thank NSERC (Natural Sciences and Engineering Research Council of Canada) for financial support.

References

[1] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[2] E. EC. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)(Text with EEA relevance). ELI*. 2016.

[3] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[5] A. Shrikumar, P. Greenside, and A. Kundaje. "Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1704.02685* (2017).

[6] M. Sundararajan, A. Taly, and Q. Yan. "Gradients of counterfactuals". In: *arXiv preprint arXiv:1611.02639* (2016).

[7] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi. "Black Box Explanation by Learning Image Exemplars in the Latent Feature Space". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 189–205.

[8] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644* (2015).

[9] S. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[10] M. T. Ribeiro, S. Singh, and C. Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[11] C. Grimsley, E. Mayfield, and J. R.S. Bursten. "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models". English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1780–1790. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.220.

[12] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. "Hierarchical attention networks for document classification". In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016, pp. 1480–1489.

[13] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, and M. Van Gerven. *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018.

[14] Q.-s. Zhang and S.-C. Zhu. "Visual interpretability for deep learning: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

[15] Y. Dong, H. Su, J. Zhu, and B. Zhang. "Improving interpretability of deep neural networks with semantic information". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4306–4314.

[16] H. Liu, Q. Yin, and W. Y. Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by A. Korhonen, D. R. Traum, and L. Màrquez. Association for Computational Linguistics, 2019, pp. 5570–5581. DOI: 10.18653/v1/p19-1560. URL: https://doi.org/10.18653/v1/p19-1560.

[17] S. Ouyang, A. Lawlor, F. Costa, and P. Dolog. "Improving explainable recommendations with synthetic reviews". In: *arXiv preprint arXiv:1807.06978* (2018).

[18] M.-T. Luong, H. Pham, and C. D. Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[19] R. Guidotti, J. Soldani, D. Neri, A. Brogi, and D. Pedreschi. "Helping your docker images to spread based on explainable models". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, pp. 205–221.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[21] R. Kohavi, G. H. John, et al. "Wrappers for feature subset selection". In: *Artificial intelligence* 97.1-2 (1997), pp. 273–324.

[22] J. R. Firth. *Selected papers of JR Firth, 1952-59*. Indiana University Press, 1968.

[23] G. King and L. Zeng. "Logistic regression in rare events data". In: *Political analysis* 9.2 (2001), pp. 137–163.

[24] R. Elshawi, M. H. Al-Mallah, and S. Sakr. "On the interpretability of machine learning-based model for predicting hypertension". In: *BMC medical informatics and decision making* 19.1 (2019), pp. 1–32.

[25] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.

[26] D. Tang, B. Qin, and T. Liu. "Document modeling with gated recurrent neural network for sentiment classification". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1422–1432.

[27] A. Rane and A. Kumar. "Sentiment Classification System of Twitter Data for US Airline Service Analysis". In: *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 01. 2018, pp. 769–773. DOI: 10.1109/COMPSAC.2018.00114.

[28] R. He and J. McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering". In: *proceedings of the 25th international conference on world wide web*. 2016, pp. 507–517.

[29] J. Zhao, Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun. "Adversarially regularized autoencoders". In: *arXiv preprint arXiv:1706.04223* (2017).

[30] R. Guidotti and A. Monreale. "Data-Agnostic Local Neighborhood Generation". In: *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2020, pp. 1040–1045.

[31] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.

[32] A. Rezaei, S. Dami, and P. Daneshjoo. "Multi-Document Extractive Text Summarization via Deep Learning Approach". In: *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*. IEEE, pp. 680–685.