EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR AUTHORSHIP ATTRIBUTION ON SOCIAL MEDIA

Antonio Theophilo[⋆†], *Rafael Padilha*[⋆], *Fernanda A. Andaló*[⋆], *Anderson Rocha*[⋆]

[⋆] Artificial Intelligence Lab. (Recod.ai)
Institute of Computing, University of Campinas, Brazil
[†] Center for Information Technology Renato Archer, Campinas, Brazil

ABSTRACT

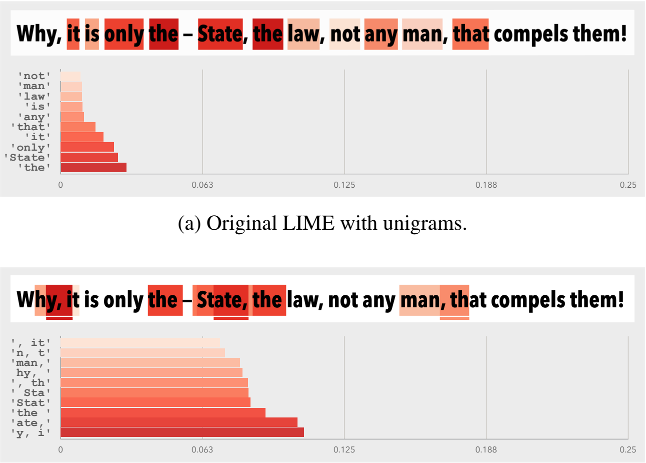
One of the major modern threats to society is the propagation of misinformation — fake news, science denialism, hate speech — fueled by social media's widespread adoption. On the leading social platforms, millions of automated and fake profiles exist only for this purpose. One step to mitigate this problem is verifying the authenticity of profiles, which proves to be an infeasible task to be done manually. Recent data-driven methods accurately tackle this problem by performing automatic authorship attribution, although an important aspect is often overlooked: model interpretability. Is it possible to make the decision process of such methods transparent and interpretable for social media content considering its specificities? In this work, we extend upon LIME — a model-agnostic interpretability technique — to improve the explanations of the state-of-the-art methods for authorship attribution on social media posts. Our extension allows us to employ the same input representation of the model as interpretable features, identifying important elements for the authorship process. We also allow coping with the lack of perturbed samples in the scenario of short messages. Finally, we show qualitative and quantitative evidence of these findings.

Index Terms— Explainable Artificial Intelligence, Authorship Attribution, Social Media, Multimedia Forensics, Text Analysis.

1. INTRODUCTION

The growth of social media in the past decade has profoundly altered how our society shares information. Social media has democratized the creation of information and allowed rapid and wide-ranging dissemination of content. Every person can instantly reach thousands of individuals with a single post, an ability that was limited to governments and large communication vehicles in the past. Due to the fast pace with which content is created and shared, we are constantly bombarded with a myriad of texts, pictures, and videos, often presenting different views about the same topic.

Unfortunately, despite facilitating access to information, social media offers an easy avenue for hate speech, anti-science statements, and large-scale misinformation. Such content is able to influence public opinion, fostering narratives aligned to particular interests in our society [1, 2, 3]. The proliferation of this kind of content is often done anonymously through fake or automated profiles, which represent more than 60 million accounts on Facebook and more than 50 million on Twitter [4, 5].



(a) Original LIME with unigrams.

(b) Our approach with character 4-grams.

Fig. 1: Explanations based on (a) LIME with unigrams, (b) our approach with character 4-grams, and corresponding relevance value per element for the same short text. Both examples consider the same authorship model trained with character 4-grams [6]. A higher value (darker background) means that the interpretable element is more relevant for the model prediction. Elements containing spaces and punctuation — underlined in (b) — are essential for the model decision and are not captured by the original LIME using unigrams.

Verifying the authenticity of profiles and what they propagate online is essential to lessen the negative impact of fake news in our society. However, due to the sheer volume of profiles and daily posts, it is unfeasible to verify all suspicious accounts manually. Therefore, recent works propose Natural Language Processing (NLP) approaches to automatic authorship attribution. This task consists of identifying the actual author of a text/message using its textual content solely, without considering other social media information and meta-data. Even though automatic authorship detection is a relatively well-studied problem by the NLP community when we consider lengthy texts (e.g., novels and articles) [7, 8, 9], the new context of social media poses several challenges for existing methods. These include a massive amount of messages with reduced size (e.g., thousands of 280-character tweets per user), platform-specific language (e.g., usernames, hashtags), non-uniform symbols (e.g., non-ASCII Unicode characters, emojis, emoticons), and expressions (e.g., Internet jargon, onomatopoeia, misspelled words).

To overcome these challenges, recent works [6, 10, 11, 12, 13]

This work was funded by: São Paulo Research Foundation (FAPESP) grants DéjàVu #2017/12646-3, #2017/21957-2, and #2018/10204-6; and by Google LARA (Latin America Research Awards).

rely on data-driven approaches to automatically learn discriminative features (e.g., syntactic, semantic, stylistic) of short text messages. The methods rely upon deep neural networks that classify the author based on different types of message encoding (words, n-grams, characters, part-of-speech tags), outperforming traditional hand-crafted features [14, 15]. While part of the community has focused its efforts on proposing highly accurate models, we investigate a different aspect that is equally important: **model interpretability**.

The decision process of a deep learning model is not transparent. However, in many forensic scenarios (e.g., an investigation or a trial), the reasoning behind its decision is essential to guarantee the trustworthiness of an automated outcome. Besides that, deep learning models are often trained with vast amounts of data from heterogeneous sources, such as social media. Such data may reflect several biases of our society that the model can undesirably capture. With this in mind, interpreting which factors are taken into account for a model decision allows us to identify biases and mitigate them before the model is applied in real-world scenarios.

The unknown nature of Artificial Intelligence models is currently one of the most significant challenges in the field [16, 17], driving an increasing interest in model transparency, fairness, and interpretability. The field of eXplainable Artificial Intelligence (XAI) has proposed several approaches for interpreting machine learning models [18]. A particular line of research focuses on model-agnostic techniques to explain the decisions made by black-box models. These methods are decoupled from specific characteristics of the target problem or architectural details, offering post hoc explanations by analyzing paired input data and output answers. In this work, we are particularly interested in LIME [19], a technique that explains model predictions over individual data samples. It consists of training an interpretable surrogate model that locally approximates the black-box decision space surrounding a sample of interest. In doing so, the method provides clear reasoning about which interpretable components of a specific sample (e.g., words of a text or regions of an image) impact its classification without requiring knowledge of the model details.

Ribeiro et al. [19] evaluate the use of LIME for texts of moderate size, classifying topics in newsgroup discussions using words as the interpretable components. Although individual words (unigrams) capture the semantics of a topic, when we consider shorter text messages, they might not be discriminative enough to account for the writing patterns of an individual, hindering the quality of the explanation. Specifically for the problem of authorship attribution of short messages, Rocha et al. [14] showed that character 4-grams — contiguous sequences of four characters — are more discriminative than many other representations, including unigrams. Furthermore, LIME depends on the number of interpretable components in the analyzed sample to generate enough training data for an adequate local approximation of decision space. When dealing with short texts, we show that using words as interpretable components can hinder the quality of explanations provided by the surrogate model.

We propose an extension to LIME that considers character n-grams as interpretable components to improve the explanations for authorship attribution models that rely on character n-grams [6, 10, 13, 14, 15]. The original LIME formulation was restricted to unigrams and character 1-gram as interpretable components. We evaluate our approach qualitatively and quantitatively in a realistic dataset of tweets, comparing it to the original LIME. Our results show that our extension can better capture the elements used initially as input and improve the generation of perturbed data of the surrogate model in the scenario of short texts, helping to explain complex authorship attribution models more clearly (Figure 1).

The remainder of the text is organized as follows. In Section 2, we discuss existing explainability approaches for authorship attribution. Next, we present our method in Section 3 and its evaluation in Section 4. Finally, in Section 5, we discuss the insights of our work and draw future lines of research.

2. RELATED WORK

The field of XAI applied to NLP has been exploring, under different lenses, how the input data influence predictions made by a model.

One research line aims at deriving influence functions [20, 21] that map how each training sample affects the prediction of a particular test sample. For example, they assess the impact of removing a sample during training in the model parameters and, consequently, predicting a test sample. Despite not highlighting which semantic or stylistic patterns of a text have made the model assign it to a specific author, this "explain by example" approach offers a global view of that text regarding the training data.

Another line focuses on mapping which parts of an input sample (e.g., words, n-grams, tokens, part-of-speech tags) are most influential when determining its class. Unlike influence functions, methods in this category offer a local view over a particular sample, pinpointing its most discriminative parts without directly relying on the relationships to other data samples.

Most works on authorship attribution interpretability fall into the second line of research. For example, Sapkota et al. [15] categorized types of characters n-grams that express an author's different writing patterns (e.g., morphological, semantic, stylistic). Their analyses highlight that the most discriminative category for authorship attribution is stylistic character n-grams that capture punctuation and affixes. Even though the authors provide interesting insights about the usage of character n-grams, their experiments are mostly manual and do not focus on identifying the patterns of a specific author nor explaining the prediction of a text.

Other approaches assess the importance of particular words or n-grams through the analysis of saliency maps [22, 23, 24]. Shrestha et al. [10] proposes a convolutional neural network (CNN) that operates over characters n-grams of tweets. By investigating saliency scores, they found interesting n-grams patterns that differentiate bots from human authors, such as the usage of links at the end of automated tweets or author-specific usage of emoticons and punctuation. Boenninghoff et al. [25] find similar writing patterns when analyzing attention-based maps generated by a siamese neural network optimized for authorship verification (i.e., identifying if the same person authored two texts).

This work investigates interpretability through the lenses of the character n-grams that compose each input text. Following recent findings attesting to the power of such representation for authorship attribution [10, 14, 15], we extend LIME [19] to consider character n-grams as interpretable components instead of limiting the explanation to character 1-gram or unigrams.

LIME offers explanations of a model f by training an interpretable surrogate model g (e.g., logistic regression or decision tree) to approximate the decision space of the original model around a sample x. As discussed in Ribeiro et al. [19], by focusing on the vicinity of x, g does not need to learn all the complex patterns learned by f and can be optimized to provide locally faithful explanations. The authors train g with a novel representation x' that encodes the presence or absence of interpretable components in x. During optimization, training samples are generated by removing one or more components from x, and their proximity to x weights the loss function.

3. PROPOSED APPROACH

Authorship attribution differs from some traditional NLP tasks (e.g., topic identification and sentiment analysis) in the sense that it seeks subtle clues of how an individual writes that go beyond the semantic meaning of the text. Not surprisingly, focusing on individual words (unigrams) and their semantic content is often detrimental to identifying authorship [15]. Furthermore, smaller messages coupled with a high number of authors in the social media domain means that models have an even more difficult task of distinguishing such writing telltales from reduced sample sizes per class. [14]. Character n-grams increase the granularity of input text by considering contiguous sequences of n characters. Besides increasing the sample size of each message, they capture the writing patterns of an author, such as their use of punctuation, abbreviations, emojis, emoticons, and capitalization [15].

With this in mind, we extend upon the original LIME formulation to consider arbitrary character n-grams instead of individual words or characters when generating an explanation. Given a small text $x \in R^d$ whose prediction by model f we want to explain, we build an interpretable representation $x' \in \{0,1\}^{d'}$ as a d'-dimensional binary vector. Each dimension encodes the presence or absence of each of the d' unique character n-grams of x. Similar to LIME, we seek a surrogate model g that operates over the interpretable representations to approximate the decision space of f in the vicinity of x. To train it, we sample perturbed instances $z' \in \{0,1\}^{d'}$ with a fraction of non-zero components of x' and their respective messages z in the original representation R^d. Model g is optimized to approximate $f(z)$ (i.e., the prediction of perturbed instances z), given as input the interpretable representations z'. As perturbations in x may generate samples far apart in input space, LIME weights the influence of instances z in the optimization concerning their similarity to x, $\Pi_x(z)$. Finally, the weighted square loss $\mathcal{L}$ is defined by:

$$\mathcal{L}(f, g, \Pi_x) = \sum_{z,z'} \Pi_x(z) \left(f(z) - g(z')\right)^2 , \qquad (1)$$

with $\Pi_x(z) = exp(-D(x,z)^2/\sigma^2)$, an exponential kernel over the cosine distance D with width σ, measuring the proximity of z to x.

As we show in Section 4, the added support for character n-grams in LIME offers **two essential benefits** to the explainability of social media authorship attribution models: 1) the use of interpretable textual elements that can better align with the model input, and 2) a better sampling space to generate the perturbed samples and, hence, the potential of generating better surrogate models. The code of our approach is freely available.[1]

4. EXPERIMENTAL EVALUATION

We evaluate our approach to interpreting model decisions for authorship attribution of small texts, comparing it to the original formulation of LIME [19]. In this section, we present the dataset and models used in our evaluation and implementation details for our approach. Next, we compare both propositions qualitatively and quantitatively.

4.1. Dataset and Authorship Attribution Models

We adopt the dataset collected by Theophilo et al. [6], which comprises 130 million messages from more than 56,000 Twitter users. In a following work [13], the authors introduce a sanitization process to remove messages from automated and multiple-user profiles (e.g., celebrities) and define data partitions to develop an authorship attribution model for sets of 50 authors, using character 4-grams representation as input. To assess our modifications to LIME, we use the latest version of the dataset, along with the corresponding authorship attribution model [13] that achieves an accuracy of 69.18% on the validation set.

We also conduct experiments considering a second authorship attribution model based on character n-grams [14], and the results are presented in the supplementary material.[2] Even though our main goal in this work is not to improve the classification performance of these models in particular, we aim to understand how it achieves such results so better strategies can be devised using clear interpretations that may improve the learning capability of the model in the future.

We generated predictions for all samples in the validation set and, from these samples, we defined two subsets for which to generate LIME explanations. The first (RAND) has 1,000 randomly chosen messages (20 messages from each one of the 50 authors), and the second (T-CONF) is the set of all correct predictions with high confidence ($>$ 0.9), consisting of a subset of 2977 samples.

We compared the standard LIME approach using words (unigrams) with our modification using character 4-grams, aligning the interpretable representation with the one used by the target model. In both LIME setups (original and extension), we used Regression Ridge as the surrogate model and set the width sigma of the exponential kernel to 25.

4.2. Interpretation Evaluation

We compare the explanations offered by our proposal and the standard unigram LIME qualitatively in Figures 1 and 2. LIME explains a model decision by presenting, for each interpretable component, how much the overall prediction probability/confidence will change if that component is removed. This can be interpreted as the importance or influence of that component in the prediction. We manually analyzed the ten elements accounting for the highest probability considering both approaches.

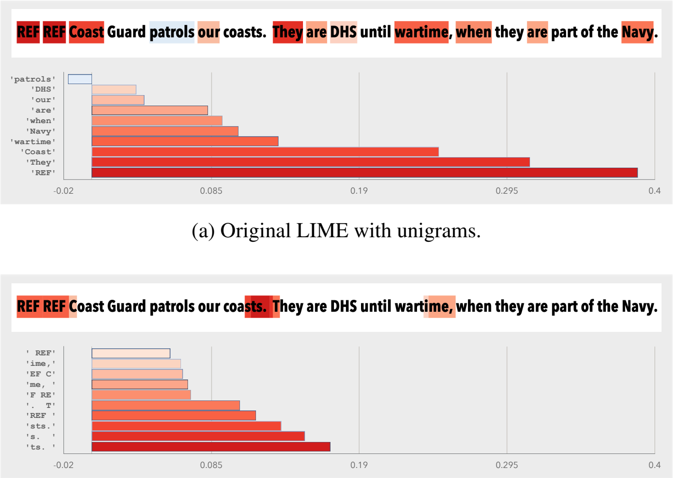
In most comparisons, we observed a scenario with little overlap between the most relevant unigrams and character 4-grams. However, the most significant finding is that the most relevant character 4-grams encompass spaces and punctuation, which are ignored by design by the unigram representation. This result highlights the type of characteristics learned by the attribution model, emphasizing its focus on writing patterns rather than semantic. Additionally, it reinforces the importance of using an explainability technique that is aligned with the target model representation. More examples can be found in the supplementary material.

4.3. Redundancy of Perturbed Samples

Explaining a prediction with LIME requires generating several perturbed samples for training the interpretable surrogate model. In our evaluation, we created 5,000 perturbed points, the same amount as Ribeiro et al. [19]. However, when applying the standard unigram LIME method to our set of short messages, it can be observed that a considerable amount of perturbed samples are identical, not generating the desired amount of distinct data to train the surrogate model. It means that using unigrams as interpretable elements for short messages restricts the number of unique samples generated by LIME.

[1] https://github.com/theocjr/social-media-forensics/tree/master/microblog_authorship_attribution/xai/lime/

[2] https://github.com/theocjr/social-media-forensics/tree/master/microblog_authorship_attribution/xai/lime/icassp_2022



(a) Original LIME with unigrams.

(b) Our approach with character 4-grams.

Fig. 2: Explanations based on (a) LIME with unigrams and (b) our approach with character 4-grams, with corresponding relevance value per element, for the same short text. Elements highlighted in blue (negative relevance) represent terms whose removal would increase the confidence of the prediction. In this example, our method captures the subtle use of two spaces between "coasts." and "They", yielding high importance to all character 4-grams with this telltale.

Table 1: Percentage of duplicated perturbed data for samples in RAND and T-CONF subsets, using unigrams and character 4-grams representations. A lower rate of duplicates means the method generates more varied and less redundant perturbed samples, which leads to better surrogate models and better explanations.

Representation	**RAND**		**T-CONF**	
	Mean	**Max**	**Mean**	**Max**
unigram	55.19 ± 19.74	89.09	51.62 ± 22.06	91.47
char-4-gram	**9.84 ± 7.98**	**29.73**	**7.96 ± 6.27**	**22.78**

The advantage of using character 4-grams is that they allow the method to generate much more non-redundant data, better exploring the input space around the target sample. Table 1 presents the percentage of duplicated perturbed data when using unigrams and character 4-grams for both subsets (RAND and T-CONF). The difference between the two approaches is significant for both random and high confident samples. This indicates that using unigrams to explain the model decisions locally is not recommended for short messages, as they restrict the number of possible perturbations. Nonetheless, RAND presents a higher number of duplicate samples than T-CONF, probably due to smaller messages being selected at random more frequently than the longer and higher-confidence samples of T-CONF.

4.4. Coverage of Explanations

Our qualitative analysis showed that stylistic patterns such as punctuation and spaces are vital to identifying who wrote a message. However, a question remains: how much do we lose in terms of information coverage when adopting unigram LIME — which naturally ignores such elements — to explain an authorship attribution model that relies on character n-grams?

To further investigate this issue, we selected the 20 most relevant character 4-grams for each author in the RAND and T-CONF subsets. We consider a character 4-grams relevant when, on a specific set (e.g., messages of a particular author), it presents higher scores according to LIME.

Table 2: Percentage of the most relevant character 4-grams that contain elements missed by unigram LIME (e.g., space, punctuation, and emojis) across all authors of RAND and T-CONF subsets. Character 4-grams capture these elements, and their writing patterns are essential to attribute authorship. However, the original LIME is unable to identify the majority of them, generating worse explanations.

RAND		**T-CONF**	
Mean	**Max**	**Mean**	**Max**
83.80 ± 10.84	100.0	81.70 ± 12.47	100.0

Using our extension, we select these relevant grams and filter, for each message, the top-10 grams in terms of LIME scores, counting how many times these grams appeared in the dataset. Several of them depict stylistic writing patterns that are not encoded by individual words.

Table 2 presents the percentage of the most relevant character 4-grams containing elements missed by unigram LIME (e.g., space, punctuation, and emojis) across all authors of RAND and T-CONF subsets. We found that, on average, more than 80% of the most relevant character 4-grams contain these ignored characters for each author. This result is in line with previous works that confirm the importance of these textual elements for authorship attribution of short messages [10, 14, 15] and support the need to use a representation that takes them into account to explain the model decision. The supplementary material provides extended analysis varying the number of relevant character 4-grams and results considering a second authorship model [14].

5. CONCLUSION

Automatic authorship attribution of small text messages is paramount to undermine the social impact of misinformation spread by fake and automated profiles in social media. However, while recent research focuses on proposing accurate machine learning models for this task, few works shed light on interpreting what influences a model into crediting the authorship of a message.

In this work, we propose an extension of LIME [19], a model-agnostic interpretability approach, for the problem of authorship attribution of short messages. In line with previous work [10, 14, 15], we propose the use of character n-grams as interpretable components of LIME, showing their importance in capturing author writing patterns and providing additional explanations of the model's decision process. This is even more important when the underlying target model uses this representation to make predictions.

We also show the limitations of the standard unigram LIME approach when dealing with short messages. The small size of these samples hinders the number of generated perturbed samples used to train the surrogate model. Using character n-grams significantly helps to overcome this problem.

For future work, we plan to adapt other XAI techniques for character n-grams, such as influence functions [20, 21] and saliency maps [22, 23, 24], and evaluate them for authorship attribution.

6. REFERENCES

[1] A. Chen, "The agency," https://www.nytimes.com/2015/06/07/magazine/the-agency.html, 2015, [Online; accessed on October 5th, 2021].

[2] F. Davey-Attlee and I. Soares, "The fake news machine. inside a town gearing up for 2020." http://money.cnn.com/interactive/media/the-macedonia-story/, 2017, [Online; accessed on October 5th, 2021].

[3] World Health Organization, "Fighting misinformation in the time of covid-19, one click at a time," https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time, 2021, [Online; accessed on October 5th, 2021].

[4] A. Heath, "Facebook quietly updated two key numbers about its user base," http://www.businessinsider.com/facebook-raises-duplicate-fake-account-estimates-q3-earnings-2017-11, 2017, [Online; accessed on October 5th, 2021].

[5] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *arXiv preprint arXiv:1703.03107*, 2017.

[6] A. Theophilo, L. A. Pereira, and A. Rocha, "A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2692–2696.

[7] P. Juola *et al.*, "Authorship attribution," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2008.

[8] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for information Science and Technology*, vol. 60, no. 1, pp. 9–26, 2009.

[9] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[10] P. Shrestha, S. Sierra, F. Gonzalez, M. Montes, P. Rosso, and T. Solorio, "Convolutional neural networks for authorship attribution of short texts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Short Papers)*, vol. 2, 2017, pp. 669–674.

[11] Z. Hu, R. K.-W. Lee, L. Wang, E.-P. Lim, and B. Dai, "Deepstyle: User style embedding for authorship attribution of short texts," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2020, pp. 221–229.

[12] B. Boenninghoff, R. M. Nickel, S. Zeiler, and D. Kolossa, "Similarity learning for authorship verification in social media," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2457–2461.

[13] A. Theophilo, R. Giot, and A. Rocha, "Authorship attribution of social media messages," *IEEE Transactions on Computational Social Systems*, 2021.

[14] A. Rocha, W. Scheirer, C. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Transactions on Information Forensics and Security (T-IFS)*, vol. 12, no. 1, pp. 5–33, 2017.

[15] U. Sapkota, S. Bethard, M. Montes, and T. Solorio, "Not all character n-grams are created equal: A study in authorship attribution," in *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 2015, pp. 93–102.

[16] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[17] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[18] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[19] 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.

[20] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894.

[21] X. Han, B. C. Wallace, and Y. Tsvetkov, "Explaining black box predictions and unveiling data artifacts through influence functions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5553–5563. [Online]. Available: https://aclanthology.org/2020.acl-main.492

[22] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in nlp," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 681–691.

[23] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[24] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3145–3153.

[25] B. Boenninghoff, S. Hessler, D. Kolossa, and R. M. Nickel, "Explainable authorship verification in social media via attention-based similarity learning," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 36–45.