Improving Visual Interpretability in NLP Short-Text Tasks: A Pre-Hoc Approach Based on Gram-Weighted Tracing

José J. Calderón[1,3],
Mario Graff[1,2], and Eric S. Tellez[1,2]

[1] INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación. Circuito Tecnopolo Sur No 112, Fracc. Tecnopolo Pocitos II, Aguascalientes 20313, México
[2] Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT), Insurgentes Sur 1582, Crédito Constructor, CDMX, México
{mario.graff, eric.tellez}@infotec.mx
[3] CIMAV Center for Research in Advanced Materials. Av. Miguel de Cervantes 120 Complejo Industrial, Chihuahua, Chih. 31136, México.
juan.calderon@cimav.edu.mx

Abstract. Explainability in the decisions made by machine learning algorithms in Natural Language Processing (NLP) tasks related to short texts (*tweets*) presents several key challenges as a result of the high dimensionality and high sparsity observed in the feature matrices generated by short texts under bag-of-words schemes, as well as inherent problems in natural language processing, such as ambiguity, typographical errors, and lack of context. To address these issues, we introduce the Interpretability by Gram-Weighted Tracing (IGWT) framework, a pre-hoc explainability model at the token contribution level. The IGWT framework leverages the advantages of intrinsically interpretable classifiers, such as linear models, while incorporating innovative techniques like the traceability of q-grams back to their original words to improve visual explainability. This approach focuses on visual interpretability, facilitating the detection of biases and patterns in the data, allowing for the optimization of the dataset before training, and offering a comprehensive framework to address the complexity of explainability in NLP tasks related to tweets, such as user profiling and supervised classification. Moreover, this approach aligns perfectly with active learning strategies, making it a suitable choice for iterative model improvement.

Keywords: XAI, Explainability, Visual Interpretability, Pre-Hoc, q-grams, Tweet User Profiling, Tweet Classification, Gram-Weighted Tracing (IGWT), Token Contribution Analysis, Active Learning.

1 Introduction

User profiling and supervised classification of short-texts, such as tweets, face a series of challenges that affect both model performance and explainability in ways that are interpretable to humans, as explained by Zhao et al. [32]. Among the most common problems are high dimensionality and excessive sparsity in the feature matrices, caused by the presence of a high number of unique words in very few tweets [7, 17]. These challenges are compounded by the inherent problems of natural language processing, such as ambiguity, typographical errors, and lack of context [31, 3, 29]. Since Bag of Words (BoW) schemes for short texts tasks produce feature matrices that are highly dimensional and sparse [29], it becomes difficult to establish and visualize the contribution of each word to the model's decisions, which affects the interpretability of the results in human terms.

To address these issues, a pre-hoc explainability model [32] at the token contribution level is proposed, leveraging the advantages of intrinsically interpretable models, such as linear models [7], while incorporating innovative techniques like the traceability of q-grams [24, 13] back to their original words to improve visual explainability. Q-grams [13] —a technique that breaks down words and texts into sequences of q characters— enhance the ability to capture contextual and ordered structures that go beyond traditional BoW approaches.

Q-grams facilitating the detection of suffixes, prefixes, variations, and typographical errors while also increase the number of related texts detected by user profiling and classification algorithms, providing significant advantages that enhance the performance of algorithms [27] [16]. However, q-grams present significant challenges in terms of explainability. Once generated, their contribution to the decision-making process becomes nearly incomprehensible in human terms due to their sheer volume and lack of semantic interpretability. For instance, given the word "inteligente," the corresponding set of q-grams (q=3)

includes ["int", "nte", "tel", "eli", "lig", "ige", "gen", "ent", "nte"]. The fragmented nature of these substrings makes it difficult to directly trace their individual impact on classification outcomes, further complicating their interpretability in machine learning models.

To address this challenge, our core proposal centers on tracing each q-gram back to its original word, unlocking human-understandable explanations into the model's decision-making process. This tracing mechanism is pivotal: it retains the computational efficiency and predictive power of q-grams while bridging the gap to human understanding. By visually mapping q-gram contributions to their source words, we preserve the performance benefits of subword features and recover the intuitive meaning of full words—enhancing interpretability without compromising robustness. Crucially, visualizing contributions enables human-readable explainability [23].

This proposed approach adopts a pre-hoc explainability model, which could also be considered hybrid as it incorporates post-hoc techniques [32] to enhance explainability from the start of the modeling process [17]. Before training the machine learning model that the researcher intents to develop -a target model-, an independent linear classifier is used as a surrogate model to evaluate the contribution of each token and its q-grams, allowing for the identification of patterns and biases present in the data previous to the training of the target model for user profiling or text classification.

The visual analysis not only enhances the level of explainability regarding the algorithm's decisions but also enables the end-user of the framework to tailor the dataset to their specific objectives, such as correcting undesirable or biased contributions or adapting the dataset to a different region or domain from the original. Moreover, this approach aligns perfectly with active learning strategies [22], making it a suitable choice for iterative model improvement.

The proposed explainability framework relies on the weights of q-grams, enabling the setting of thresholds on the weights to be considered. This feature will be used to demonstrate the feasibility of the proposal by applying different thresholds to the dataset and comparing the results.

In summary, this proposal introduces an innovative approach to enhancing the explainability of profiling and classification by explicitly addressing the challenges posed by high dimensionality and the high sparsity inherent in short texts. By leveraging intrinsically interpretable models and techniques for tracing q-gram contributions, it offers an approach that not only facilitates the understanding of model decisions, but also enables the efficient refinement or adaptation of the dataset as required by the researcher.

2 State of the Art

2.1 Q-grams tokenization

As analyzed in [24], q-grams capture syntactically relevant information by representing character sequences, making them particularly useful for informal and unstructured text. Their ability to handle common spelling errors—frequent in social media—enhances model robustness against noisy data. Moreover, since they do not rely on lexical boundaries, q-grams adapt more effectively to linguistic variations such as abbreviations, slang, or colloquial language.

In their experimental analysis [24], authors demonstrate that models using q-grams for tokenization achieve significant improvements in accuracy and output quality. These findings support the use of q-grams as an effective technique in NLP, especially for tasks involving short-text and high variability, such as tweet analysis.

To further validate these findings, we conducted a two-stage experiment aimed at assessing the impact of q-gram tokenization in short-text NLP tasks. In the first stage, standard tokenization (without q-grams) was applied to establish a baseline. In the second stage, q-gram tokenization was performed. Fig. 1 presents the differences in f1-score, precision, and recall for each dataset, relative to the baseline. Positive values indicate performance gains, while negative values reflect a decrease. The results reveal a consistent trend of improved metrics when using q-grams, reinforcing their applicability within our proposed approach.

2.2 Challenges in NLP Tasks related to Tweets

Tweets profiling and classification present a challenge due to the high diversity of contents, ambiguity, limited length, and lack of context in these microtexts. These factors exacerbate common issues in NLP, such as typographical errors, regional differences in language use and words out of vocabulary, which reduce the quality of the data for extracting accurate and reliable information [1].

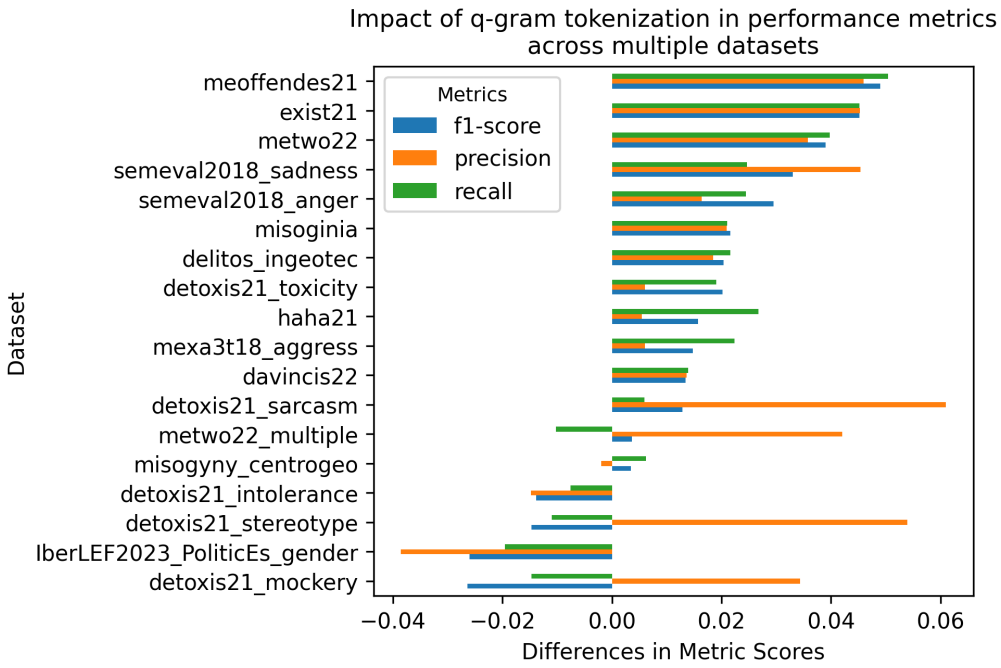


Fig. 1: Differences in performance metrics observed after applying q-gram tokenization to NLP datasets. The experiment involved two stages: first, tokenization was performed without q-grams to establish a baseline; then, q-gram tokenization was applied. The values plotted represent the difference in each metric (f1-score, precision and recall) for each dataset, relative to the baseline. Positive values indicate improvement, while negative values suggest a performance drop.

2.3 Dimensionality Reduction and Feature Selection Techniques

Moreover, the bag-of-words schemes used for tweets representation often produce feature matrices with high dimensionality and sparsity, leading to a low standard deviation[19] in the token coefficients calculated by weighting schemes like Term Frequency-Inverse Document Frequency (Tf-Idf)[12]. Short texts, such as tweets, exhibit these distinctive characteristics in their matrix representation. High dimensionality arises from the wide variety of unique terms in the corpus, while high sparsity occurs because each text contains a limited number of words, resulting in feature matrices with many null values. Low standard deviation is observed due to the uniformity in term frequency, as most words appear only once in each document. These properties exacerbate common issues in natural language processing, such as typographical errors and regional differences in language use, which further reduce the quality of data for extracting accurate and reliable information[19].

This makes it challenging to identify the most influential tokens and complicates the interpretability of the model [29, 15]. Advanced tokenization techniques, such as q-grams —as discussed in the previous section— have become indispensable for capturing linguistic patterns and improving accuracy in short-text processing algorithms, and can even enhance explainability by specifically identifying the most important subwords (q-grams) in a prediction [29]; however, their use further increases the dimensionality and complexity of the data, making it more difficult to identify key features and interpret how these patterns contribute to the model's decisions [20, 16].

As explained in [26], to address the challenges related to high dimensionality in short-text schemes, common strategies focus on applying methods for regularization, feature reduction, and feature selection. Methods like l1 and l2 regularization (Elastic Net) help control overfitting and further reduce complexity in high-dimensional spaces, and techniques such as mutual information, chi-square, and Principal Component Analysis (PCA) reduce the feature space by transforming the data into a smaller set of uncorrelated components, retaining the most important information without compromising model performance.

However, a challenge with all these reduction and selection techniques in high-dimensional spaces is that multiple feature subsets can achieve similar performance metrics [14]. These subsets may not be unique, and some may overlook features that are important for specific instances, leading to a model that

performs well globally but misses key features needed for local or instance-specific predictions, which is crucial for the interpretability of the prediction in those cases.

These reduction and selection techniques are effective, but they must be applied with extreme caution to avoid compromising explainability.

2.4 Visual Explainability in High-Dimensional Spaces

At the same time, recent studies [9, 30] have shown that combining these techniques along with visual explainability approaches allows for a more balanced analysis of models, providing clarity in global patterns and facilitating the understanding of decisions at the individual level.

Numerous studies highlight the importance of explainability in building reliable and transparent models [32, 2]. This can be achieved through natural language explanations, visualization of key instances and features, or contribution graphs, which help interpret the most relevant factors. Visual techniques, such as ranking tokens by their importance in predictions, assist in identifying the most influential elements, enhancing the ability to detect biases and errors [11].

Visualization techniques, such as heatmaps and bar charts, are useful for analyzing global patterns [18], while techniques based on t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) enable exploration of relationships between instances in lower-dimensional spaces [8].

However, although these techniques can provide a global understanding of the data, they are not sufficient to offer clear and detailed explanations for specific predictions in high-dimensional spaces.

2.5 Explainability in Artificial Intelligence

With the rapid rise of Artificial Intelligence (AI) in various domains, the need for explainability in AI models has become increasingly important [32]. Explainable Artificial Intelligence (XAI) emerged as a field dedicated to making the decisions of AI models understandable to humans, addressing concerns about how complex models—especially black-box models like deep learning—arrive at their predictions. This is crucial not only for trust and transparency but also for improving and refining the models based on their decision-making processes.

To address the challenge of understanding decisions, tools like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) were developed. These tools have gained widespread recognition for their ability to offer both global and local explanations of model predictions. LIME works by creating local surrogate models around a prediction, helping to explain individual outcomes by approximating the decision boundary in the vicinity of the instance. SHAP, based on cooperative game theory, attributes a Shapley value to each feature, indicating its contribution to the prediction, making it valuable for understanding feature importance both globally and for individual predictions [2].

However, despite the success of explainability tools like SHAP and LIME, they encounter significant challenges in high-dimensional, large-scale, and sparse/low-density environments. Their algorithms, while powerful, are computationally intensive, becoming increasingly expensive, and their performance degrades as the number of features grows. Moreover, the complexity of visually representing a vast number of features complicates the task of providing clear and interpretable results. These two issues highlight the need for more optimized or hybrid solutions that can handle the demands of high-dimensional data without compromising visual explainability [28].

For instance, in the case of a simple tweet such as *"#masterchefmx y salen los putos a tener fantasías con el chef irlandés"*, a vector with 201 q-grams is generated. Among these, only 9 correspond to complete and readable words, such as 'chef', 'fantasías', and 'irlandeses'. The remaining 192 q-grams, however, consist of fragmented sequences like *'#m', 'alen', 'a'*, and *'con'*, which are difficult to interpret in human-readable terms.

Figure 2 presents the SHAP output for this example tweet, both using and not using q-grams. It is evident that the visual representation of SHAP becomes increasingly complex and even unusable, despite the short length of the tweet, due to the high number of generated q-grams.

This challenge escalates exponentially in user profiling tasks, where tweets from a single user must be grouped based on author-related similarities. In such cases, grouped tweets often produce over 25,000 q-grams, making their visualization through SHAP virtually impossible as shown in figure 2. Moreover, the computational cost becomes prohibitively high.

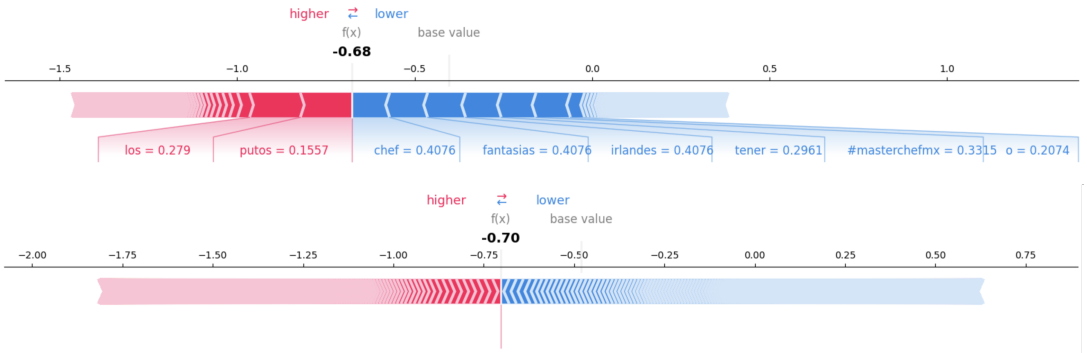


Fig. 2: SHAP output displaying the tokens and their contributions for the given example tweet. The top graph without q-grams accurately displays all tokens. On the other hand, the bottom graph incorporates q-grams for the same example tweet but fails to visualize them properly due to the high number of generated q-grams, rendering it ineffective and null in terms of explainability

2.6 Pre-hoc and Post-hoc Explainability Strategies

Explainability can be implemented through different approaches [2]. Pre-hoc strategies integrate interpretability into the model's design, which sometimes involves sacrificing accuracy. Pre-hoc models are *interpretable by design*, meaning their structure allows decisions to be easily understood. For instance, the hierarchical structure of decision trees makes it easy to interpret decisions at each node, while the coefficients of linear models are directly interpretable as the contribution of each feature to the prediction.

On the other hand, post-hoc strategies are applied after the model has been trained, especially for complex and difficult-to-interpret models (such as neural networks). These techniques generate explanations without modifying the model's structure. Post-hoc models can use surrogate models, which are interpretable models designed to approximate the behavior of a complex model. For example, LIME trains a linear surrogate model around a specific prediction, allowing it to identify which features most influence that local prediction. However, the downside of surrogate models is that they are approximations, meaning they may not fully capture the complexity of the original model or may oversimplify it.

Lastly, hybrid strategies combine pre-hoc techniques, which ensure explainability from the start, with post-hoc techniques that provide more detailed explanations [5]. This approach seeks to balance explainability with model performance and is becoming increasingly common in the field of XAI.

Aligned with combined explainability strategies, we propose a pre-hoc approach that leverages a linear model during the preprocessing stage, independent of the target model intended for use. By using interpretable coefficients, this approach facilitates the identification of biases, problematic patterns, and anomalies in the dataset. Also, it allows researchers to refine the dataset to suit specific objectives, such as adapting it to a different domain or aligning it with particular research goals. This flexibility ensures that the dataset is both optimized and tailored to meet the requirements of the task at hand.

Additionally, this approach is well-suited for Active Learning cycles [22]. After each training iteration, the data can be reanalyzed, allowing for continuous adjustments and improving the model's efficiency and accuracy as it learns actively through the interactive explainability of its decisions.

In summary, while traditional visualization techniques and dimensionality reduction methods are useful for general analysis, and tools like LIME and SHAP enhance local interpretability, there remains a need to strike a balance between dimensionality reduction and explainability in order to achieve models that are both effective and interpretable in the context of short-text tasks, such as tweets profiling and classification.

2.7 Embedding Schemes and Transformers

In this proposal, we chose q-gram tokenization [24] due to its robustness in handling common issues in short-texts (tweets) —as discussed in the previous section—. By breaking words into sequences of characters, q-grams can detect and correct typographical errors and linguistic variations. Unlike embedding schemes such as Word2Vec or FastText, q-grams are particularly effective with Out of Vocabulary Out of Vocabulary (OOV) words [6], such as neologisms, hashtags, or invented terms, which are very common in

tweets. It is crucial to include these capabilities in user profiling and supervised classification of tweets. Also, embeddings require intensive pretraining, which increases complexity and computational cost. In contrast, q-grams allow for clear visual inspection, making it easier to interpret how subwords contribute to the prediction.

Although transformers, like Bidirectional Encoder Representations from Transformers (BERT) [12], are extremely effective in many NLP tasks, their use in user profiling and supervised classification of tweets may be unnecessarily computationally expensive, and they lack inherent interpretability [32] requiring additional techniques to understand their decisions [11]. In contrast, q-grams combined with linear estimators, such as Support Vector Machine (SVM), have shown to achieve comparable performance in NLP tasks focused on short-texts, with significantly lower computational cost and inherent explainability.

Furthermore, this proposal seeks to develop an explainability framework focused on active learning, where q-grams and linear classifiers enable an iterative analysis of data beyond the reach of modern classifiers such as those based in embeddings and transformers. This pre-hoc approach optimizes the dataset in the early stages of processing, improving the efficiency and quality of any target model.

3 Our Approach

Building upon the challenges and preliminary solutions outlined in 2 (State of the Art), this proposal introduces IGWT, a visual explanatory framework specifically designed to address the high-dimensional and sparse nature of tweet data. At its core, IGWT integrates q-gram tracing with linear models, applied during the preprocessing phase to tailor the dataset to task-specific requirements via an active learning cycle.

To enhance visual interpretability, we compute the contribution of each q-gram and then trace the most relevant ones back to their original words. This mapped traceability enables a human-readable association between influential q-grams and the lexical units they originate from. For instance, the word *intelligent* produces q-grams (q=3) such as [*' in', 'int', 'nte', ..., 'ent', 'nt '*]. By mapping these q-grams back to the word, we enhance interpretability by allowing a direct association between the extracted features and their linguistic context. In addition, this approach enhances explainability without compromising the performance advantages of q-grams—namely, their ability to capture morphological variation and handle out-of-vocabulary terms effectively.

Figure 3 illustrates this process using the word *estúpida*. Each q-gram's weight is derived by combining its Tf-Idf score with the corresponding coefficient assigned by the linear classifier.

This explainability mechanism is primarily pre-hoc, as it is applied before training the target model, to analyze and optimize the dataset at an early stage. Nevertheless, by employing a surrogate model—typically associated with post-hoc techniques—during preprocessing, this approach combines the benefits of both paradigms.

3.1 Datasets

Our research leverages the vast number of datasets available for tweet analysis tasks, forums and competitions, with Iberian Languages Evaluation Forum (IberLEF) and Conference and Labs of the Evaluation Forum (CLEF) being among the most notable sources. These datasets offer the advantage of being precleaned, structured, and labeled, which facilitates tasks such as supervised classification. However, depending on the dataset and task, additional preprocessing may be required. Traditional techniques—such as stopword removal, lowercasing, and punctuation filtering—are applied when necessary to improve data quality and model performance.

In user profiling tasks, for example, a clustering step using the K-Means algorithm was performed to group tweets by user, enabling profile-level classification. To enhance the effectiveness of this process, standard preprocessing techniques were applied prior to clustering, helping reduce noise and improve the coherence of the resulting user groups.

3.2 Text Transformation with Micro Text Classifier (μTC)

Once the datasets have been selected, the next step is to transform the text into a vector space supported by the BoW schemes, ensuring that the transformation process includes the fragmentation of words into q-grams. For this transformation, we have chosen the μTC framework [25] to convert datasets into vector-label pairs. This choice is driven by three key factors: i) μTC seamlessly integrates with the scikit-learn

workflow, facilitating its use within our development pipeline; ii) μTC has demonstrated exceptional performance in various tasks and competitions related to user profiling and supervised classification of tweets, delivering outstanding results; and iii) it encapsulates essential tasks such as normalization, segmentation, advanced tokenization and transformation into vector spaces, significantly streamlining the overall process.

Among its advanced tokenization capabilities, μTC allows for and controls the decomposition of words into q-gram sequences ranging from 1 to 5 characters, which not only improves prediction accuracy but also serves as the foundation for implementing explainability at the subword level. While μTC facilitates this advanced tokenization, the traceability of each q-gram back to its original word —a core objective of our proposal— is achieved through additional methods built on top of the decomposition performed by μTC.

3.3 Dimensionality Reduction and Feature Selection

As reviewed, feature engineering techniques are essential for reducing and optimizing vector spaces used by NLP algorithms. For user profiling and supervised classification of tweets, our proposal has selected methods that take into account the assigned label and perform well in high-dimensional spaces, such as mutual information and chi-square. In our approach, we apply these techniques to exclude non-significant tokens in a massive, quick, and computationally inexpensive way, streamlining the process. Although this filtering step alone may not significantly reduce dimensionality in all cases, its very low computational cost and minimal impact on interpretability justify its application.

3.4 Integration of an Interpretable Linear Model (LinearSVC)

In general terms, training a linear classifier for binary classification involves finding a hyperplane that separates the two classes in the feature space (for multiclass classification, this could becomes a one-vs-rest problem). For a new instance, the decision function is a linear combination of the features weighted by their respective coefficients, which determines on which side of the hyperplane it lies and, therefore, to which class it belongs. Specifically, Linear Support Vector Classification (LinearSVC)[4] is an implementation of the linear classifier based on the SVM, specialized in finding the optimal hyperplane that maximizes the distance (or margin) between classes.

LinearSVC is suitable for classification problems where the classes are linearly separable and handles high dimensionality well, making it ideal for supervised classification of tweets, and consequently, for our proposal. LinearSVC enhances interpretability by assigning coefficients to each token, revealing the degree of its contribution to the classification outcome.

Furthermore, the selection of LinearSVC was based on specific criteria, as it meets other functionalities sought in this proposal: i) the calculation of coefficients and the decision function are integrated into the algorithm, so no additional computations are required; ii) it supports l1 regularization (Lasso), which reduces small coefficients to zero and automatically removes features below a set threshold; iii) it integrates easily into the preprocessing phase; and iv) it is independent of the target model chosen for user profiling or classification.

In this way, by combining the explanatory properties of LinearSVC with the traceability of q-grams back to their original words in the tweet, our approach enhances the granularity of visual interpretability for each tweet in the dataset. This allows the proposed IGWT framework to facilitate, in human-understandable terms, the identification of misclassified or mislabeled instances, defective tokens, biases and any other insight at the individual instance level, while also providing valuable insights into the model's global behavior.

3.5 Interactive Web-Based Framework

As part of the innovation and practical application of the proposed IGWT framework, an interactive website has been developed that allows researchers to upload and analyze datasets of tweets for user profiling tasks, supervised classification, or related tasks. This platform displays the results directly on screen, facilitating a detailed and visual analysis of each tweet with its predicted and pre-labeled classification.

The framework organizes and presents the data in a multi-level structure, including details such as the decision function, true and predicted classes, and the q-grams, visualized their importance using

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

heat-maps. Researchers can dynamically adjust parameters such as threshold for the decision function and threshold for the weight to optimize visualization and results.

This data structure is highly reusable for various additional procedures, including, for instance, the implementation of embedding techniques to explore relationships between the q-grams. This opens a wide spectrum of possibilities and allows for adapting and expanding the analysis to a variety of advanced approaches in natural language processing and machine learning.

The site also serves as an ideal platform for active learning, allowing users to make iterative adjustments, observe how these impact explainability in NLP tasks almost in real time, and continuously refine the model according to their specific research needs. This capability for iteration and adjustment turns the website into a powerful tool for experimentation and optimization of the dataset prior to the final training of tweets, thereby improving the understanding, accuracy, and interpretability of the target model.

The IGWT framework is designed for researchers with a good understanding of the dataset's domain and the target model, ensuring that any experimentation aligns with both the dataset's characteristics and the model's objectives. A screenshot of the interface is provided in Appendix B to illustrate how tokens and q-gram contributions are visually presented in practice, showcasing the interpretability features available in the interactive environment.

3.6 Algorithm – Token-Level Interpretability via Q-Gram Tracing

Algorithm 1 presents the interpretability workflow proposed for tweet analysis. Its purpose is to enable a visual and interpretable understanding of how each lexical unit (token or word) contributes to the model's predictions, by aggregating the individual contributions of its constituent q-grams. This understanding is both quantitative—since it is based on the weighted contributions assigned to each q-gram—and qualitative, as the tracing mechanism links these subword units back to their original linguistic context. As a result, researchers can identify and mitigate potential biases, detect mislabeled data, and extract critical insights prior to applying the target model to the dataset.

Algorithm 1: Q-Gram-Based Interpretability Framework

Input: Tweet dataset D, parameters: sample size s, decision threshold θ, weight threshold τ
Output: Visualization-ready interpretability structure
Step 1: Data Input and Configuration
Load dataset D; configure parameters $\{s, \theta, \tau\}$
Step 2: Preprocessing and Tokenization
foreach *tweet* $t \in D$ **do**
  Tokenize t into words and q-grams ($q \in \{-1, 2, 3, 4\}$);
  Compute TF-IDF weights for each token;
Step 3: Feature Selection and Dimensionality Reduction
Apply Mutual Information or Chi-Square to select top-k informative tokens
Step 4: Train Linear Model
Fit LinearSVC on selected features to obtain weight vector w and bias b
Step 5: Apply L1 Regularization
Discard tokens where $|w_i| < \tau$
Step 6: Compute Decision Scores and Classify
foreach *instance* x **do**
  Compute $Score(x) = \sum_i TF\text{-}IDF_i(x) \cdot w_i + b$;
  Predict class: 1 if $Score(x) > \theta$, else 0;
Step 7: Q-Gram Tracing to Original Words
foreach *q-gram* g **do**
  Map g back to its originating word(s) in t
Step 8: Build Interpretability Structure
foreach *tweet* t **do**
  Store predicted class, score, token contributions, and q-gram-word mappings;
  Save as hierarchical dictionary (see Appendix)
Step 9: Visualize Results
Render heatmap contributions of tokens and q-grams;
Display original words with their associated influence

This algorithmic facilitates efficient data analysis and enables quick access to adjust or validate interpretation outcomes. Besides, the stored structure (JSON format) is also adaptable for future research, as the data is organized and readily accessible.

4 Experimental Results

In this section, we evaluate the effectiveness of the proposed IGWT framework. Emphasizing that the strength of IGWT lies in its ability to utilize the coefficients of the linear estimator to maintain competitive performance in models, while also leveraging and combining the traceability of q-grams to their original words to provide efficient visual aids that enhance the explainability of decisions. Additionally, it is important to highlight that IGWT can process a large number of instances simultaneously, further increasing the identification of the most influential words and q-grams and, consequently, its explanatory capacity. All of this is focused on addressing the challenges of high dimensionality and high dispersion generated by short-text in NLP tasks, such as user profiling in tweets.

4.1 Explainability Example of a Tweet

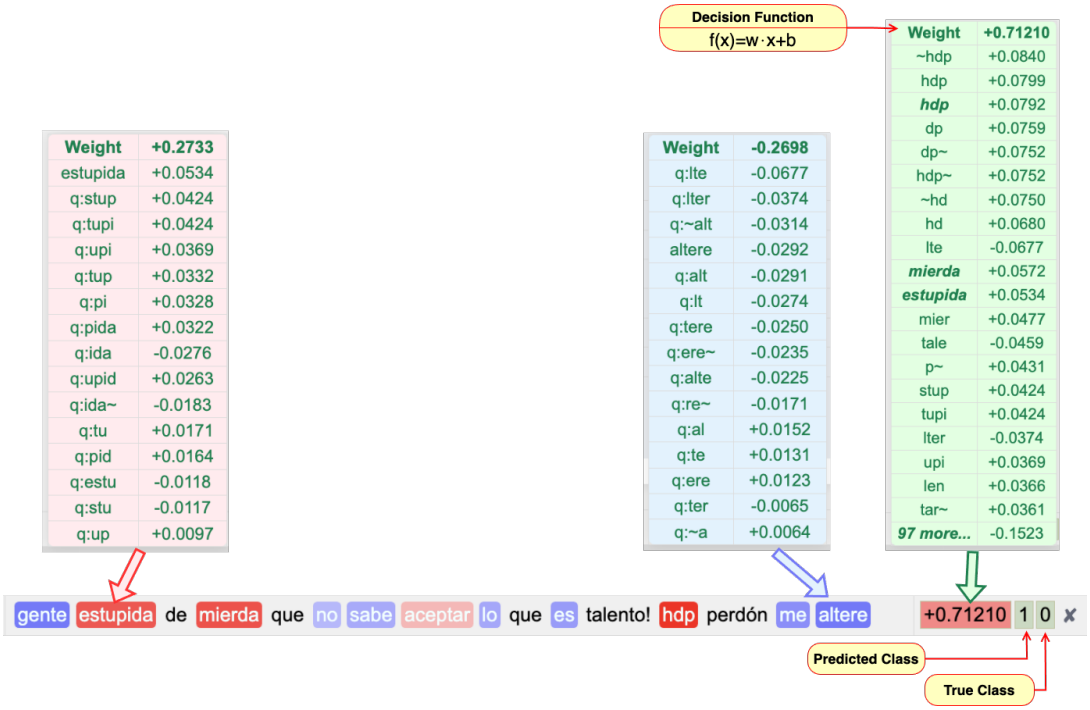


Fig. 3: Example tweet with heatmaps indicating token importance through color intensity: red for positive (offensive) tokens and blue for negative (non-offensive) tokens. Detailed breakdown of the q-grams and their weights for the tokens *estupida* and *altere*, as well as an analysis of all q-grams (117 in total) in the tweet alongside the calculated decision function.

Figure 3 shows an example of explainability using the proposed IGWT framework, applied to a tweet from the MeOffendEs dataset for the binary classification of offensive and non-offensive tweets in Mexican Spanish [21]. Tokens—understood as words or lexical units—are highlighted according to their contribution to each class, as determined by their coefficients. These coefficients are obtained through the q-gram tracing mechanism, which aggregates the individual contributions of all q-grams that make up each token. Offensive tokens appear in red, non-offensive ones in blue, while those with low or null relevance remain unhighlighted, allowing the most influential lexical units to visually stand out.

Given its color intensity, it is observed that tokens like *estupida* and *mierda* have a high contribution to the offensive class, while *altere* contributes negatively. Words such as *perdón* do not meet the relevance

Fig. 4: Visualization of a specific instance for user profiling. This instance displays the grouped tweets for a user and the contribution of each token, facilitating the interpretation of the prediction. In this way, the IGWT framework allows for the explainability analysis of all instances in the dataset. Appendix B includes a representative segment featuring multiple tweet instances processed by the framework.

threshold, possibly due to their use in sarcastic contexts in Mexican Spanish. Additionally, the token *talento!* has a high local Tf-Idf weight but a low global contribution.

The decision function value indicates that the tweet is offensive. A green table breaks down the n most influential q-grams, including those corresponding to complete words, such as *gente* and *mierda*. This explainability exercise is complemented by specific tables that highlight how q-grams, even with a higher contribution than certain tokens, influence the classification, providing greater transparency in the model.

Just as figure 3 presents a specific tweet for a classification task, figure 4 provides a detailed visualization of the grouped tweets for an author in a profiling task. A heatmap is used to highlight the most important tokens according to the prediction made by the surrogate model, indicated by the letter P. The letter T represents the true label, and the decision function is also displayed according to its importance. These elements enhance the explainability of the decision.

It is important to emphasize that figures 3 and 4 present only a fragment of the visualization for a single instance of their respective tasks and datasets. However, unlike other interpretability tools that are focused to global explanations or single-instance interpretation with a small number of features, the proposed framework is capable of displaying all instances in the dataset under examination, as shown in appendix B. Additionally, the significance level of the highlighted tokens is controlled by a threshold parameter applied to the coefficients to be considered, while the number of instances displayed is regulated by a decision function threshold parameter.

4.2 Experimental Evaluation of Q-Gram Traceability for Token-Level Interpretability

Motivation and Conceptual Basis: Figure 3 shows a representative tweet instance processed by the proposed IGWT framework. The visualization includes a heatmap that highlights two tokens—*estúpida* and *altere*—along with all the q-grams that compose them. This example illustrates how each token's

contribution to the model's prediction is computed by aggregating the weights of its constituent subword units, enabled by the q-gram traceability mechanism. Notably, the sum of the q-gram weights for these tokens exceeds the weight assigned to the token as a single unit, emphasizing that—without this aggregation—they would not reach the threshold required to be considered influential in the model's decision. As a result, their visibility would be reduced, thereby weakening the overall interpretability of the instance. This effect is particularly relevant in short-text classification tasks, such as tweet analysis, where lexical variability and limited context make it more difficult for individual tokens to be recognized as significant.

Limitations of the Traceability Mechanism: Before detailing the experiment, it is important to clarify that the traceability mechanism is constrained to aggregating q-grams that are strictly contained within the boundaries of a given word. It does not incorporate other linguistic properties of q-grams, such as their ability to span across word boundaries or include skip-grams. While this restriction preserves a clear and interpretable mapping between q-grams and lexical units—essential for visual explainability—it may omit contextual or structural patterns that inter-word or skip-based q-grams could capture. Investigating the incorporation of such extensions without compromising interpretability or computational efficiency remains an open direction for future research.

Experimental Objective and Hypothesis: The goal of the experiment is to evaluate whether q-gram traceability improves token-level interpretability by enhancing the measured importance of words within text classification models. The central hypothesis is that mapping q-grams to their original words provides greater explanatory value than relying solely on word-level representations. Specifically, it is hypothesized that the aggregate weight of q-grams corresponding to a given word will exceed the weight assigned to that word when treated as a single unit.

Experimental Setup: The experiment uses a labeled dataset of tweets encompassing diverse themes and text structures to ensure robustness. The preprocessing phase includes two tokenization scenarios:

1. **Scenario 1**: Tokenization is performed using full words.
2. **Scenario 2**: Tokenization is performed using character-level bi-grams, tri-grams, and quad-grams extracted as subword units within each word.

This design allows for a direct comparison between traditional word-based representations and subword-based representations using q-grams, enabling an evaluation of their respective effects on model interpretability.

Modeling and Feature Representation: In both scenarios, TF-IDF is used to vectorize the input data. A linear classifier (LinearSVC) is trained independently on each representation. In Scenario1, model coefficients are assigned to full words; in Scenario2, coefficients are assigned to q-grams.

Aggregation via Traceability: To enable comparison, the traceability mechanism links each word from Scenario1 to its constituent q-grams in Scenario2. The contribution of a word in Scenario 2 is computed by summing the weights (model coefficients) of all q-grams derived from that word. This aggregation allows for a fair comparison of token-level importance across the two representations.

Weight Comparison and Statistical Analysis: A paired t-test is conducted to compare the weights of words in both scenarios. The test assesses whether the aggregated q-gram weights in Scenario2 are significantly greater than the corresponding word-level weights in Scenario1. Metrics such as the mean difference, standard deviation, and p-value are calculated to evaluate statistical significance.

Results and Visualization: Figure 5 presents the results using boxplots for the training and testing sets across three datasets: *CheckWorthiness*[4], *MeOffendEs*[21], and *PoliticEs*[10]. These plots compare the distribution of word weights in both scenarios, using a predefined threshold to identify tokens considered relevant for interpretability. The results show that the q-gram-based representation consistently yields a larger number of tokens above the importance threshold, supporting the hypothesis that traceability enhances the identification of influential words.

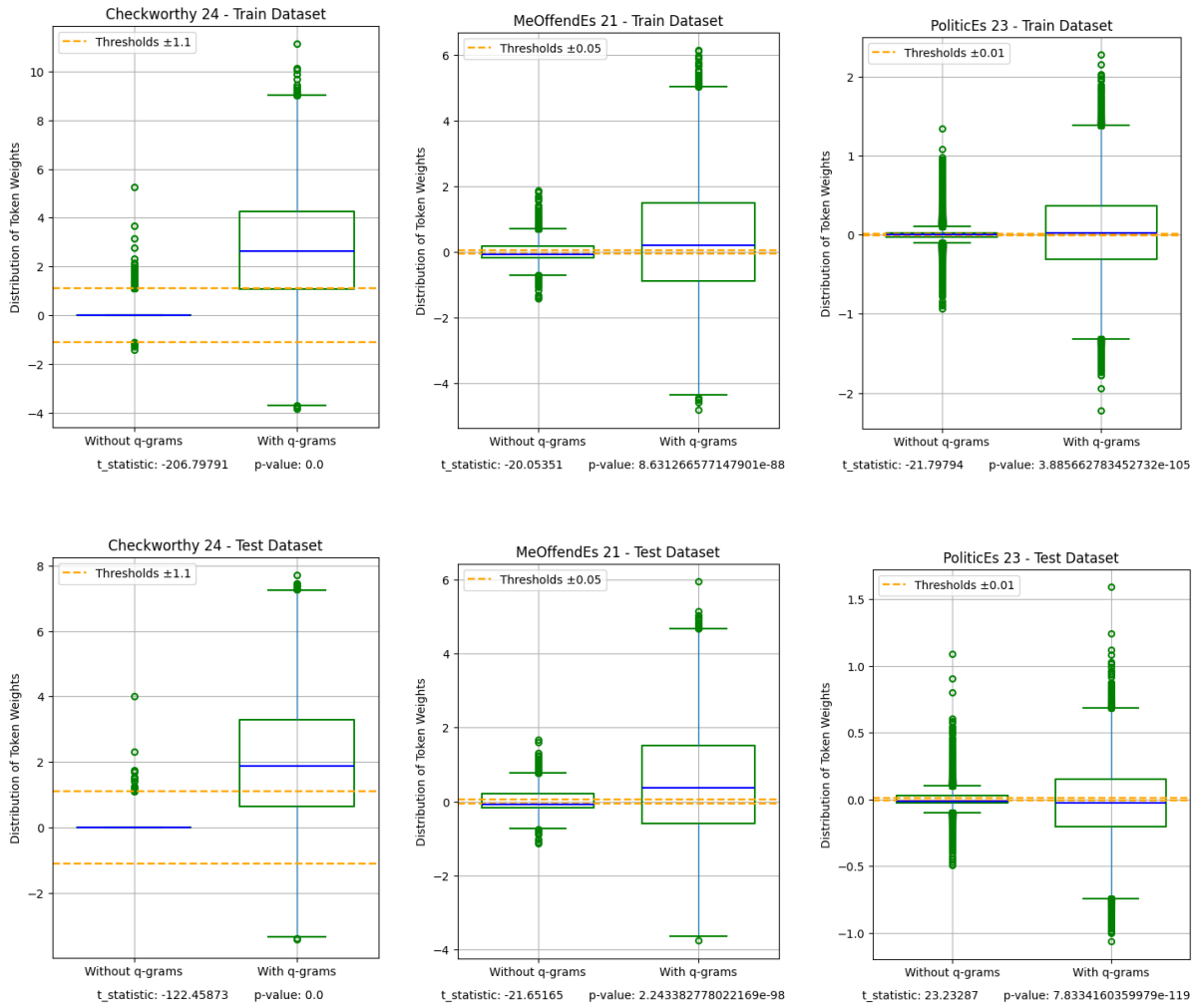


Fig. 5: Comparison of word importance distributions between Scenario 1 (word-level tokenization) and Scenario 2 (q-gram-based tokenization) across three datasets: *CheckWorthiness*, *MeOffendEs*, and *PoliticEs*. The boxplots display the distribution of token weights for training and testing sets. A predefined threshold (dashed line) is used to identify tokens considered relevant for model interpretability. The results show that Scenario 2 consistently yields a greater number of tokens with weights outside the predefined threshold range, supporting the effectiveness of q-gram traceability.

Conclusion: This experiment provides quantitative evidence that the traceability of q-grams to their original words improves both the measured importance and interpretability of lexical units in text classification models. By enabling the aggregation of subword contributions, the model can better recognize and explain the influence of individual terms—particularly in short-text NLP tasks where sparsity and variability are prevalent. These findings validate the effectiveness of the proposed traceability mechanism as a pre-hoc interpretability strategy.

4.3 Comparison with Post-Hoc Interpretability Methods

Post-hoc interpretability techniques such as SHAP and LIME have become standard tools for analyzing the decision-making process of complex models. However, these approaches introduce limitations when applied to high-dimensional and sparse textual data—particularly in the context of short-text classification, such as tweet analysis.

In contrast, the proposed interpretability workflow (Algorithm 1) offers several advantages by integrating explainability directly into the data representation and modeling pipeline. Table **??** summarizes the main differences:

Criterion	IGWT (Proposed)	SHAP	LIME
Interpretability Phase	Pre-hoc (integrated into preprocessing)	Post-hoc	Post-hoc
Token-level Explanation	Yes (via q-gram aggregation)	Approximate	Approximate
Linguistic Traceability	Yes (q-gram → word mapping)	No	Limited
Computational Cost	Low	High	Medium
Scalability to Many Instances	High	Low	Low
Support for Sparse Data	Native (TF-IDF + Linear Model)	Requires sampling	Requires sampling
Model Dependency	Works best with linear models	Model-agnostic	Model-agnostic

Table 1: Comparison Between the Proposed Approach (IGWT) and Post-Hoc Methods

While SHAP and LIME provide general-purpose, model-agnostic interpretability, they often struggle with short texts due to their reliance on perturbation-based sampling, which becomes unreliable in high-dimensional, sparse feature spaces. Furthermore, visualizing thousands of q-grams generated from grouped tweets (e.g., in user profiling tasks) becomes computationally prohibitive under post-hoc frameworks.

In contrast, our approach leverages the strengths of linear models to compute exact token contributions, and introduces a q-gram tracing mechanism that maps subword units back to original words. This not only improves interpretability but also enables a human-readable, context-aware explanation of model behavior—before training the target model—making it suitable for both data exploration and bias correction.

4.4 Applications of IGWT: Enhancing Active Learning and Computational Efficiency

The IGWT explainability model is structured around four key aspects: (1) the fragmentation of words into q-grams as subword units, (2) the computation of individual q-gram weights, (3) the traceability of these q-grams back to their original words, and (4) the visualization of their aggregated contributions to the model's decision through color-coded heatmaps.

In addition, it is computationally efficient, leveraging the coefficients of the self-interpretable linear model. These characteristics allow IGWT to not only enhance explainability but also make it suitable for Active Learning applications, where the model iteratively refines its understanding by analyzing the most informative tokens and instances.

The following sections provide a structured overview of the potential uses of the IGWT highlighting its versatility, advantages, and applications in the fields of Machine Learning (ML) and NLP on tasks related to short-text, as well as its impact on both research and practical implementations.

Efficient Uncertainty Sampling in Active Learning: IGWT enables the rapid detection of instances where the model exhibits high uncertainty or ambiguity, such as in reviews with contradictory terms (e.g., positive and negative words in the same sentence). This facilitates the application of sampling strategies, such as Least Confidence Sampling or Margin Sampling, to prioritize the manual annotation of these cases and focus on the most informative data. Additionally, its low computational cost allows for frequent model retraining, agile incorporation of new instances, and continuous refinement of the model without incurring high computational expenses.

Data Selection with Explainability and Error Analysis: The traceability of q-grams and the heatmaps generated by IGWT make the contributions of words in the model's decisions highly interpretable. This helps experts to:

- Select samples where the model makes errors or overemphasizes irrelevant words.
- Identify patterns in misclassified instances, exposing words that disproportionately contribute to errors related to tokenization or data preprocessing.
- Detect and correct ambiguous instances (hard-negatives), improving model performance and increasing its robustness. For example, in fake news detection, IGWT identifies articles that appear trustworthy but contain misleading terms.

Bias and Unexpected Pattern Detection: IGWT shows words that excessively influence predictions, exposing potential biases in the model. This serves as a guide to:

- Avoid unfair or unbalanced decisions, contributing to the creation of fairer and more ethical models.
- Detect biases toward certain names or terms in hiring tools.
- Identify unexpected patterns that may indicate issues in the dataset or the model.

Domain Adaptation and Model Optimization: IGWT analyzes heatmaps across different datasets to identify key linguistic changes and adjust weights in domain-specific terminology. This ensures that transferred models maintain their accuracy in new contexts. For example:

- A legal text classifier can be adapted from U.S. to European documents.
- In medicine, IGWT identifies technical terms crucial for accurate diagnoses, allowing experts to fine-tune the model according to the specific needs of the domain.

Intelligent Data Augmentation and Rapid Interpretability: IGWT uses heatmaps to identify key words and generate new synthetic training samples. These samples retain the most influential words while introducing variations, improving the diversity and representativeness of the dataset. For example, in sentiment analysis, IGWT generates new reviews using synonyms of key words such as *amazing* or *fantastic*. Additionally, its ability to generate heatmaps and decision explanations quickly and efficiently reduces the time required for model-based decision-making without sacrificing precision.

Conclusion: As demonstrated through its potential applications, the impact of the IGWT framework extends beyond interpretability into practical improvements in model development workflows. The IGWT framework supports early detection of low-quality samples, facilitates uncertainty sampling, and guides data refinement—key aspects in active learning strategies. Its linear structure and low computational overhead further enable efficient processing in large-scale or resource-constrained NLP tasks. These properties position IGWT not only as an interpretability tool, but also as a lightweight, data-centric component that enhances learning efficiency and model robustness.

5 Discussion of Results and Conclusion

The experimental findings validate the core premise of the IGWT framework: that q-gram traceability significantly enhances both the visual and quantitative interpretability of short-text classification models. By decomposing tokens into subword units, assigning individual weights via a linear estimator, and reaggregating them into their original lexical units, IGWT offers a unique pre-hoc interpretability mechanism that maintains model transparency without compromising computational efficiency.

As visualized in Figure 5, the distribution of token weights across multiple datasets shows that the q-gram tokenization consistently produces a greater number of tokens exceeding the interpretability threshold. This result supports the central hypothesis that subword-level representations capture richer semantic and structural information. The statistical analysis reinforces this finding: a paired t-test confirmed that, for the majority of tokens, the sum of q-gram contributions was significantly higher than the weight assigned to the same token in its original form. For example, across datasets such as *MeOffendEs* and *CheckWorthiness*, mean weight differences were consistently positive, with p-values below 0.01, indicating strong statistical significance.

Additionally, Section 4.4 outlines the broader impact of IGWT beyond static interpretability. The framework's ability to operate efficiently across multiple instances, identify influential or problematic tokens, and guide data refinement positions it as a valuable asset in active learning workflows. Unlike post-hoc methods, which struggle with high-dimensional feature spaces and incur significant computational costs, IGWT integrates interpretability directly into the modeling workflow. This integration enables tasks such as uncertainty sampling, error analysis, bias detection, and domain adaptation to be addressed from within the framework itself—facilitating more informed model development.

In summary, the IGWT framework not only achieves its original goal of improving visual interpretability in NLP tasks involving short-text but also establishes itself as a lightweight, scalable, and adaptable tool. Its pre-hoc design, grounded in traceable subword features and interpretable linear modeling, makes it especially suited for real-world NLP scenarios where transparency, efficiency, and iterative refinement are essential.

6 Future Work

While the IGWT framework has shown promising results in enhancing interpretability and supporting active learning strategies, several directions remain open for future exploration.

First, further research is needed to assess the impact of q-gram traceability on downstream NLP tasks performance, particularly in terms of accuracy, f1-score, and robustness across imbalanced or noisy datasets. A systematic evaluation would help quantify the trade-offs between interpretability and predictive power.

Second, expanding the traceability mechanism to account for filtered or removed tokens would improve the framework's applicability in data cleaning and preprocessing workflows. This would require integrating information about the role of absent features in the model's decision-making process—an aspect typically overlooked in current approaches.

Third, while the current implementation is designed for linear models due to their interpretability and efficiency, adapting IGWT to work with neural architectures (e.g., CNNs or transformers) could allow for broader applicability in more complex NLP tasks. This adaptation would involve developing surrogate mechanisms to extract token-level contributions from non-linear decision boundaries.

Finally, future work could explore additional properties of q-grams, such as cross-token spans and skip-grams, to capture richer morphological or contextual patterns. Incorporating these features in a controlled and explainable manner could further enhance the model's ability to handle lexical variability without sacrificing interpretability or computational efficiency.

Availability of Resources

The source codes for the Interpretability by Gram-Weighted Tracing (IGWT) framework, along with reproducible experiments, are publicly available in the GitHub repository at:

Not yet...

7 Acknowledgements

The first author thanks to MSc Rodrigo Dominguez and Professor Jacques Savoy for their invaluable comments and suggestions.

References

1. AminiMotlagh, M., Shahhoseini, H., Fatehi, N.: A reliable sentiment analysis for classification of tweets in social networks. Social network analysis and mining **13**(1), 7 (2022)
2. Anjomshoae, S.: Context-based explanations for machine learning predictions. Ph.D. thesis, Umeå University (2022)
3. Ashfaq, F., Bajwa, I.S.: Natural language ambiguity resolution by intelligent semantic annotation of software requirements. Automated Software Engineering **28**(2), 13 (2021)
4. Barrón-Cedeño, A., Alam, F., Chakraborty, T., Elsayed, T., Nakov, P., Przybyła, P., Struß, J.M., Haouari, F., Hasanain, M., Ruggeri, F., et al.: The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In: European Conference on Information Retrieval. pp. 449–458. Springer (2024)
5. Biswas, B., Mukhopadhyay, A., Kumar, A., Delen, D.: A hybrid framework using explainable ai (xai) in cyber-risk management for defence and recovery against phishing attacks. Decision Support Systems **177**, 114102 (2024)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)
7. Chen, L.M., Xiu, B.X., Ding, Z.Y.: Multiple weak supervision for short text classification. Applied Intelligence **52**(8), 9101–9116 (2022)
8. Do, V.H., Canzar, S.: A generalization of t-sne and umap to single-cell multimodal omics. Genome Biology **22**(1), 130 (2021)
9. Dogra, V., Singh, A., Verma, S., Kavita, Jhanjhi, N., Talib, M.: Understanding of data preprocessing for dimensionality reduction using feature selection techniques in text classification. In: Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021. pp. 455–464. Springer (2021)
10. Garcia-Díaz, J.A., Jiménez Zafra, S.M., Martín Valdivia, M.T., García-Sánchez, F., Ureña López, L.A., Valencia García, R.: Overview of politices at iberlef 2023: Political ideology detection in spanish texts (2023)

11. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)
12. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
13. HaCohen-Kerner, Y., Miller, D., Yigal, Y.: The influence of preprocessing on text classification using a bag-of-words representation. PloS one **15**(5), e0232525 (2020)
14. Heidary, K., Atluri, V., Bland, J.: Performance evaluation of machine learning algorithms in reduced dimensional spaces. Journal of Cyber Security **6**(1), 69–87 (2024). https://doi.org/10.32604/jcs.2024.051196
15. Karajeh, O., Lourentzou, I., Fox, E.A.: Multi-view graph-based text representations for imbalanced classification. In: International Conference on Theory and Practice of Digital Libraries. pp. 249–264. Springer (2023)
16. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR) **54**(3), 1–40 (2021)
17. Pattanayak, P.K., Tripathy, R.M., Padhy, S.: A semi-supervised approach of short text topic modeling using embedded fuzzy clustering for twitter hashtag recommendation. Discover Sustainability **5**(1), 56 (2024)
18. Peeters, J., Thas, O., Shkedy, Z., Kodalci, L., Musisi, C., Owokotomo, O.E., Dyczko, A., Hamad, I., Vangronsveld, J., Kleinewietfeld, M., et al.: Exploring the microbiome analysis and visualization landscape. Frontiers in Bioinformatics **1**, 774631 (2021)
19. Peng, H., Pavlidis, N., Eckley, I., Tsalamanis, I.: Subspace clustering of very sparse high-dimensional data. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 3780–3783. IEEE (2018)
20. Pintas, J.T., Fernandes, L.A., Garcia, A.C.B.: Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review **54**(8), 6149–6200 (2021)
21. Plaza-del-Arco, F.M., Casavantes, M., Escalante, H., Martín-Valdivia, M.T., Montejo-Ráez, A., Montes-y-Gómez, M., Jarquín-Vásquez, H., Villaseñor-Pineda, L.: Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)
22. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
23. Salih, A., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Menegaz, G., Lekadir, K.: Commentary on explainable artificial intelligence methods: Shap and lime. arXiv preprint arXiv:2305.02012 (2023)
24. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O.S., Villaseñor, E.A.: A case study of spanish text transformations for twitter sentiment analysis. Expert Systems with Applications **81**, 457–471 (2017)
25. Tellez, E.S., Moctezuma, D., Miranda-Jiménez, S., Graff, M.: An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems **149**, 110–123 (2018)
26. Van Otten, N.: L1 and l2 regularization explained, when to use them & practical how to examples. `https://spotintelligence.com/2023/05/26/l1-l2-regularization/` (May 2023), last accessed: Octubre 24, 2024
27. Villa-Pérez, M.E., Trejo, L.A., Moin, M.B., Stroulia, E.: Extracting mental health indicators from english and spanish social media: A machine learning approach. IEEE Access **11**, 128141–128152 (November 2023). https://doi.org/10.1109/ACCESS.2023.3332289
28. Vimbi, V., Shaffi, N., Mahmud, M.: Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. Brain Informatics **11**(1), 10 (2024)
29. Wang, Y., Wang, S., Yao, Q., Dou, D.: Hierarchical heterogeneous graph representation learning for short text classification. arXiv preprint arXiv:2111.00180 (2021)
30. Wu, H., Chen, Y., Zhu, W., Cai, Z., Heidari, A.A., Chen, H.: Feature selection in high-dimensional data: an enhanced rime optimization with information entropy pruning and dbscan clustering. International Journal of Machine Learning and Cybernetics pp. 1–44 (2024)
31. Yadav, A., Patel, A., Shah, M.: A comprehensive review on resolving ambiguities in natural language processing. AI Open **2**, 85–92 (2021)
32. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology **15**(2), 1–38 (2024)

A Appendix: Example of JSON Structure Generated by the IGWT

Basic and trimmed example of an instance from a German tweet dataset.

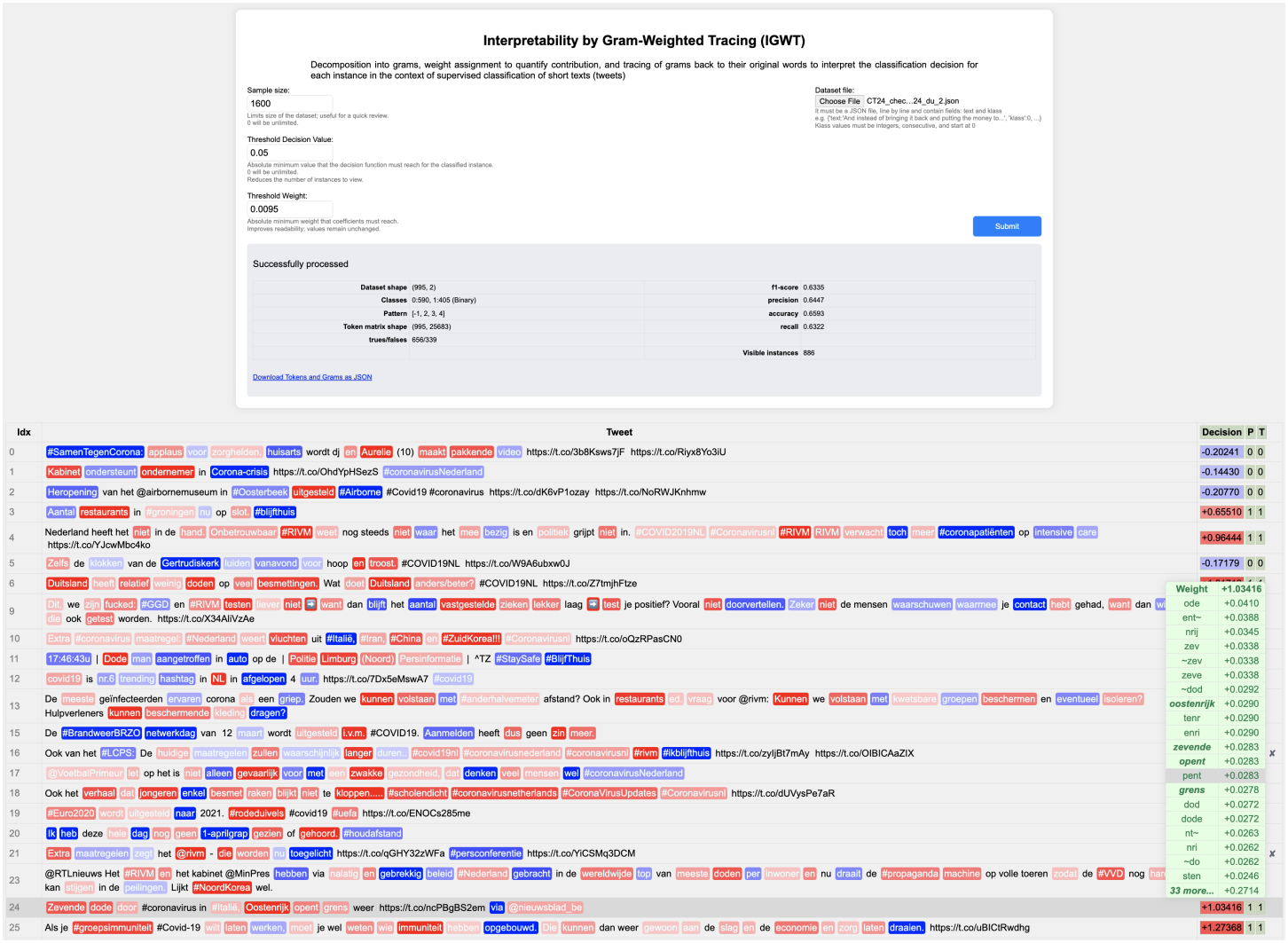

```
{
    "101": {
        "prediction_klass": "1",
        "decision_value": "0.6579841325",
        "true_klass": "1",
        "data": {
            "restaurants": {
                "weight": 0.7325947434,
                "grams": {
                    "q:rest": 0.0358046721,
                    "q:taur": 0.0487939050
                },
                "color": "#FF0000"
            },
            "groningen": {
                "weight": 0.0179637958,
                "grams": {
                    "q:inge": 0.0212614873,
                    "q:gro": 0.0107364526
                },
                "color": "#FFC6C6"
            }
        }
    }
}
```


Explanation of the Example:
The instance `"101"` represents a classification prediction with:

- **Instance "101":**
  - **prediction klass:** The predicted class by the model is "1".
  - **decision value:** The decision value calculated for this instance is "0.6579841325", indicating the confidence level in the prediction.
  - **true klass:** The actual class of the tweet is "1", useful for comparative evaluation.
- **data (Details at the word and q-gram level):**
  - **Word "restaurants":**
    - **weight:** Total weight of the word is 0.7325947434, representing its contribution to the decision function.
    - **grams:**
      - **q:rest:** Contributes a weight of 0.0358046721.
      - **q:taur:** Contributes a weight of 0.0487939050.
    - **color:** Assigned color is #FF0000, which reflects the predicted class, with the intensity indicating the magnitude of the contribution to the prediction, facilitating interpretive visualization on the web interface.
  - **Word "#groningen":**
    - **weight:** Total weight of the word is 0.0179637958, representing its contribution to the decision function.
    - **grams:**
      - **q:inge:** Contributes a weight of 0.0212614873.
      - **q:gro:** Contributes a weight of 0.0107364526.
    - **color:** Assigned color is #FFC6C6, used to facilitate interpretive visualization on the web interface.

B Appendix: User Interface

User Interface of the IGWT Web Tool for Tweet Interpretability



Screenshot of the IGWT Web Service User Interface. It displays the input data: a dataset of tweets in German, the sample size, and the established thresholds. The metrics obtained with the LinearSVC surrogate model and the link to download the complete structure are clearly visible. In the instance space, 25 tweets out of the 886 filtered from a total of 995 are showcased. On the left, the breakdown of tokens and q-grams for instance number 24 is highlighted in green. Each instance features a color code varying according to the class and the intensity defined by the corresponding coefficient.