# POLITECNICO DI TORINO

**Master's Degree in Computer Engineering**

Master's Degree Thesis

# An innovative methodology to experimentally compare explainable AI solutions for Natural Language Processing


**Supervisors**
Prof. Tania CERQUITELLI
Ph.D Salvatore GRECO

**Candidate**
Victor SINAPI

Academic Year 2020/2021

*A mamma e papà,*
*A chi non ha trattenuto le lacrime,*
*A chi le ha trattenute,*
*A chi mi vuole bene.*

**Abstract**

In recent years there has been a great development of artificial intelligence (AI). From agriculture to finance passing through healthcare, the potential advantages of using these algorithms are extremely high. Our life is influenced daily by AI decisions, just think of the recommendation systems for films and TV series, suggestions for purchases in an e-commerce or even the mechanisms of targeted advertising. But there is an important distinction to be made, although the performance of an artificial intelligence is very high there are areas in which for various reasons the AI decision cannot be accepted.

The reason for this difference is due to the construction of artificial intelligence, they work like black boxes, take an input and return an output, the results provided are often excellent, there is no knowing why a decision has been made, a fundamental element in many sectors.

For various reasons, the demand for interpretability and explainability of artificial intelligence models has increased, an open question is still how to evaluate the goodness of an explanation and how different models of explanation can be compared.

The goal of my thesis work was to define a methodology to make a comparison between various systems of explanation of artificial intelligence models, in the field of natural language processing, both in quantitative and qualitative terms. I determined what were the most appropriate criteria to compare explanations, I defined metrics to be able to quantitatively measure these criteria. I also proceeded to measure the qualitative criteria concerning the explanation through a survey.

I have applied this methodology on three different explanation frameworks, respectively LIME, T-EBAnO and Shap. For each of these comparison experiments were performed on three different tasks and datasets, in order Sentiment Analysis on Movie review, Toxicity Detection on Comment and Topic Classification on News Article.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Artificial intelligence is not just a fashionable term, nowadays its influence in our life is greater than ever: from autonomous driving, to the selection of the Netflix catalog, there are more and more machine learning algorithms that take decisions for us in order to improve our user experience. The sectors in which this is applied are the most diverse, but despite the performance of this in terms of accuracy are excellent, sometimes even exceed the performance of a human being, there are many sectors in which the applicability of these algorithms is still to be limited.

The reasons for these limitations are to be found in the basic functioning of the machine learning algorithms, basically they are black boxes, they take an input and return an output, although this is most often correct, we do not know on the basis of what it was elaborated, often fundamental information. "The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks" Doshi-Velez and Kim [2017].
Crucially, therefore, that the model not only provides the prediction but also **how** it arrived at it, a correct prediction is only a partial solution to the original problem. The following reasons have increased the demand for interpretability and explainability of artificial intelligence models.

**Human Curiosity**: When something unexpected happens, our environment's mental model updates, finding an explanation for the unexpected event. For example, if a person feels sick every time he eats red berries. He will update his mental model and decide to avoid them. When using opaque

machine learning models in research, scientific findings are not revealed if the predictions are not explained. Understanding why certain behaviors and predictions were created by machines is an important step in learning, which can be achieved with interpretability and explanations.

**Detecting bias**: A common mistake of machine learning models is to inherit a bias from the training dataset. We absolutely do not want to have a model that discriminates against underrepresented minorities or that in any case makes correct predictions for the wrong reasons. In this case, interpretability becomes a powerful tool for detecting these errors.
A famous example is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), is an ML implementation to determine the risk of reiteration of a crime by offenders COMPAS has repeatedly expressed a human-like bias towards race, incorrectly predicting a double the relapse rate in black people compared to white people. That is, the rate of black false positives is double that of whites Fuchs [2018].
**Model improvement**: Audit and debug are two very useful operations for improving a machine learning model, but to be performed they require the model to be interpretable. Even in low-risk tasks such as a movie recommendation system, when things go wrong, having an interpretation of the prediction can help determine the cause of the problem.

**Social Acceptance**: The lack of information related to the decision-making process of an AI model is a factor that discourages people from trusting and using the models in higher-risk tasks, such as the medical, legal or financial fields. If a machine learning algorithm advises a user to use a certain drug, he or she will be reluctant to trust the choice if he does not know the reasons for it. Granting or not granting a mortgage is a decision that could be delegated to an artificial intelligence but the user must have a reason for the outcome.

These are the reasons that push more and more the diffusion of eXplainable Artificial intelligence (xAI), systems that aim to push artificial intelligence towards a more transparent and understandable dimension, while keeping the performances unchanged. From the research point of view, although the first discussions on the subject date back to decades ago, we have had a surge in recent years also following the entry into force in May 2018 of the General Data Protection Regulation (GDPR) which in article 22 establishes

that natural persons have the right not to be subjected to decision-making processes based exclusively on automated processes (including profiling) and furthermore the criteria for reaching such decisions must be disclosed in order to guarantee the right of objection.

The main objective of my work is to perform a comparison between different methods of eXplainable Artificial Intelligence to determine the strengths and weaknesses of each and which ones are best to use in the relevant situation. I focused on artificial intelligences that perform text classification tasks. I have considered different algorithms, specifically Lime, Shap, T-EBAnO, generating for each of them different explanations on different tasks. I defined some metrics to evaluate the latter, after which the results of the texts were compared to evaluate the different approaches.

To present the work done, the thesis was organized into several chapters. Therefore, the next chapters will be organized as follows:

- Chapter 2 - State of the art
  This chapter shows an overview of the current state of the literature of eXplainable Artificial Intelligence, regarding the different algorithms and techniques that are used to explain a model, then going on to show the current state of the art on the possible approaches to measurement methodologies for an explanation.

- Chapter 3 - Proposed Methodology
  In this part the methodology used to conduct the experiments will be shown, the criteria that have been chosen to make the comparison and the metrics have been used will be presented.

- Chapter 4 - Experimental Results
  In this chapter we will show the configuration of the experiments and what were the results obtained

- Chapter 5 - Conclusion and future work

# Chapter 2

# State of the art

This chapter shows an overview of the current state of the literature of eXplainable Artificial Intelligence, regarding the different algorithms and techniques that are used to explain a model, then going on to show the current state of the art on the possible approaches to measurement methodologies for an explanation.

## 2.1 Taxonomy of Interpretability Methods

There are several criteria by which a method for the interpretability of a machine learning algorithm can be categorized

### 2.1.1 Intrinsic vs Post-Hoc

This criterion discriminates the explanations between those that have been obtained through internal mechanisms of the ML model that we want to explain (intrinsic) and those obtained by applying a method after model training (post hoc).Molnar [2019]
Lipton states that these criteria answer two different questions, intrinsic tells us *how the model works*, while post-hoc tells us *what else can the model tell us*. Lipton [2017]
When we are dealing with machine learning models that are interpretable thanks to their structure, we refer to intrinsic interpretability, a classic example can be sparse linear models or decision trees. This in-model interpretability can be achieved through imposition of constraints on the model, such as sparsity, monotonicity, causality, or physical constraints that come

from the domain knowledge Rudin [2019].
Post-hoc interpretability, on the other hand, concerns the use of explanation methods, performed only after model training. Usually these methods are not linked to the main model of artificial intelligence.

### 2.1.2 Model-specific vs model-agnostic

The distinction between model agnostic methods and model specific methods is also of fundamental importance. This criterion can only be applied to post-hoc methods. It is determined by the visibility that the explanation method has. The model specific interpretation tools are those that in order to elaborate their explanation require in addition to the input and output of the machine learning model, also the internal components of the model itself, for example the interpretation of regression weights in a linear model is a model -specific interpretation, since the interpretation of intrinsically interpretable models is always model-specific Molnar [2019]. Given this peculiarity, a model specific method is not applicable to all models.
On the contrary, the model agnostic methods have no visibility inside the model, to generate the explanation they only need to know the input and output of this, nothing more. They must be totally decoupled from the ML model then by definition, these methods cannot have access to the model inner workings, such as weights or structural information.

### 2.1.3 Results of explanation methods

This criterion allows us to distinguish a method of explanation, observing the type of result that is proposed to us Molnar [2019].
The results cited below represent the vast majority of interpretability methods of a machine learning model, although there may still be other ways to provide an explanation of an ML model, such as rule sets, question-answering or an explanation written in natural language.

- **Feature summary statistic**: Several interpretation methods provide summary statistics for each feature. Often it is returned a single number per feature, such as feature importance, in other cases the result is more complex, such as the pairwise feature interaction strengths, which consist of a number for each feature pair.

Often the feature summaries make sense if they are visualized, losing relevance if presented in other ways, e.g., partial dependence plots are not intuitive if presented in tabular format. Carvalho et al. [2019]

- **Model internals**: This is the category where explanations of intrinsically interpretable models fall. An example of this kind can be the learned tree structure (the features and thresholds used for the splits) of a decision tree or the weights of a linear model.
  It may happen that a method has an overlap between feature summary statistic and model internals. In the case of a linear model, for example, the weights are at the same time summary statistics for the features and model internals.

- **Data point**: There are several methods that work more on unstructured data, such as images or text, which return data points as an explanation.
  These methods require for efficient operation that data points have meaning and can be interpreted by themselves

- **Intrinsically interpretable model**: A final solution to provide an explanation for a black box model is to generate a surrogate model that is inherently interpretable.
  The explanation of the surrogate model will then be a reference for the explanation of the original model

- **Local vs Global**: With this criterion we evaluate the scope of the explanation method, distinguish whether the instrument explains the entire ML model (global) or only a single decision of the latter (local).

## 2.2 Scope of Interpretability

The prediction process can be divided into different portions, based on which of these an interpretation method wants to explain we can determine its scope.

### 2.2.1 Algorithm Transparency

Algorithm transparency mainly needs to know the algorithm, without taking into account the data or learned model, answering the question How does

the algorithm create the model? Carvalho et al. [2019]. An example is the ordinary least squares method.
In particular Algorithm transparency is about how the algorithm learns a model from the data and what kind of relationships it can learn from it. This refers to how the algorithm (which generates the model itself) works but not to the specific model that is learned in the end and not to how individual predictions are made.

### 2.2.2 Global, Holistic Model Interpretability

At this level we refer to a global understanding of how the model makes decisions, of a holistic type through the analysis of all the components and all the parameters learned from the model. In particular, it is necessary to understand the distribution of the model output, analyzing in detail how the driven model makes its decisions, in order to be able to interpret a global model. Carvalho et al. [2019]
This kind of interpretability in practice is extremely difficult to achieve Molnar [2019]

### 2.2.3 Global Model Interpretability on a Modular Level

While achieving global interpretability on a holistic level is very difficult, it is relatively easy for some models to consider this on a modular level. For example, we can think of linear models for which we can easily interpret the weights, or for a decision tree it is easy to understand the branches. In this case the question we want to answer is how do parts of the model affect predictions Molnar [2019]. Obviously this type of approach is not suitable for those models with too much dimensionality of features or with opaque features. Honegger [2018]

### 2.2.4 Local Interpretability for a Single Prediction

In this case we analyze in detail a single instance of an explanation, trying to answer the question "Why did the model make a certain prediction for an instance?". Looking locally, it is possible to reduce the complexity of the task, as at this level the explanation depends linearly on a few characteristics. Carvalho et al. [2019]

### 2.2.5 Local Interpretability for a Group of Predictions

To explain a group of predictions, two opposite approaches are available, in the first the whole is considered globally by applying global methods, in the other case the predictions are considered individually, after which the results are aggregated. Lipton [2019]

## 2.3 Properties of Explanation

In order to be able to judge how good a method of explanation or the explanation itself is, we can rely on the criteria introduced by Robnik-Sikonja and Bohanec listed below. A quantitative measurement of these is still being researched. Robnik-Šikonja and Bohanec [2018]

### 2.3.1 Properties of Explanation Methods

- **Expressive Power**: refers to the structural output that is generated by the explanation method. Examples can be if-then rules, a weighted sum, decision trees, natural language or something else.

- **Translucency**: This criterion indicates to what extent the explanation model looks inside the model, for those model specific methods that require among the parameters internal components of the model there will be a high level of translucency, on the contrary for model agnostic methods that require only input and output, the translucency level will be 0. The advantage of using translucent methods is that the explanation is given based on more data, potentially it will be better.

- **Portability**: Describes the range of models for which an explanation method can be applied. Tendentially this criterion is specular and opposite to the previous one, methods with low translucency will be much more transportable, on the contrary translucent methods will be applicable on fewer models.

- **Algorithmic Complexity**: This criterion is extremely relevant within the applicability of an explanation method, particularly when computation time is a bottleneck in generating explanations. It focuses on describing the computational complexity of the explanation method.

### 2.3.2 Properties of Individual Explanations

- **Fidelity**: This is certainly one of the fundamental criteria to be analyzed when evaluating whether an explanation is good or not. Evaluate how closely the explanation approximates the prediction of the machine learning model. Are the characteristics identified by the explanation the ones on which the model actually based its choice?

- **Comprehensibility**: This component is also of fundamental importance, it is relative to the human, in particular it investigates the extent to which a person is able to fully understand the explanation. It is extremely difficult to measure, but at the same time very important for an explanation method.

- **Consistency**: Evaluate to what extent two explanations performed with the same methodology obtain similar results, when they are performed on two models driven by the same task and which produce similar results.

- **Stability**: As opposed to consistency comparing explanations between models, stability focuses on evaluating two similar instances for a fixed model. In this case, the similarity between the features identified is evaluated. It is always appreciated that an explanation method has high stability. Low stability represents another variance of the method of explanation.

- **Accuracy**: This criterion is taken into consideration for methods of explanations that associate a surrogate model as an explanation to the black box model. In this case we wonder how well unseen data is predicted? For machine models that have a high accuracy, a high accuracy of the explanation is required, the opposite can be accepted that the explanation has a low accuracy if that of the model is too.

- **Certainty**: This criterion is related to the certainty of the machine learning model, often in fact the model in addition to returning the prediction as output, also offers a confidence value of the latter.

- **Degree of Importance**: This criterion evaluates to what extent the explanation method has taken into consideration the features that had greater weight for the machine learning model

- **Novelty**: This criterion is very close to the concepts of efficiency and satisfaction. Describe the subjective degree of novelty of information provided to the explainee. Langer et al. [2021]

- **Representativeness**: This criterion evaluates in what measure an explanation method evaluates a model, if it can explain the black box model in its entirety or if only one instance of prediction is explained.

## 2.4 Human-friendly Explanations

Since humans are the recipients of all explanations, it is important to analyze the factors that make a good explanation human-friendly. This means, for instance, making sure that the explanation is easy to understand and not complex. For his study, Miller surveyed various publications on the subject of explanations. He found that most of the work on this topic mainly uses the researchers' intuition on what constitutes an appropriate explanation for humans.

- **Contrastiveness**: Most of the time, humans do not ask why a particular prediction was made. Instead, they tend to think about the factors that need to change in order for the prediction to be successful. An explanation that shows a contrast between a reference point and an instance is preferable. However, this explanation is also considered application-dependent since it requires a reference object.

- **Selectivity**: Instead of covering the entire list of causes of an event, people prefer to select one or two main explanations. This phenomenon, known as theRashomon Effect, occurs when people prefer to select different causes for a prediction.

- **Social**: The concept of explanation states that the context in which the interaction occurs determines the type of explanation that is provided. Generally, the objective of an explanation is to provide a compelling and accurate response to the target audience.

- **Focus on the abnormal**: People tend to focus on certain abnormal causes to explain events. If these were eliminated, the results would have been different. If a feature value of a prediction is abnormal in any sense, then the feature should not be ignored in the explanation.

Even if other features have the same influence as the abnormal one, it should still be included in the explanation.

- **Truthful**: Good explanations are true in the real world. Not all of them are true, and it is important that the explanation is formulated with the intent of making sense.
- **Consistent with prior beliefs of the explainee**: People tend to ignore information that contradicts their prior beliefs. This effect, known as confirmation bias, is associated with a set of beliefs that vary depending on the individual. It is possible to avoid this bias by being truthful when explaining something that is contrary to your prior beliefs.
- **General and probable**: A good cause can explain a lot of events and could be considered a good explanation. However, in the absence of an unusual cause, it is rare for a general explanation to be considered a good one.

## 2.5 Evaluation of Interpretability

Some research has been done in this regard, but there is still no real consensus on what interpretability in machine learning can be and there is no way to measure it. Doshi-Velez and Kim [2017] proposed three different levels to evaluate interpretability:

- **Application level evaluation (real task)**: the explanation must be added to the product and then tested by the end user. As an example, consider fracture detection with a machine learning component that identifies and marks fractures on X-rays. Radiologists will need to test fracture detection software in order to evaluate the model. To ensure that everything goes smoothly, you need a good experimental setup and understand how to evaluate quality. Fundamental to all this is that the human being is good at explaining the decision.
- **Human level evaluation (simple task)**: this assessment is at the simplified application level. These experiments are not conducted by domain experts, but by inexperienced ones. The experiments are therefore made cheaper and it remains easier to find testers. For example, you can show more explanations to a user who will then choose the best one.

- **Function level evaluation (proxy task)**: there is no need to use people. It works best when the model class used is previously evaluated by some other person. Taking an example, it may be known that end users understand decision trees. Therefore, the depth of the tree could be the approximation of the quality of the explanation. The shorter the trees, the better the explainability score will be. A constraint should be added with the predictive performance of the tree remaining good, without decreasing too much compared to a larger tree.

## 2.6 Goals of Interpretability

- **Accuracy**: The goal is to connect the given explanation method with the prediction derived from the model. If not achieved, the explanation would not be useful.

- **Understandability**: This goal is related to how easily an explanation is understood by an observer. Generally, an explanation is not usable if it is not understandable.

- **Efficiency**: This condition shows the time needed for a user to grasp the concept. It is commonly argued that any model is interpretable if it has infinite time.A good explanation should be brief and understandable. It should be able to be understood in a limited amount of time. Generally, the more understandable an explanation is, the more it is likely to be grasped.

## 2.7 Metrics for the measure of the goodness of an explanation

As previously stated in the literature, we do not find a clear reference for establishing criteria for the definition of criteria on which to evaluate the validity of an explanation, nor is it even more difficult to find quantitative metrics for measuring the latter.
Let's see in this section, those that have already been proposed. Specifically those adapted to unstructured data in the domain of natural language processing.

- **Selectivity**: Proposed by Montavon Montavon et al. [2018], with this metric it quantitatively measures the fidelity of an explanation, defined X as the sum of all the features of a text to be analyzed, and f(X) as the prediction function of an artificial intelligence algorithm, we evaluate how quickly f(X) decreases when removing important features determined by the explanation.
  Practically, the value of f (x) is determined on a graph by removing each feature and measuring the area under the curve.

- **Recall**: With this metric we evaluate among all the features considered as influential by the explanation, how many of these were actually relevant, by means of a division. Obviously, in order to calculate this metric, it is necessary to know which characteristics are actually important for the model.We find an application of this in the following study Ribeiro et al. [2016a].

- **Switching point**: It has been proposed for a comparison of explanation methods in the domain of natural language processing Nguyen [2018], it is defined as the quantity of features to be removed in order to change the prediction of an artificial intelligence model.

## 2.8 xAI methods

In this section we will show three different methods of explanation present in the literature, their characteristics and their functioning will be analyzed in detail.

### 2.8.1 LIME

Lime is an acronym for Local Interpretable Model-Agnostic Explanations, Ribeiro et al. [2016b] is a model agnostic explanation method that manages to act on multiple domains, providing a local explanation of the model prediction. The technique attempts to understand the model by by approximating it locally with an interpretable linear model.

It can therefore be performed on any model with minimal effort on the development side, in fact it only requires the implementation of a function

that associates the model's output with the model's input. Provides explanations that are locally faithful within the surroundings or vicinity of the observation/sample being explained.

### How does it works?

In practice, once LIME receives a prediction model and a sample, the following two steps are performed

- **Sampling and obtaining a surrogate dataset**: A totally random input sample perturbation is performed following a normal distribution, by default 5000 perturbations are generated from the sample, and a prediction is performed for perturbed samples.

- **Feature Selection from the surrogate dataset**: Then it uses a feature selection technique to obtain the top important features. This technique is based on the distance of the perturbed sample from the original sample and the relative prediction difference

In the figure you can see an example of the explanation produced by LIME

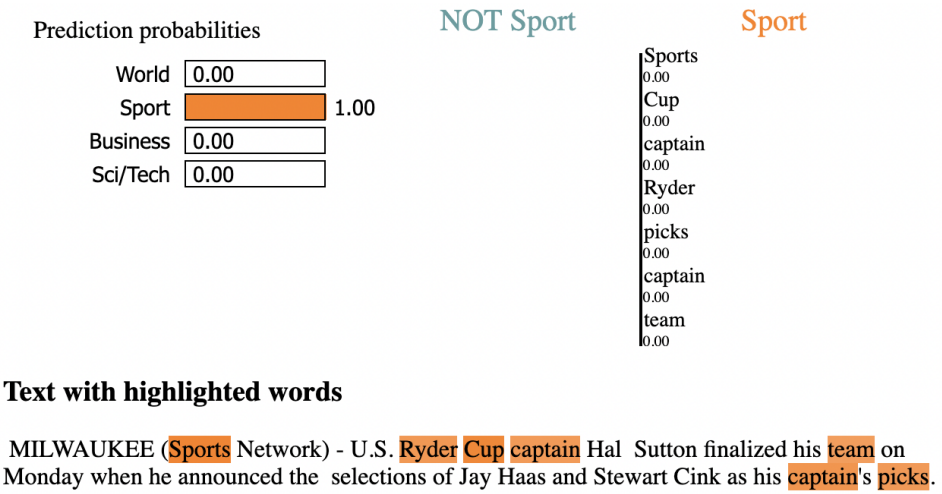


Figure 2.1. Lime Explaination

### 2.8.2 T-EBAnO

T-EBAnO Ventura et al. [2021] is the acronym for Text-Explaining BlAck-box mOdels, consists of a fairly recent explanation framework that allows to obtain both local and global explanations for predictions performed in the domain of Natural Language Processing.

It is a model specific method, although its implementation is not excessively complex from the development point of view, it only requires the implementation of an interface

#### How does it works?

Given an input text and a model, this framework performs three different feature extraction techniques, one model specific and two model agnostic, which are respectively the following:

- **Multi-layer Word Embedding (MLWE)**: This technique, the only one of the three that takes into account the internal knowledge of the model, extracts the features based on the weights they have in the internal layers.
- **Part-of-Speech (PoS)**: In this case, a semantic meaning is given to the words that make up the input text and are considered grouped according to their relative type (for example nouns, adjectives, pronouns, etc.).
- **Sentence-based**: Also in this case the semantic meaning of the input text is valued and the features are considered based on the sentence they belong to

The input text is then perturbed, removing the various groups of features extracted and their weight is determined by evaluating how the prediction varies.

### 2.8.3 SHAP

SHAP Lundberg and Lee [2017] acronym for SHapley Additive exPlanation, also consists of a framework of explanations for black box models. Similarly to LIME it is model agnostic and can be applied on different domains.

### How does it works?

This methodology is based on a principle borrowed from game theory, of shapley values, a concept of solution used to assign a reward to each player present in a coalition, according to the marginal contribution he makes to it. A possible solution to this calculation consists in making an average of all the marginal contributions of the player over all the possible orders of the players present in the coalition.

$\phi(i, v) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} v((\pi, i) \cup \{i\}) - v((\pi, i))$

$\phi(i, v)$ indicates the reward received by the player i

$v$ is the characteristic function (contribution of the set of players in order on the outcome).

$\Pi_N$ is the set of all possible orderings of the elements of N or permutations.

B $(\pi, i)$ is the set of players that precede player i in the order taken into consideration.

By applying this concept in the field of machine learning models, we transform the players into the features we want to extract and define the contributions of these variables as SHAP value.

# Chapter 3

# Proposed Methodology

In this part the methodology used to conduct the experiments will be shown, the criteria that have been chosen to make the comparison and the metrics have been used will be presented.

## 3.1 Definition of metrics

As we saw in the previous chapter, there is no mathematical definition of interpretability, much less a way to measure it has not been uniquely defined. Criteria have been defined in the literature to judge how good an explanation is, on the basis of these a comparison can be made.

It is not clear for these properties how to measure them correctly, so one of the challenges is to formalize how they could be calculated.
What is clear that not all criteria can be measured in the same way, I have therefore decided on several two different criteria strategies to underestimate the. For those objective categories that do not require the presence of a human I have used the so-called Functionally-grounded Evaluation Doshi-Velez and Kim [2017], that is, I have implemented very specific quantitative metrics, while for the more subjective categories I have measured them with human support through the use of a survey.

In this part I present the metrics I have chosen for my work.

### 3.1.1 Percentage of highlighted text

This metric is used to measure the dimensionality of an explanation, it is determined by measuring the ratio between the number of words highlighted and the total words present in the text. For structured data the dimensionality of the explanation is often associated with the cardinality of the latter, but it is not possible to do this in unstructured data. It also partially gives us a measure of the comprehensibility of the explanation.

#### Example - Percentage of highlighted text

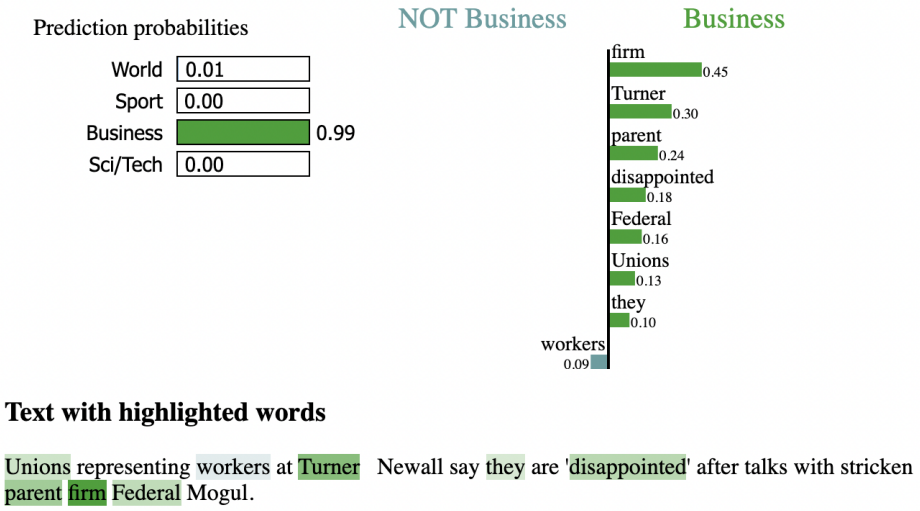


Figure 3.1. Lime Explaination

In the explanation of Lime shown in figure, 7 words out of a total of 14 are highlighted, the percentage of highlighted text will therefore be 50

### 3.1.2 Variation of prediction

The goal of this metric is to measure the fidelity of the explanation, then to determine how much what highlights the explanation was actually decisive for the black box model we are trying to explain.
It is determined by considering the difference between the prediction between the text and the perturbed text. Where for perturbed text we consider the text to which the words that the explainer highlighted have been removed

In my experiments I have considered both absolute and relative variation

**Example - Variation of prediction**

Let's consider the following text:

"The film was very good"
Original text

Suppose the prediction for this text is: positive 80 %
The explanation for this prediction consists of the following set: good
The perturbed text will be as follows:

"The film was very"
Perturbed text

Prediction is now: positive 10In this hypothetical case the absolute prediction variation will be 0.7 while the relative one will be 87.5%

### 3.1.3 Execution time

This metric is determined by measuring the execution time of the explanation algorithm, it gives us a measure of the algorithmic complexity at the time level.

### 3.1.4 Score

This metric was defined by me to measure the fidelity of the explanation, it is a value assigned to the single explanation that varies between 0 and 1. Decreases if too many words are highlighted and increases as the prediction variation increases.

It is defined as the harmonic mean between the complementary of the percentage highlighted words and the variation of the prediction

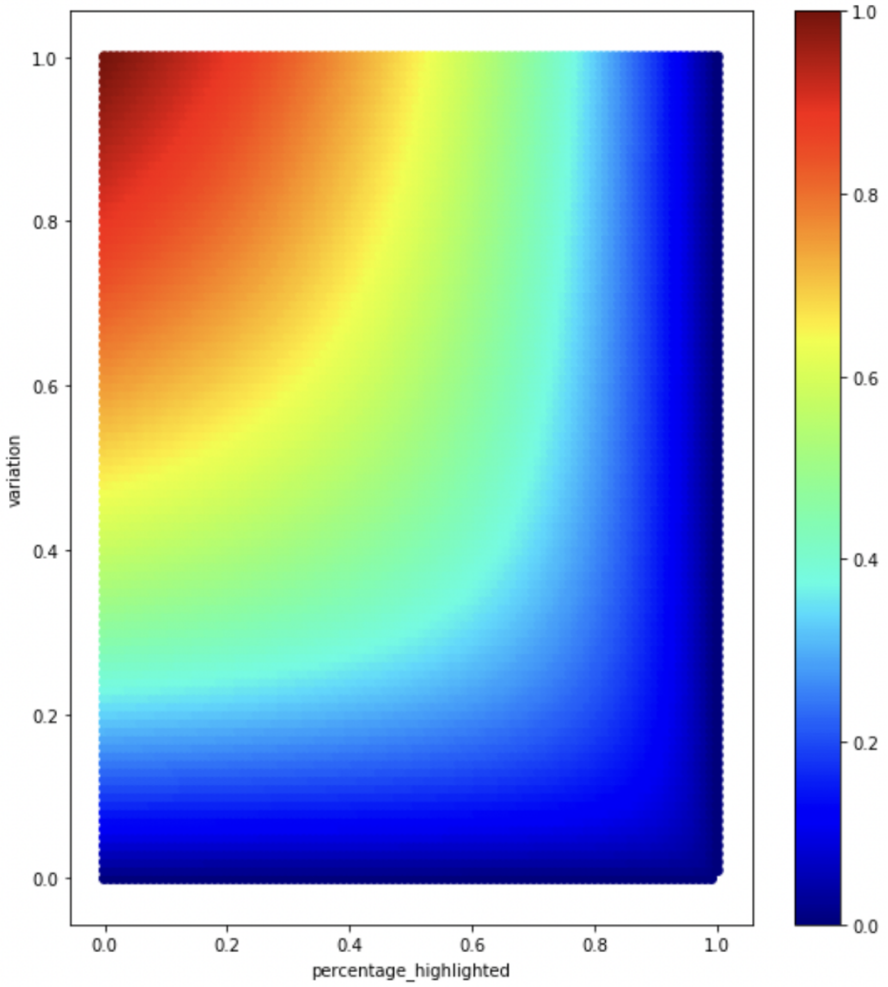


Figure 3.2. Score

$$score = \frac{2}{\frac{1}{1-Percentage\,of\,highlighted\,text} + \frac{1}{1-Variation\,of\,prediction\,relative}}$$

I used this type of formula to emphasize small values and less importance to large ones, the idea of combining the dimensionality of the explanation with the variation of prediction, comes from the works of Montavon et al. [2018] and Nguyen [2018], in both these cases however there is it is an assumption that it is not always possible, ie. that the features of the explanation have a criterion by which they can be ordered, for example in the

case of T-EBAnO this is not possible.

## 3.2 Survey

As we have seen, there are some criteria that cannot be evaluated with automated metrics, some require the presence of a human in order to be evaluated. In my work I have focused on the evaluation of three more subjective criteria, which are the following:

- Clarity: The degree to which the explanation is understood by man, specifically implies that the explanation is unambiguous and that the explanation is presented in a simple and compact form Zhou et al. [2021]
- Trustworthiness: The ability for the explanation to be believed in or accepted by the user as an honest representation or correct description. Sperrle et al. [2021]
- Effectiveness: The degree to which the explanation is successfully conveying the decision-making process of the model. Sperrle et al. [2021]

In order to be able to make a more complete comparison, also evaluating these aforementioned factors, I developed a survey through a questionnaire aimed at specific and non-specific users. The questionnaire with the relative results will be presented specifically in the next chapter.

## 3.3 Comparison framework

In order to facilitate the comparison process, I have developed an architecture that allows to automate it, this starting from the saved explanations manages to generate a visual comparison and a tabular summary for each of the analyzed texts.

### 3.3.1 Visual comparison

The visual comparison focuses on a single prediction, explained by several explainer and shows the following fields:

- Highlighted Text: It is a representation of the text in which all the features relevant for the explanation are highlighted

- Explanations: They are the single features identified
- Perturbed probabilities: The change in absolute probability for the predicted class, between the prediction on the original text and that on the perturbed text
- Time: The time for the elaboration of the explanation

An example follows

**INPUT INFO**

**Original Text**

A giant 100km colony of ants which has been discovered in Melbourne, Australia, could threaten local insect species.

**Original Probabilities** **Original Label**

**lime [ [0.01518, 0.00047, 0.00092, 0.98342] ]**
**ebano [ [0.01518, 0.00047, 0.00092, 0.98342] ]** **4**

| | Lime | T-Ebano (MLWE) | T-Ebano (POS) | T-Ebano (SEN) | SHAP |
|---|---|---|---|---|---|
| Highlighted Text | A giant 100km colony of ants which has been discovered in Melbourne, Australia, could threaten local insect species. | A giant 100km colony of ants which has been discovered in Melbourne , Australia , could threaten local insect species . | giant 100km colony ants which has been discovered Melbourne Australia could threaten local insect species | A giant 100km colony of ants which has been discovered in Melbourne , Australia , could threaten local insect species . | A giant 100km colony of ants which has been discovered in Melbourne, Australia, could threaten local insect species. |
| Explanations | local 0.12235772717221423<br>ants 0.19306331997782009<br>insect 0.2063741487408064<br>species 0.5950059478110611 | Feature 3 - nPIR 0.918 - cover ratio 6/21<br>Feature 7 - nPIR 0.939 - cover ratio 7/21<br>Feature 12 - nPIR 0.904 - cover ratio 9/21<br>Feature 10 - nPIR 0.78 - cover ratio 8/21<br>Feature 14 - nPIR 0.996 - cover ratio 15/21<br>Feature 13 - nPIR 0.018 - cover ratio 12/21<br>Feature 4 - nPIR 0.014 - cover ratio 9/21<br>Feature 8 - nPIR 0.012 - cover ratio 10/21<br>Feature 11 - nPIR 0.01 - cover ratio 11/21 | Feature 6 - nPIR 0.508 - cover ratio 8/21<br>Feature 11 - nPIR 0.175 - cover ratio 10/21<br>Feature 13 - nPIR 0.154 - cover ratio 7/21<br>Feature 12 - nPIR 0.141 - cover ratio 7/21<br>Feature 14 - nPIR 0.137 - cover ratio 7/21<br>Feature 1 - nPIR 0.071 - cover ratio 6/21<br>Feature 8 - nPIR 0.012 - cover ratio 3/21<br>Feature 16 - nPIR 0.012 - cover ratio 5/21<br>Feature 10 - nPIR 0.012 - cover ratio 3/21 | Feature 0 - nPIR 0.742 - cover ratio 21/21 | insect species |
| Perturbed Probabilities | 0.983 -> 0.678 : -0.305 | 0.983 > 0.08 : -0.904 | 0.983 > 0.546 : -0.437 | 0.983 > 0.263 : -0.72 | 0.983 -> 0.968 : -0.015 |
| Time | 108 sec | 1 sec | 1 sec | 1 sec | 2 sec |

Figure 3.3. Example of Visual comparison

### 3.3.2 Overall comparison

This tool, on the other hand, takes care of accumulating a large batch of explanations in a tabular format. For each prediction it shows the input text, prediction and metrics discussed above. Specifically, there are these items:

- Text ID
- Original text

- Original prediction
- Percentage of highlighted text
- Variation of prediction (Absolute)
- Variation of prediction (Relative)
- Time of elaboration
- Score

# Chapter 4

# Experimental Results

In this chapter we will show the configuration of the experiments and what were the results obtained

## 4.1 Model

The reference model with which all the explanations were elaborated was Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. [2018], uses a transformer-based machine learning architecture for natural language processing (NLP).
The reason that motivated this choice is mainly the fact that the tranformer-based architecture used by the model is nowadays one of the most widespread as regards natural language processing.

## 4.2 Datasets and Task

For the experiments of my thesis work I decided to compare the various explanation algorithms on different tasks in order to have a more heterogeneous result, the datasets considered are three IMDB, Civil Comment and AgNews below I present them in detail

### 4.2.1 Sentiment Analysis of IMDB Movie Reviews

#### Task

Sentiment Analysis is part of Text Mining, that is the set of Data Mining techniques aimed at analyzing unstructured texts, in natural language. You look at the written texts, analyzing in particular the level of positivity or negativity, that is their polarity. The main challenge is to try to capture sarcasm, irony and all the other characteristics typical of natural language

#### Dataset

The dataset considered is the Large Movie Review Dataset Maas et al. [2011], it includes 50000 movie reviews, collected by IMDB, separated into 25000 for the train set and 25000 for the test set, as far as sentiment is concerned, 25000 are positive and 25000 are negative. The class (positive or negative) was determined by looking at the score that users of the platform gave to the film, greater than or equal to 7 out of 10 the class is considered positive, less than or equal to 4 out of 10 the class is considered negative, the reviews with an intermediate score and were not included in the dataset.

For the experiments of my work, a set of 100 texts was considered in a random way.

#### Preprocessing

A preprocessing function was applied for all texts in this case quite simple, which was used to remove mentions, URLs, hashtags, special characters, html tags and merge multiple whitespace.

#### Training

The following parameters were used for the training phase:

- number of epoch: 3
- learning rate : $2 * 10^{-5}$

- classification metrics : accuracy
- weight decay: 0.01
- warm up steps: 500

The results obtained at the various eras are shown in the table below

| Epoch | Training Loss | Validation Loss | Accuracy |
|---|---|---|---|
| 1 | 0.315700 | 0.241507 | 0.920920 |
| 2 | 0.167700 | 0.278578 | 0.934560 |
| 3 | 0.044500 | 0.340723 | 0.934440 |

Table 4.1. Training result for the sentiment analysis task

### 4.2.2 Toxicity Detection on Civil Comment

#### Task

In this case the goal is to determine whether a text is clean or not, the discriminant of this classification is the presence within it of various factors such as severe toxicity, obscene, threat, insult, identity attack, and sexual explicit.

#### Dataset

The reference dataset in this is Civil Comment Borkan et al. [2019] includes within a large number of samples, in particular 97320 comments in the test set 1804874 in the train set and 97320 for the validation test. Each comment is associated with 7 discreet attributes ranging from 0 to 1 to indicate the presence within the comment of the elements that make it toxic, specifically toxicity, severe toxicity, obscene, threat, insult, identity attack, and sexual explicit. For the training phase I did not use all the samples, I randomly selected 30000 for the training and 10000 for the test, as a reference label I considered the toxicity attribute which I transformed into a binary value> 0.5 the comment is considered toxic, <= 0.5 the comment is considered non-toxic.

For the comparison experiments I considered 100 comments trying to distribute the toxicity value evenly, specifically I considered 25 with a toxicity value between 0 and 0.25, another 25 with a toxicity value between 0.26 and 0.50, another 25 with a toxicity value between 0.51 and 0.75, finally the last 25 with a toxicity value between 0.76 and 1.

### Preprocessing

In this case the preprocessing phase of the initial text is more complex than the other tasks, as some comments may contain disguised words that need to be reconstructed (example: in many online forums users replace the word "shit" with "sh * t").
So in addition to the common steps, including removing spaces, special characters, clean bad case words and cleaning repeated words it was necessary to add the cleaning steps of rare words, fix misspell words.

### Training

The following parameters were used for the training phase:

- number of epoch: 3
- learning rate : $2 * 10^{-5}$
- classification metrics : accuracy
- weight decay: 0.01
- warm up steps: 500

The results obtained at the various eras are shown in the table below

| Epoch | Training Loss | Validation Loss | Accuracy |
|---|---|---|---|
| 1 | 0.177500 | 0.138713 | 0.950300 |
| 2 | 0.098200 | 0.204556 | 0.958600 |
| 3 | 0.030600 | 0.257283 | 0.956100 |

Table 4.2. Training result for the toxicity detection task

### 4.2.3 Topic Classification on AG News

#### Task

Unlike the two previous tasks discussed above, which deal with a binary classification problem, this is a multiclass classification problem, given a text, we want to determine which category it belongs to among the following World, Sports, Business, Sci / Tech.

#### Dataset

For this task I used AG's news topic classification dataset built by Xiang Zhang Zhang et al. [2016], it includes 120000 journal articles classified for the train set and 7600 samples for the test set. For training the model used in my experiments I used all the samples.

As for the texts used for the comparison experiments, I considered a set of 100 samples taken randomly

#### Preprocessing

Also in this case a preprocessing function was applied for all texts in this case, as for the sentiment analysis task it was quite simple, which was used to remove mentions, URLs, hashtags, special characters, html tags and merge multiple whitespace.

#### Training

The following parameters were used for the training phase:

- number of epoch: 1
- learning rate : $2 * 10^{-5}$
- classification metrics : accuracy
- weight decay: 0.01
- warm up steps: 500

The results obtained at the various eras are shown in the table below

| Epoch | Training Loss | Validation Loss | Accuracy |
|---|---|---|---|
| 1 | 0.184000 | 0.220449 | 0.945000 |

Table 4.3. Training result for the topic classification task

## 4.3 Experimental configuration

In this part I will show how the different explanation frameworks that I have considered in my work have been configured

### 4.3.1 LIME

#### Saving explanation

As for Lime, I needed a way to save the explanation, with the standard version this can be saved to an html file, but since I needed to be able to easily access the fields of the explanation, and also that some information was missing in the explanation fundamental for the work that I should have done (for example the execution time or the variation of prediction), I created a framework that once the explanation was generated, calculated all the values I needed and saved them on a .json file.

Below is an example of how the json file is composed


```
{
   "metadata":{
      "report_id":11,
      "execution_time":316.4436058998108,
      "num_features":9,
      "num_samples":5000
   },
   "input_info":{
      "original_text":" LOS ANGELES (Reuters) - Apple Computer Inc.&lt;AAPL.O&gt; o
      Tuesday began shipping a new program designed to let users  create real-time
      motion graphics and unveiled a discount  video-editing software
```

```
    bundle featuring its flagship Final Cut  Pro software.",
    "original_label":4,
    "original_prediction":[
       9.61708283284679e-05,
       7.305831968551502e-05,
       0.004930940922349691,
       0.9948998093605042
    ]
 },
 "local_explanations":{
    "local_explanations":[
       [
          "software",
          249,
          0.012554812720904321
       ],
       [
          "bundle",
          204,
          0.010660301357927685
       ],
       ...
       ,
       [
          "Apple",
          25,
          0.018262609466637516
       ]
    ],
    "prediction_without_positive":[
       0.00011764620285248384,
       0.0001036649991874583,
       0.011049083434045315,
       0.9887295365333557
    ],
    "prediction_without_negative":[
       9.61708283284679e-05,
       7.305831968551502e-05,
       0.004930940922349691,
       0.9948998093605042
    ]
```

```
   }
}
```

### Explainer Set-Up

The configuration of lime was very simple it only required the implementation of a classifier prediction probability function, which takes a list of d strings and outputs a (d, k) numpy array with prediction probabilities, where k is the number of classes.

### Explainer parameters

The part of choosing the parameters for files was more delicate, there are two very relevant ones, which are the num_samples which is the size of the neighborhood to learn the linear model and the num_features which would be the maximum number of features present in explanation, for choosing the most appropriate parameters. I have carried out several experiments and I have kept those with the best results which are:

- **num_samples** : 5000
- **num_features** : 45% of the features present in the text

The experiments were carried out on the sentiment analysis task, it results from these as it can also be seen in the table and graphs below that num samples has a strong impact on the computational time for the elaboration of the explanation, while num features influences the fidelity of the latter.

| | **5000 samples** | **500 samples** |
|---|---|---|
| **Score** | 0.43 | 0.27 |
| **Time** | 300 s | 78s |

Table 4.4. LIME parameters comparison

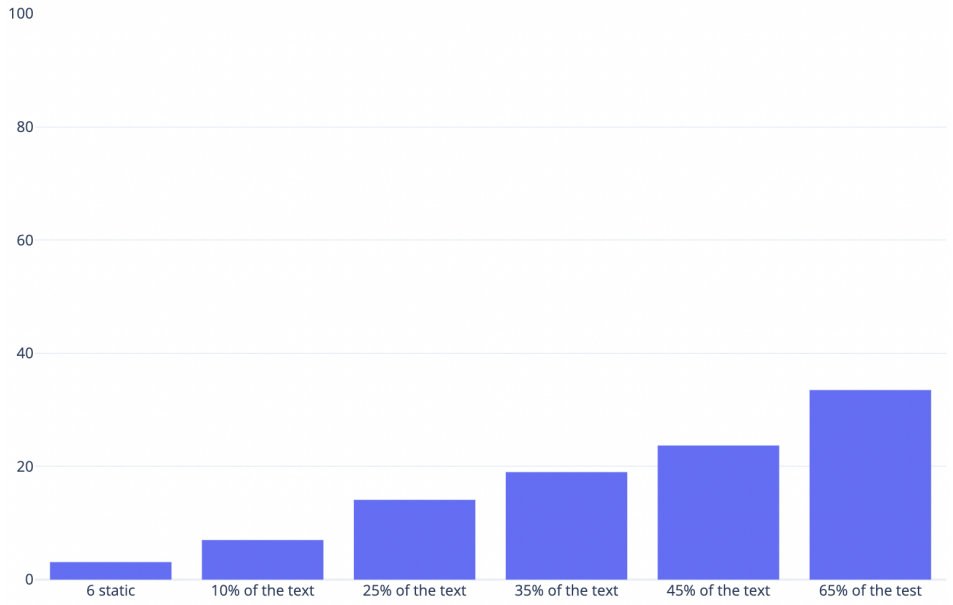


Figure 4.1. Percentage of highlighted test among different parameters

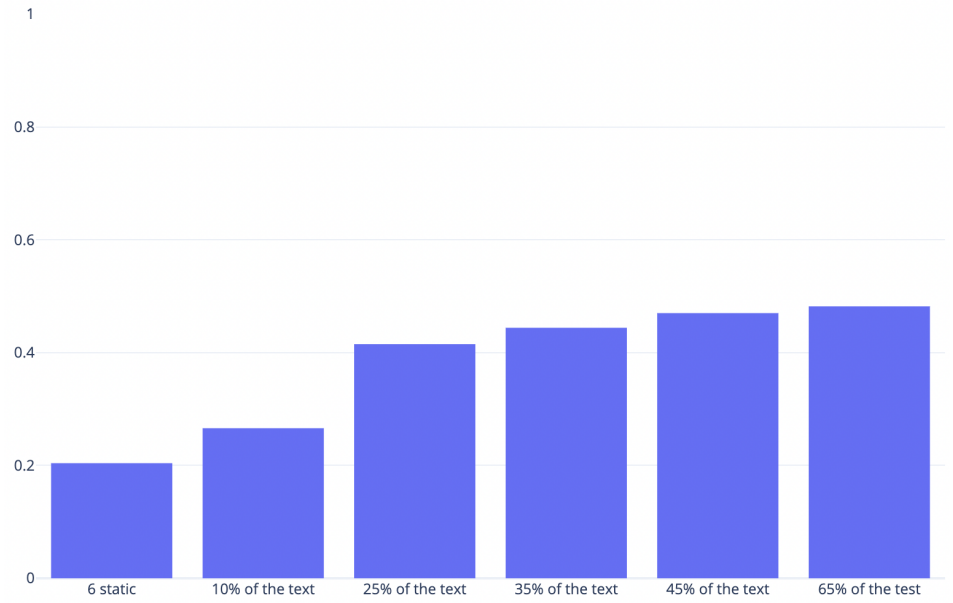


Figure 4.2. Variation of prediction among different parameters

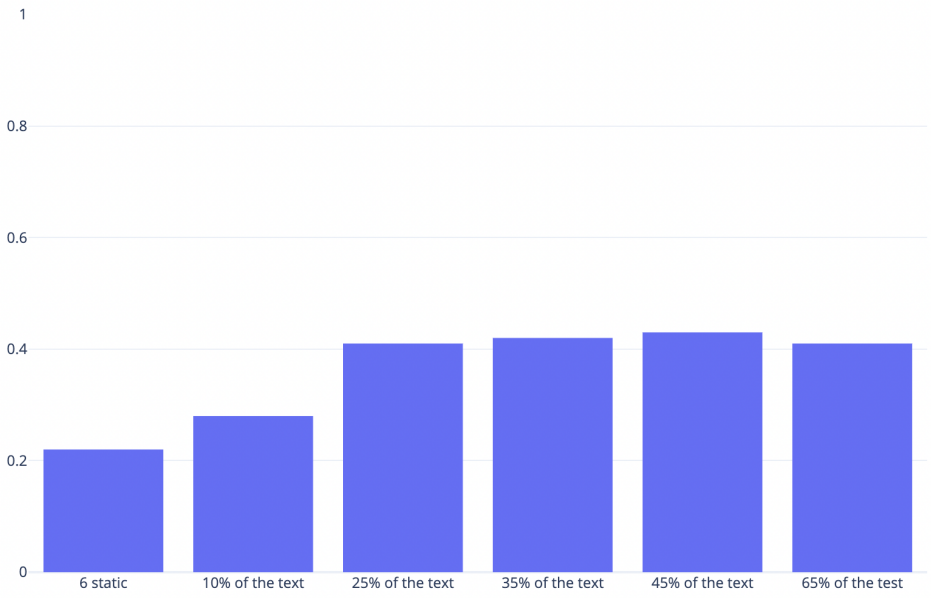


Figure 4.3. Score among different parameters

### 4.3.2 Ebano

#### Saving explanation

With this explain it was not necessary to add anything, the basic explanation could be saved on a json file containing all the information needed to perform the measurements and comparisons.


```
{
   "metadata":{
      "report_id":0,
      "start_time":[
         1630482324.6599784
      ],
      "execution_time":35.736748933792114,
      "flag_pos":true,
      "flag_sen":true,
      "flag_mlwe":true,
      "flag_combinations":true
   },
```

```
"input_info":{
   "raw_text":"Sergei Eisenstein's most famous movie has truly withstood the tes
   "cleaned_text":"Sergei Eisenstein's most famous movie has truly withstood the
   "preprocessed_text":"Sergei Eisenstein ' s most famous movie has truly withst
   "positions_tokens":[
      "Sergei",
      "Eisenstein",
      ...
      "as",
      "\"",
      "Potyomkin",
      "\"",
      "."
   ],
   "original_probabilities":[
      0.0029660449363291264,
      0.997033953666687
   ],
   "original_label":1,
   "expected_label":1
},
"local_explanations":[
   {
      "local_explanation_id":0,
      "feature_id":0,
      "feature_type":"POS",
      "feature_description":"Adjectives",
      "positions_tokens":{
         "5":"famous",
         "19":"mutiny",
         ...
         "147":"\""
      },
      "combination":1,
      "perturbation_id":0,
      "perturbation_type":"Removal Perturbation",
      "perturbed_text":"Sergei Eisenstein ' s most movie has truly withstood the
      "original_probabilities":[...],
      "perturbed_probabilities":[...],
      "original_top_class":1,
      "perturbed_top_class":1,
```

```
            "class_of_interest":1,
            "nPIR_original_top_class":-0.00022412422824782778,
            "nPIRP_original_top_class":0.06962960177511696,
            "nPIR_class_of_interest":-0.00022412422824782778,
            "nPIRP_class_of_interest":0.06962960177511696,
            "nPIRs":[...],
            ...
        },
        ...
        ,
        {
            "local_explanation_id":341,
            ...
        }
    ]
}
```


### Explainer Set-Up

For the elaboration of the explanations with T-EBAnO it was necessary to implement an interface with different functions that allow the explainer not only as in the previous case to know the prediction of the model for a set of texts, but also to know the values of the weights of the layers interior.

### Explainer parameters

For the generation of the explanations with T-EBAnO the configuration of the parameters was minimal, the only parameter to configure is the number of layers to consider when evaluating the word embedding of the model. In this case I have chosen to use the last four layers since T-EBAnO has already been tested on BERT and these turned out to be the best values Ventura et al. [2021].

## 4.3.3 Shap

### Saving explanation

Also for SHAP I needed an alternative way to save the explanation, in this case the download of the proposed standard html view was not even foreseen, and in any case they were not available. As I did for Lime, also in this

case I implemented an additional framework that once the explanation was generated, calculated all the necessary values for comparison and aggregated them together in a .json file.

An example of the composition of the latter is given below


```
{
   "metadata":{
      "report_id":37,
      "execution_time":864.8188180923462
   },
   "input_info":{
      "original_text":"Chupacabra: Dark Waters has to rank as one of the most insip
      "original_label":0,
      "original_prediction":[
         0.9993243217468262,
         0.0006756837829016149
      ]
   },
   "local_explanations":{
      "values":[
         0.0,
         0.03225784301757813,
         -0.0356640100479126,
         0.008340668678283692,
         -0.006902575492858887,
         -0.036200428009033205,
         ...
         0.0
      ],
      "data":[
         "",
         "Chu",
         "pa",
         "ca",
         ...
         "to ",
         "see ",
         "this",
         ".",
         ""
```

```
      ],
      "positive":"ChucaDark of ever my opinion Davies <<An excellent Davies Davies
      "negative":"I had expected least passable substantially/>? The acting incredi
      "prediction_without_positive":[
         0.9993267059326172,
         0.0006733709014952183
      ],
      "prediction_without_negative":[
         0.9973498582839966,
         0.0026501694228500128
      ]
   }
}
```


### Explainer Set-Up

Since this is also a model agnostic method, as well as Lime, the set-up did not require a particular effort, the only requirements were to provide a prediction function and the model tokenizer. The prediction function receives a batch of texts as input and returns an array with the various predictions of each text for each label.

### Explainer parameters

The determination of the best parameters also in this case is not trivial, there is only one to choose but it is very relevant, specifically it is the "algorithm" parameter, it refers to algorithm used to estimate the Shapley values, in this case the possible choice was between "partition" and "permutation". To do this, several experiments were carried out and the choice fell on "permutation".

The experiments were carried out on the sentiment analysis task, it results from these as it can also be seen in the table below that using the argument of "partition" you get a performance increase in terms of processing time at the expense of a cost to be paid on the fidelity of the explanations. On the contrary, by choosing permutation, better performances are obtained on the fidelity part of the explanation, in fact we get a higher score but the processing times are extended.

| | Partition | Permutation |
|---|---|---|
| **Score** | 0,49 | 0,09 |
| **Elaboration Time** | 495 seconds | 3 seconds |

Table 4.5. SHAP parameters comparison

## 4.4 Results

In the next section I will present the numerical results of the metrics already defined in the previous chapter. These will be divided by task.

### 4.4.1 Sentiment Analysis of IMDB Movie Reviews

#### Percentage of highlighted text

The graph below shows a clear misalignment of the values regarding this metric between the various explanators, the best performances are obtained by shap with a median value of 8.85%, slightly less than double for MLWE of T-EBAnO which obtains a median value of 14%, the rest are aligned on 24%. For the explanations of T-EBAnO the values are much more scattered, the variance is more contained for LIME and SHAP. This derives from configuration factors, while the number of features to be selected in LIME and SHAP was set manually, for T-EBAnO this was determined without configuration. Refer to the table in the appendix to view all the values of the experiments.

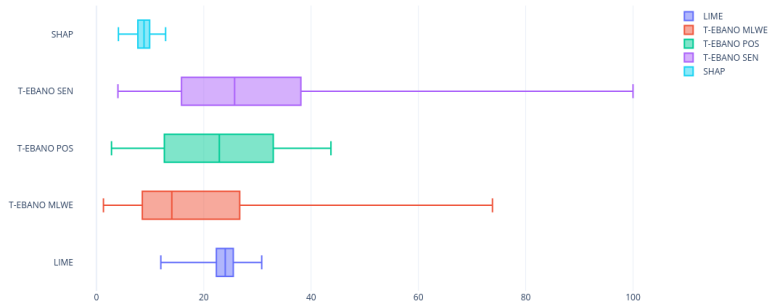


Figure 4.4. Sentiment analysis - percentage of highlighted text

### Variation of prediction

From the graph below we can observe various phenomena. In general, there is a bias towards extreme values. On this aspect the best performance is obtained from files that with MLWE obtain 68 samples in the last interval, it means that 68% of the time has identified the set of words that totally upset the prediction, they are aligned on about 40 LIME SHAP and POS samples . As for the first interval (the one in which the prediction does not change) there are about 40% of the LIME and SHAP samples, better performance for MLWE also in this case with half of the samples.
It is also important to note the presence of LIME and SHAP samples in the negative range, the one in which the probability predicted instead of decreasing increases.

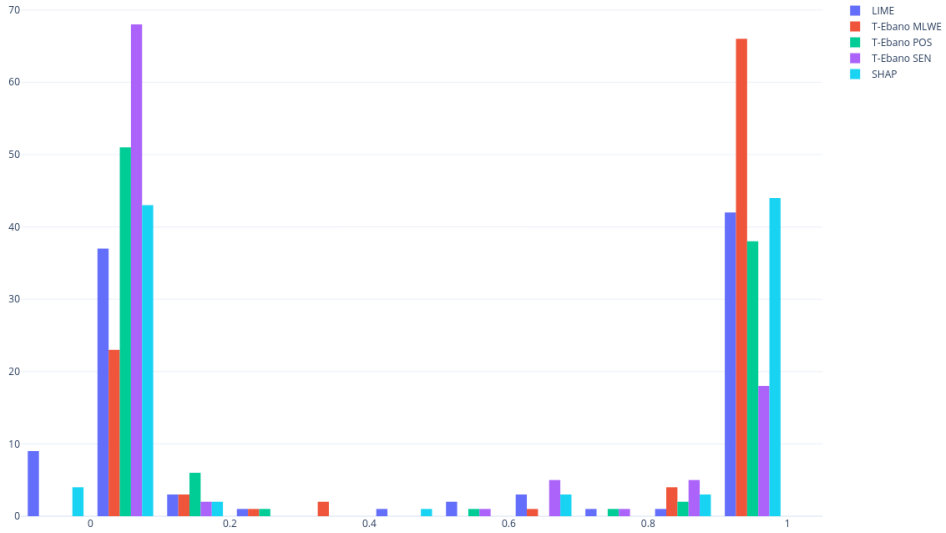


Figure 4.5. Sentiment analysis - absolute variation of prediction

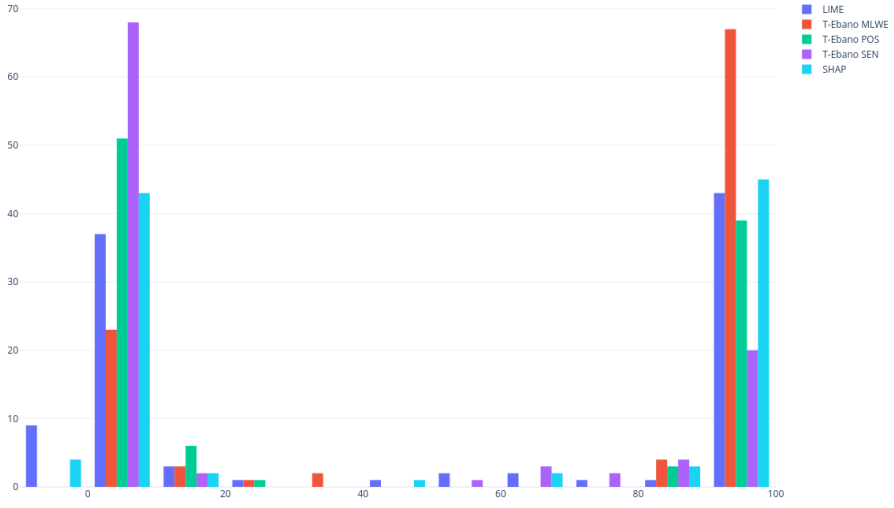


Figure 4.6. Sentiment analysis - relative variation of prediction

### Execution time

From the point of view of the elaboration time of the explanation, the best performances in this case belong to T-EBAnO with a median value of 34 seconds and an extremely limited variance (maximum value 91 seconds). About 9 times slower LIME which generates explanations with an average of 304 seconds. Worst performance for SHAP with a median value of 415 seconds and extreme values reaching 3000 seconds.

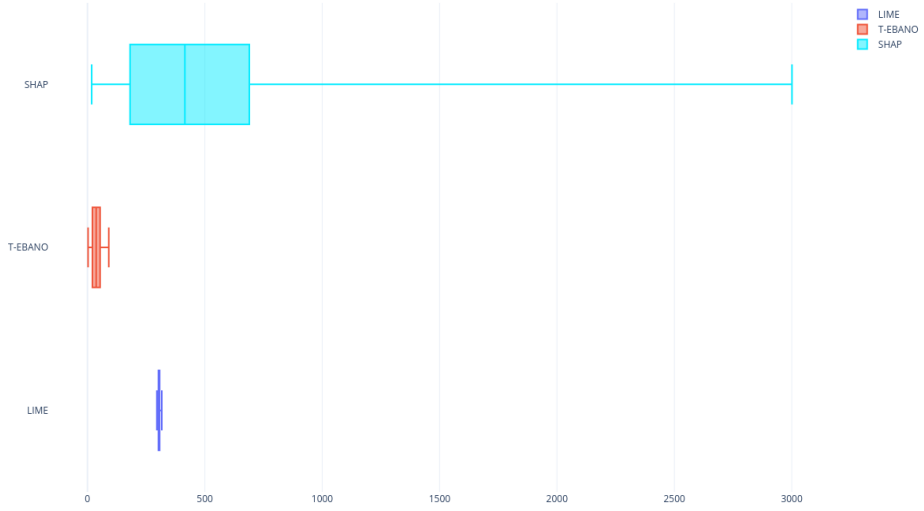


Figure 4.7. Sentiment analysis - elaboration time

### Score

The results shown in the graph also show a certain polarization for this metric, the best results are certainly obtained by T-EBAnO MLWE (the only model specific method), which obtains 80% of the explanations with a score greater than 0.8 and only 20% with a value less than 0.1. As for LIME and SHAP, about 45% of their explanations score less than 0.1, which means that almost half of their explanations highlight features that are not actually relevant to the model.

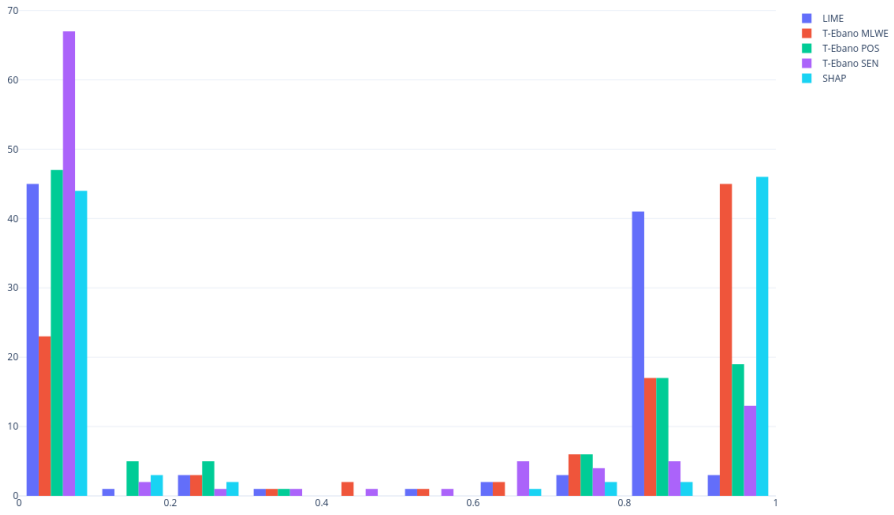


Figure 4.8. Sentiment analysis - score

### 4.4.2 Topic Classification on AG News

#### Percentage of highlighted text

We can see also in this case from the graph below a clear misalignment of the values regarding this metric between the various explanators, once again SHAP obtains the best performance with a median value of 8.6%, slightly more than double for LIME with 17.1 % of highlighted text, followed by T-EBAnO with MLWE showing 28% in median, worse results for POS and SEN as this dataset includes short texts, sometimes even of a single sentence. Also in this case with T-EBAnO we get much more scattered values, the reasons are the same as in the previous task. In the table in the appendix it is possible to consult all the test values.

#### Variation of prediction

Also in this case the polarization phenomenon is present even if less marked, MLWE obtains again the best performances with 60% of the samples in the last interval and 80 samples in the positive half. Opposite situation for SHAP with over 61 values in the first interval. As for LIME we have 57% of

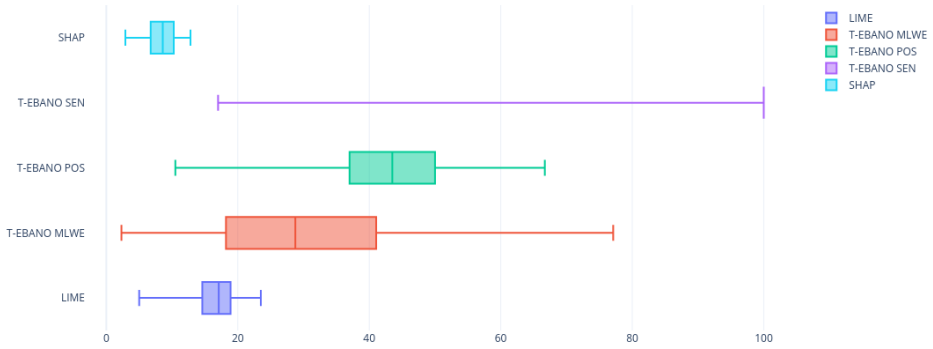


Figure 4.9. Topic classification - percentage of highlighted text

the samples in the worst 5 intervals. Here too we can observe the presence of LIME and SHAP samples in the negative range.

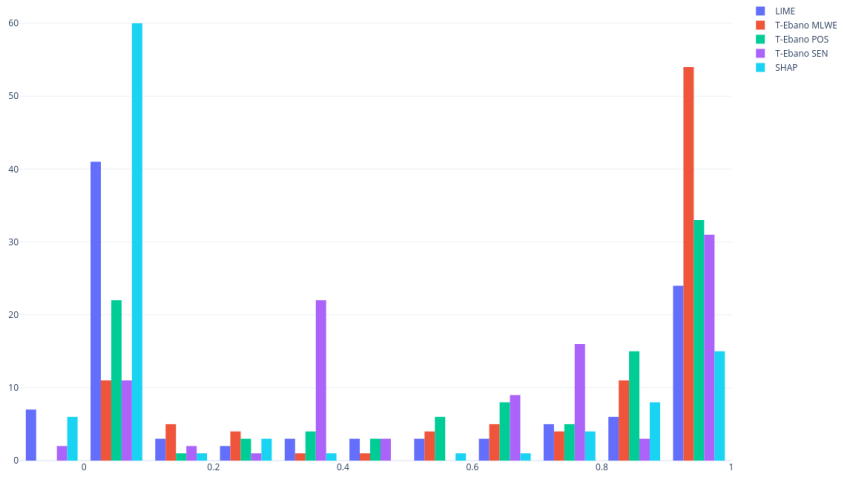


Figure 4.10. Topic classification - absolute variation of prediction

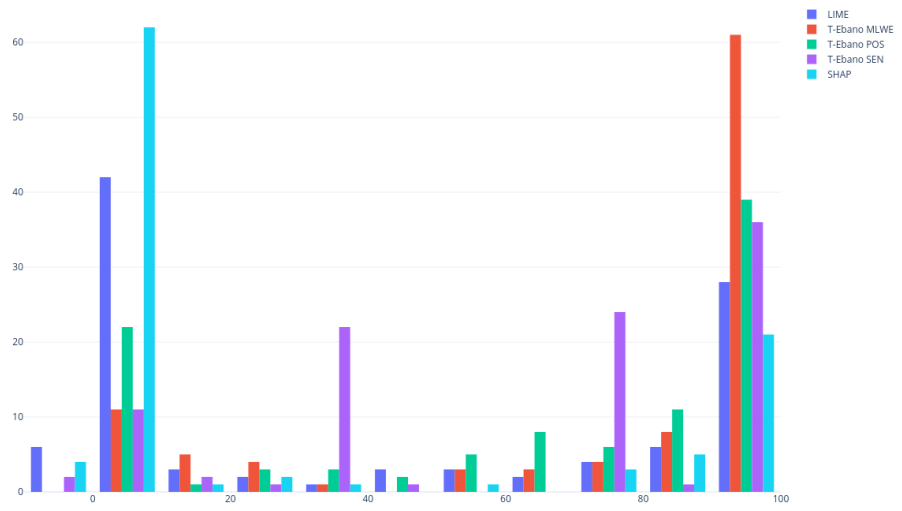


Figure 4.11. Topic classification - relative variation of prediction

### Execution time

For this task the processing times of the explanation are decidedly lower than the previous one, this is due to the composition of the dataset which in this case includes much shorter text to be processed. However, T-EBAnO achieves the best performance with a median time of 3 seconds, followed by SHAP with 6 seconds, 80 times slower than LIME with a median time of 250 seconds.

### Score

As can be seen from the graph shown, also for this task T-EBAnO MLWE obtains excellent results, in this case they are not polarized, but are distributed for the most part in a normal way between the values of 0.5 and 1, where about three quarters fall samples. Good results also for T-EBAnO POS for which there are more than 65% of the samples with a value greater than 0.5. As for LIME and SHAP, there are very polarized results with respectively 40% and 65% of the samples with a score lower than 0.5.

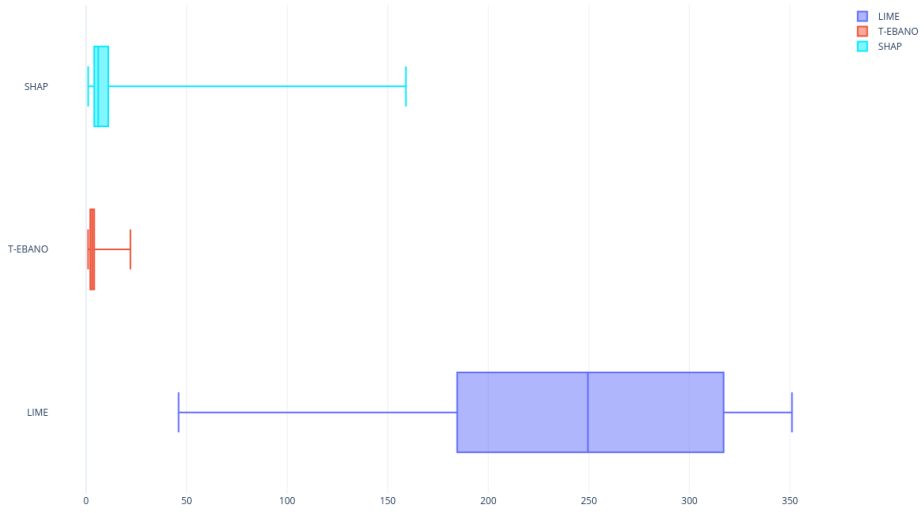


Figure 4.12. Topic classification - elaboration time

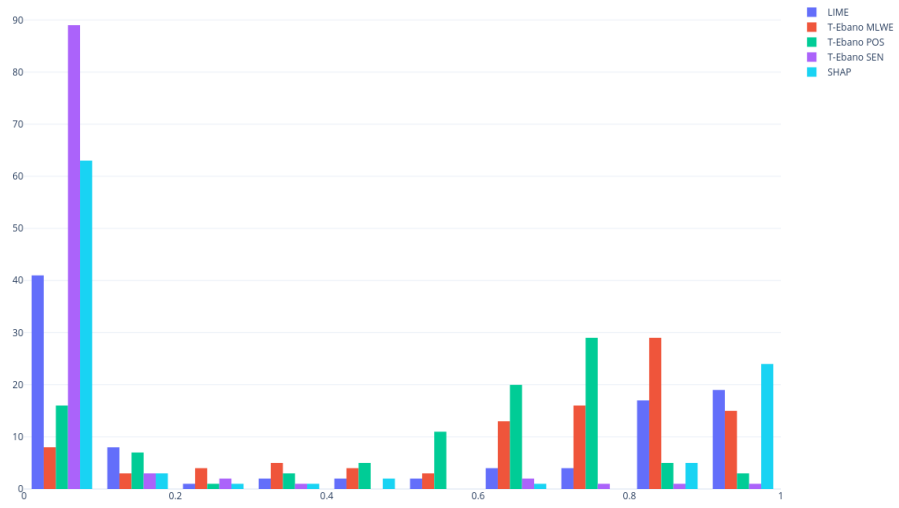


Figure 4.13. Topic classification - score

### 4.4.3 Toxicity Detection on Civil Comment

#### Percentage of highlighted text

As in previous cases, the best values are obtained from SHAP with a median value of 11%, MLWE and LIME are aligned respectively with the median values of 25% and 29%. The phenomena of greater variance for T-EBAnO are also found here.

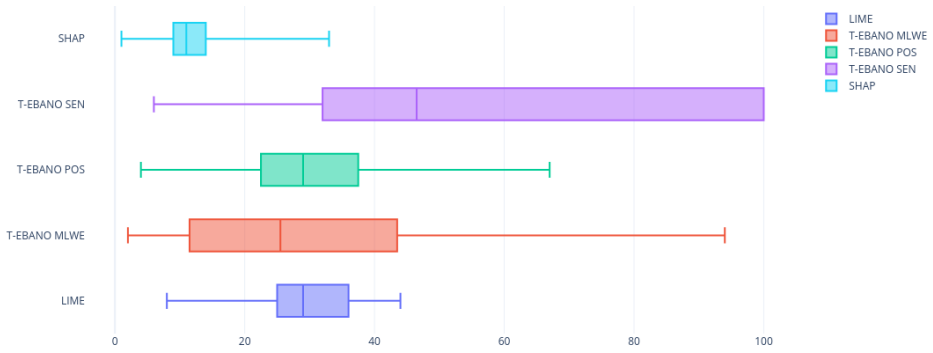


Figure 4.14. Toxicity Detection - percentage of highlighted text

#### Variation of prediction

As for the prediction variation, we note in this case very aligned and polarized performances, in general all the explanators have about 30% of the samples in the best range, and about 60% in the worst. Compared to the other tasks, the performances are generally lower, this is due to the very nature of the task which precludes the possibility of identifying a set of words that, removed from a non-toxic text, makes it become toxic. Also in this case there are samples of shap and lime that change the prediction in the negative.

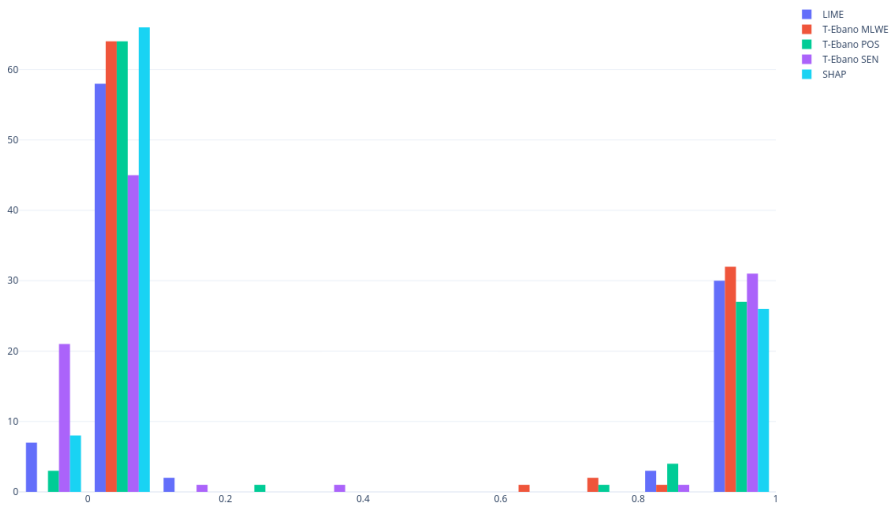


Figure 4.15. Toxicity Detection - absolute variation of prediction

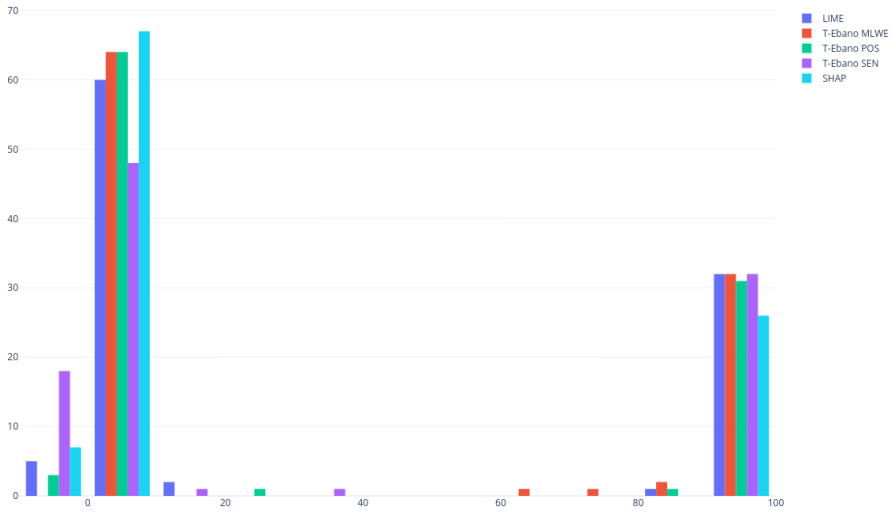


Figure 4.16. Toxicity Detection - relative variation of prediction

### Execution time

We can see from the graph below how T-EBAnO has an excellent performance in terms of execution time, average time of 3 seconds, SHAP and LIME obtain performance on average 10 times worse, with median times of 35 seconds and 191 seconds respectively.

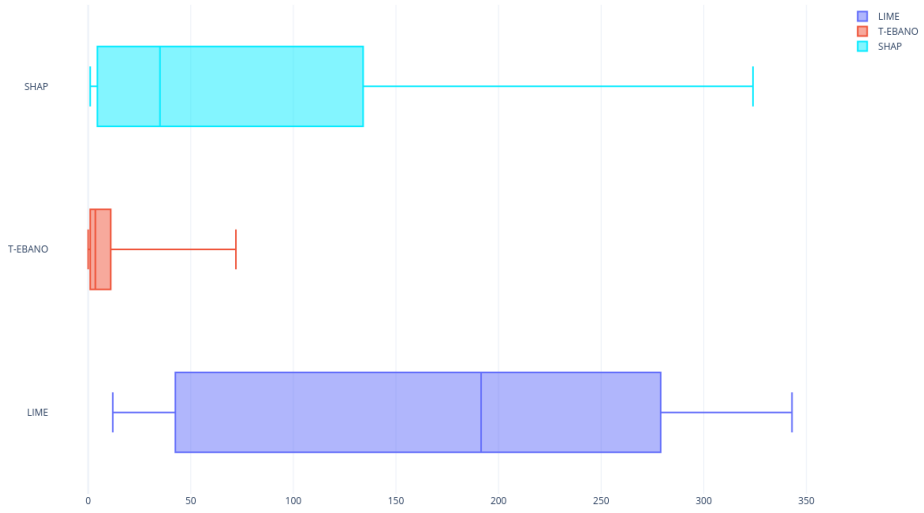


Figure 4.17. Toxicity Detection - elaboration time

### Score

As far as the score on this task is concerned, we obtain the worst results of the three, certainly, as previously mentioned, the nature of the task has a great influence given the definition of the metric itself. The latter in fact enhances explanations that with a reduced number of words are able to change the prediction, but in this case for texts that do not contain toxicity it is difficult to find a set of words, of a clean text, which if removed manages to make it toxic. . As you can see from the graph, the values are very polarized and there are similar distributions between the three methods, the worst of all is SHAP with 70% of the samples with a score lower than 0.1, followed by LIME with about 65%, and finally T -EBAnO with 60%.

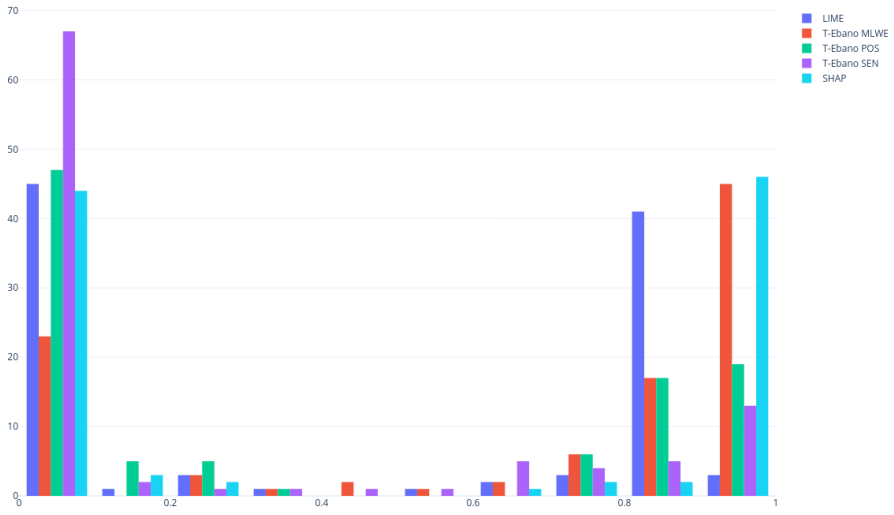


Figure 4.18. Sentiment analysis - score

## 4.5 Survey

In this session we will see how the questionnaire was set up and the results obtained

### 4.5.1 Survey configuration

The configuration of the questionnaire is divided into three macro areas shown below

#### Profiling

In this part, information about the person's background and his confidence with the topics that will be dealt with in the questionnaire are requested. The questions required are as follows

- What is your level of education?
- Are you familiar with machine learning and / or artificial intelligence?

- How many years of experience do you have in this field?
- Are you familiar with Natural Language Processing?
- Are you familiar with explainable artificial intelligence models?

**Tutorial**

A short two-minute video tutorial has been included to explain to the person who has to carry out the questionnaire, the reasons for the questionnaire, the questions that will be asked and on the basis of what to answer.

**Requests**

In this part some texts are shown, for which an artificial intelligence algorithm performed a category prediction. The texts are 12, and are distributed with 1 of the sentiment analysis task, 6 of the toxicity detection task and 5 of the topic classification task. For each of these, the original text, the predicted label and the relative probability are shown.

The first question is "How much do you agree with this prediction?"

Then three different explanations of the same text will be shown, performed by three different explanation methods, respectively LIME, SHAP and T-EBAnO.
Furthermore, for each of these it is indicated how the prediction varies by removing the highlighted words from the explanation.

Finally, the user is asked to answer the following 4 questions with a score from 1 to 5, where 1 is the minimum value and 5 the maximum.

- How human readable is the explanation?
  By Human readable here we mean, do you understand what the explanation tries to tell? Are all the points of the explanation understood?
- How effective is this explanation?
  Does it include all and only the words relevant to the prediction?

- How complete is this explanation?
  Are there all the elements that allow you to understand on the basis of what the prediction was made?

- Considering this explanation, how much do you agree with the original model prediction?
  Do you still believe that the choice you made earlier is correct?

### 4.5.2 Survey results

A partial analysis of the questionnaire shows the following results.
The users who participated in the survey are distributed as shown in the graphs below. Of those who said they were familiar with AI, 22% have 1-2 years of experience. Furthermore, 89% of the respondents are not familiar with natural language processing.

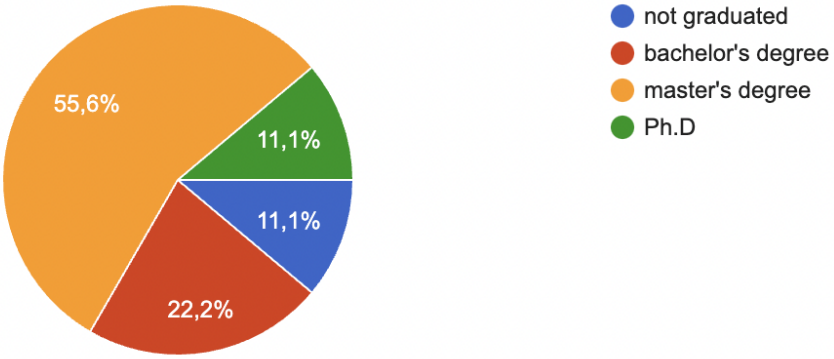


Figure 4.19. Level of education

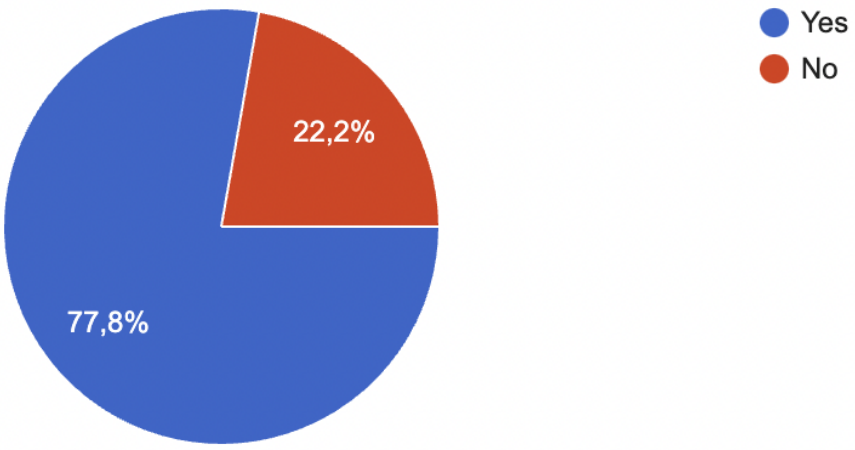


Figure 4.20. Familiar with machine learning and/or artificial intelligence

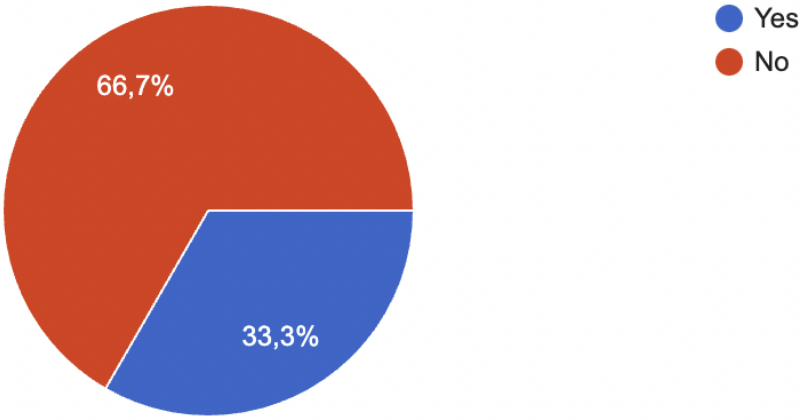


Figure 4.21. Familiar with explainable artificial intelligence models

The results obtained are analyzed below.

### How human readable is the explanation?

On this first question, excellent results were obtained for LIME and T-EBAnO which out of 108 answers obtained the maximum score 54 and 57 times respectively, with average values of 4.19 and 4.37. Worst results for SHAP but which still has a positive average of 3.47 and only 14 votes with a score lower than 3.

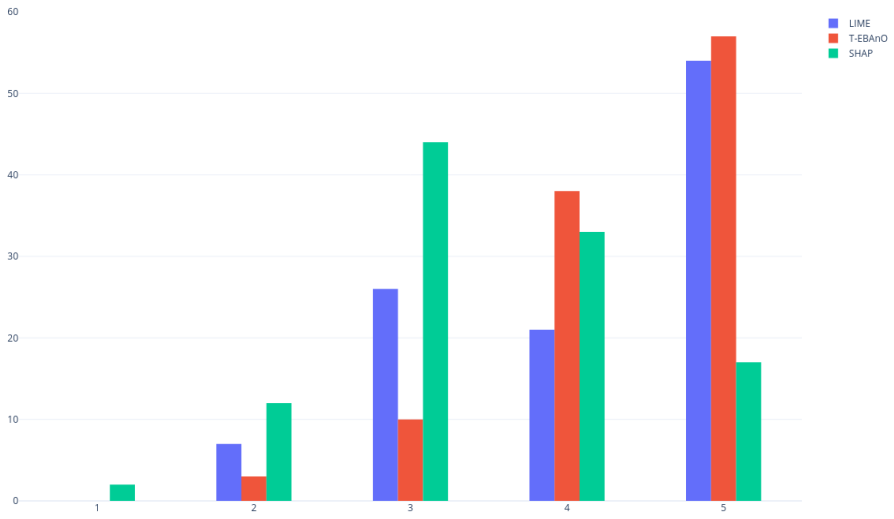


Figure 4.22. How human readable is the explanation?

### How effective is this explanation?

In this case there are clearly better results for T-EBAnO with 89 votes with a score strictly higher than 3 and an average value of 4.22. Scores aligned for LIME and SHAP, slightly in favor of lime, the average values in this case are 3.69 and 3.37 respectively.

### How complete is this explanation?

On this criterion, all three methods obtain less good results than the two previously considered, but nevertheless obtain positive judgments. The best values this time too were achieved by T-EBAnO with an average value of

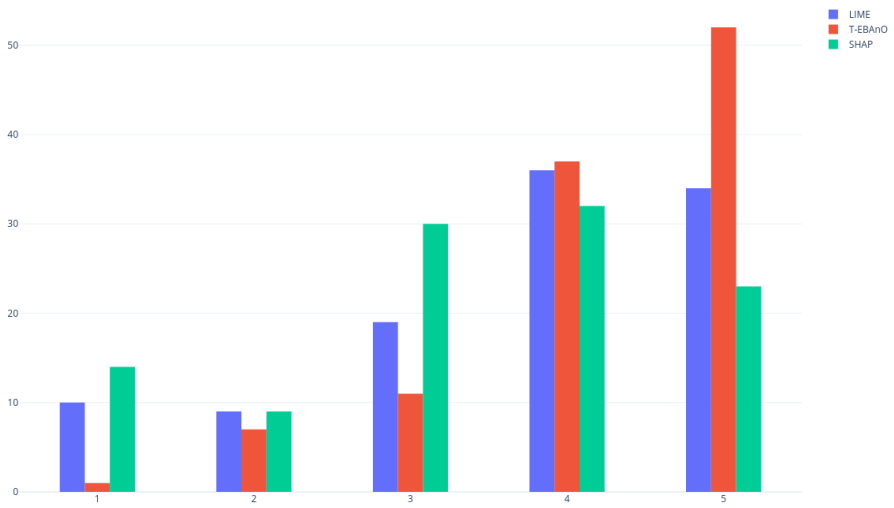


Figure 4.23. How effective is this explanation?

3.91 and 97 opinions out of 108 with a score greater than or equal to 3. LIME follows with an average value of 3.35 but 32 negative judgments (score less than 3 ). Finally SHAP gets an average score of 3.25 and only 25 negative reviews.

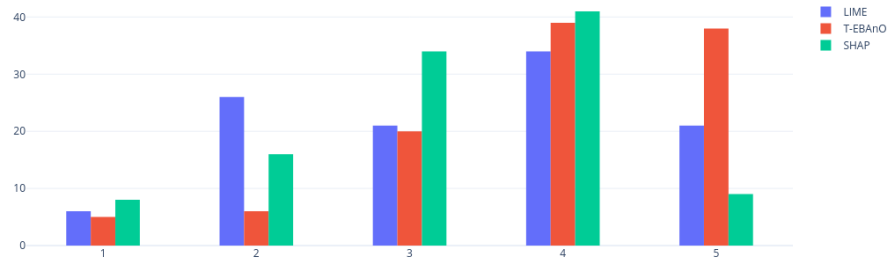


Figure 4.24. How complete is this explanation?

**Considering this explanation, how much do you agree with the original model prediction?**

Regarding this last question it is interesting to consider the variation of confidence in the model before and after the user has seen the explanation, the results in absolute value are shown in the graph. All the methods analyzed succeed for more than half of the time to change the user's opinion, this means that after having seen the explanation, the user has a more complete view of the model. Better results for T-EBAnO which obtains a non-zero value 78 times, followed by SHAP and LIME which manage to obtain values other than 0, respectively 58 and 57 times.

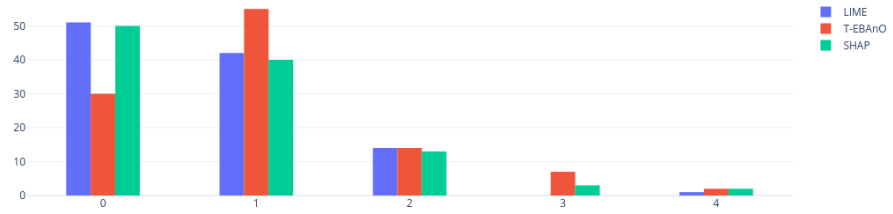


Figure 4.25. Change in confidence in the model after explanation

# Chapter 5

# Conclusion

## 5.1 Conclusion and future work

The original objective of this thesis was to define a methodology experimentally to compare explanations for machine learning models in the domain of natural language processing, and through these to make an in-depth comparison. We have seen how in recent years the demand for interpretability and explainability of artificial intelligence models has undergone an exponential increase, various reasons for this, which are articulated from improving the performance of the model itself to satisfying human curiosity. An increase in the frameworks for explaining artificial intelligence models has followed, but the area of determining how good an explanation is has remained little explored.

In the literature, studies are aligned on a specific taxonomy to categorize the methods of explanation, as regards the evaluation of their goodness, since there is not even a precise mathematical definition of interpretability, there is no clear method of measurement. On this point of view, instead, criteria have been introduced that can be used to judge how good an explanation is and on the basis of which different explanations can be compared. It is not clear in practice how these can be measured correctly, so one of the biggest challenges of this work was precisely to formalize a method for calculating these parameters. In this direction there is only one study that performs an evaluation of the explanations in the domain of natural language processing, but it presents limitations of the point of view of the quantity of criteria and metrics considered and also for the evaluation of the subjective part of the

explanation. Other studies in this direction are very limited and above all not applicable in most cases to unstructured data.

For the quantitative measurement of the criteria, four different metrics have been defined, some of these specific for the domain of natural language processing, respectively the percentage of highlighted text, the absolute and relative prediction variation, the score and the processing time of the explanation. It was therefore possible to evaluate several criteria such as fidelity, comprehensibility, degree of importance and complexity. To measure the more subjective criteria, which require a human evaluation, such as clarity, trustworthiness and effectiveness, a survey was developed and distributed through a questionnaire intended for both ordinary people and people with expertise in the sector.

Subsequently, the comparison was performed, three different methods were considered, one model specific and two model agnostic, respectively T-EBAnO, LIME and SHAP. Local explanations were generated on a BERT model for three different tasks which are sentiment analysis on IMDB review, toxicity detection on civil comment and topic Classification on AG News, in total 900 explanations for 300 texts were considered.The previously defined metrics were also calculated for these.What can be deduced by looking at the performance of the metrics is that a model specific explanation method obtains better performance in most cases, the explanations of T-EBAnO obtain in 47% of cases an excellent score value between 0.8 and 1, although regarding LIME and SHAP this value drops to 35% and 34%. From the point of view of the execution time, there is also a clear advantage of T-EBAnO which on average took 16 seconds to elaborate an explanation, over ten times slower SHAP which obtained an average time of 195 seconds, worse LIME performance with an average time of 238 seconds.

On the qualitative parameters front, a partial analysis of the survey results highlights a general tendency of users to appreciate and benefit from the explanations provided by all three methods. In all three cases, more than 50% of the time the user, after seeing the explanation, is more aware in his judgment regarding the prediction performed by the artificial intelligence, in this area the best result is obtained by T- EBAnO, which in 72% of cases managed to change the user's opinion, the latter value drops to 53% for LIME and SHAP.

Future developments of this thesis can evolve in different directions, certainly it can extend into the domain, currently only NPLs have been considered but it would be interesting to evaluate the metrics defined here on another type of data, such as audio files or images. Another possibility is to consider further methods of explanation and further models, in this work only experiments have been carried out only on BERT, and surely this methodology can be applied on other models.

# Appendix A

# Annex

## A.1 Sentiment analysis on IMDB review - Full Tables

**Percentage of highlighted text**

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |
| 0 | 19.2% | 19.2% | 18.7% | 31.1% | 9.6% |
| 1 | 24.5% | 43.1% | 30.3% | 26.6% | 9.3% |
| 2 | 24.0% | 5.6% | 17.4% | 4.7% | 10.7% |
| 3 | 23.4% | 11.2% | 35.7% | 53.9% | 10.8% |
| 4 | 26.4% | 44.0% | 12.1% | 30.8% | 7.7% |
| 5 | 24.9% | 23.7% | 20.9% | 31.1% | 7.9% |
| 6 | 22.2% | 47.9% | 41.9% | 34.2% | 7.7% |
| 7 | 20.5% | 73.8% | 32.8% | 42.6% | 7.4% |
| 8 | 18.6% | 1.3% | 5.9% | 15.7% | 8.9% |
| 9 | 26.1% | 6.1% | 18.5% | 10.8% | 10.3% |
| 10 | 26.9% | 5.6% | 20.0% | 13.0% | 11.2% |
| 11 | 22.7% | 19.7% | 37.1% | 28.8% | 8.3% |
| 12 | 20.7% | 5.6% | 10.3% | 19.7% | 6.1% |
| 13 | 21.0% | 21.5% | 15.9% | 22.7% | 10.7% |
| 14 | 23.7% | 6.7% | 15.6% | 10.7% | 9.4% |
| 15 | 22.6% | 13.7% | 39.7% | 36.3% | 5.5% |
| 16 | 22.9% | 6.5% | 9.8% | 16.3% | 9.0% |

| | | | | | |
|---|---|---|---|---|---|
| 17 | 21.5% | 27.3% | 33.7% | 23.4% | 10.2% |
| 18 | 24.5% | 28.5% | 33.1% | 23.8% | 8.6% |
| 19 | 23.8% | 8.5% | 24.3% | 100.0% | 9.0% |
| 20 | 29.9% | 15.6% | 33.8% | 51.9% | 7.8% |
| 21 | 23.3% | 2.5% | 4.6% | 8.5% | 10.6% |
| 22 | 28.7% | 35.7% | 17.5% | 32.2% | 9.4% |
| 23 | 23.3% | 57.5% | 43.3% | 24.2% | 6.7% |
| 24 | 22.8% | 26.1% | 24.9% | 7.5% | 8.7% |
| 25 | 23.0% | 12.8% | 23.0% | 27.7% | 7.4% |
| 26 | 25.6% | 18.6% | 33.3% | 56.6% | 8.5% |
| 27 | 24.0% | 2.6% | 31.3% | 9.2% | 10.5% |
| 28 | 27.6% | 42.2% | 43.7% | 20.1% | 7.5% |
| 29 | 23.2% | 10.5% | 7.7% | 16.6% | 8.8% |
| 30 | 19.1% | 1.9% | 12.2% | 9.4% | 10.7% |
| 31 | 20.5% | 59.0% | 22.4% | 19.2% | 8.3% |
| 32 | 24.4% | 3.4% | 4.1% | 13.8% | 10.6% |
| 33 | 14.3% | 9.1% | 15.6% | 50.6% | 7.1% |
| 34 | 25.2% | 10.5% | 29.4% | 43.4% | 9.8% |
| 35 | 23.4% | 48.4% | 39.1% | 12.5% | 7.8% |
| 36 | 25.0% | 20.4% | 10.4% | 33.8% | 9.6% |
| 37 | 25.6% | 33.2% | 10.6% | 8.0% | 10.6% |
| 38 | 22.5% | 12.5% | 3.7% | 4.0% | 9.1% |
| 39 | 26.2% | 16.4% | 8.6% | 14.5% | 9.9% |
| 40 | 12.0% | 4.2% | 9.2% | 7.7% | 4.2% |
| 41 | 22.1% | 4.1% | 7.5% | 5.6% | 10.5% |
| 42 | 25.2% | 27.6% | 11.8% | 91.1% | 7.3% |
| 43 | 24.1% | 6.3% | 10.7% | 17.9% | 8.0% |
| 44 | 27.2% | 16.5% | 12.0% | 31.0% | 8.9% |
| 45 | 25.4% | 16.1% | 33.9% | 89.8% | 8.5% |
| 46 | 26.3% | 11.6% | 26.0% | 23.7% | 10.1% |
| 47 | 26.0% | 18.7% | 30.9% | 61.8% | 8.9% |
| 48 | 24.0% | 43.7% | 33.2% | 18.3% | 9.6% |
| 49 | 16.8% | 6.2% | 7.1% | 17.3% | 8.4% |
| 50 | 24.7% | 11.8% | 10.8% | 25.8% | 8.6% |
| 51 | 22.7% | 9.8% | 37.9% | 53.8% | 4.5% |
| 52 | 25.9% | 33.3% | 29.6% | 68.5% | 9.3% |

| | | | | | |
|---|---|---|---|---|---|
| 53 | 23.6% | 11.8% | 15.7% | 26.8% | 8.7% |
| 54 | 25.1% | 14.4% | 24.3% | 46.9% | 7.0% |
| 55 | 23.3% | 11.3% | 13.8% | 107.5% | 12.9% |
| 56 | 26.4% | 62.5% | 38.2% | 72.2% | 6.9% |
| 57 | 26.4% | 11.8% | 13.2% | 39.6% | 8.3% |
| 58 | 28.1% | 12.4% | 18.5% | 27.3% | 11.2% |
| 59 | 21.2% | 15.3% | 40.7% | 30.1% | 8.1% |
| 60 | 26.1% | 8.4% | 9.4% | 18.3% | 9.9% |
| 61 | 24.7% | 30.4% | 25.3% | 21.5% | 10.1% |
| 62 | 24.6% | 37.7% | 40.7% | 22.8% | 8.4% |
| 63 | 25.0% | 14.7% | 25.0% | 20.7% | 8.7% |
| 64 | 23.2% | 2.1% | 2.8% | 8.7% | 10.0% |
| 65 | 25.2% | 18.7% | 38.1% | 11.6% | 7.1% |
| 66 | 21.8% | 8.2% | 9.5% | 11.9% | 9.9% |
| 67 | 24.3% | 10.8% | 31.4% | 18.4% | 10.3% |
| 68 | 27.4% | 11.1% | 24.7% | 15.3% | 7.3% |
| 69 | 25.0% | 11.0% | 28.7% | 25.7% | 8.1% |
| 70 | 20.1% | 19.5% | 31.5% | 34.2% | 6.0% |
| 71 | 23.5% | 20.4% | 35.4% | 54.9% | 8.0% |
| 72 | 30.8% | 21.8% | 23.3% | 36.1% | 10.5% |
| 73 | 27.5% | 4.1% | 11.1% | 13.5% | 10.5% |
| 74 | 20.5% | 10.7% | 8.9% | 11.6% | 7.1% |
| 75 | 25.5% | 9.5% | 16.0% | 59.5% | 9.5% |
| 76 | 22.9% | 13.3% | 25.0% | 78.2% | 6.9% |
| 77 | 25.0% | 47.2% | 13.7% | 44.8% | 9.0% |
| 78 | 23.7% | 17.2% | 36.6% | 40.9% | 7.5% |
| 79 | 21.3% | 4.7% | 20.7% | 16.0% | 11.2% |
| 80 | 23.5% | 32.2% | 34.2% | 23.5% | 8.7% |
| 81 | 25.5% | 13.6% | 17.0% | 25.5% | 7.7% |
| 82 | 20.0% | 12.7% | 13.1% | 29.4% | 9.4% |
| 83 | 20.1% | 22.2% | 33.3% | 81.3% | 4.9% |
| 84 | 24.6% | 8.1% | 13.8% | 15.0% | 10.2% |
| 85 | 19.9% | 18.9% | 36.4% | 51.9% | 9.2% |
| 86 | 27.5% | 30.6% | 21.2% | 28.0% | 4.1% |
| 87 | 24.7% | 13.2% | 16.6% | 33.6% | 8.5% |
| 88 | 27.2% | 20.9% | 22.8% | 16.5% | 8.9% |

| | | | | | |
|---|---|---|---|---|---|
| 89 | 21.6% | 24.6% | 29.1% | 26.6% | 7.5% |
| 90 | 23.5% | 46.2% | 23.5% | 30.3% | 12.1% |
| 91 | 22.2% | 57.4% | 22.2% | 35.2% | 6.5% |
| 92 | 27.6% | 6.1% | 27.2% | 17.5% | 7.9% |
| 93 | 23.4% | 4.3% | 6.1% | 17.7% | 9.1% |
| 94 | 25.3% | 16.7% | 35.8% | 42.8% | 10.9% |
| 95 | 20.7% | 4.3% | 23.3% | 4.7% | 9.9% |
| 96 | 21.4% | 12.6% | 22.6% | 27.0% | 6.9% |
| 97 | 25.7% | 32.4% | 33.1% | 39.2% | 6.1% |
| 98 | 22.9% | 28.0% | 24.2% | 10.8% | 9.6% |
| 99 | 25.3% | 8.6% | 29.0% | 37.0% | 9.3% |

Table A.1: Percentage of highlighted text - Sentiment analysis

### Variation of prediction

| | **LIME** | **T-EBAnO** | | | **SHAP** | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | | | **MLWE** | **POS** | **SEN** | |
| | **Absolut Variation** | | | | | **Relative Variation** | | | | |
| 0 | -0.03 | -0.99 | -0.04 | -0.05 | -0.99 | -3% | -100% | -4% | -5% | -99% |
| 1 | -0.59 | -0.98 | -0.01 | 0.00 | -0.13 | -59% | -98% | -1% | 0% | -13% |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | -0.99 | 0% | 0% | 0% | 0% | -99% |
| 3 | -0.06 | 0.00 | -0.01 | -0.92 | -0.01 | -6% | 0% | -1% | -92% | -1% |
| 4 | 0.00 | -0.18 | 0.00 | 0.00 | 0.00 | 0% | -18% | 0% | 0% | 0% |
| 5 | -0.98 | -0.99 | -0.98 | 0.00 | -0.06 | -99% | -99% | -98% | 0% | -6% |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 7 | 0.00 | -0.02 | 0.00 | 0.00 | 0.00 | 0% | -2% | 0% | 0% | 0% |
| 8 | -0.63 | -0.63 | -0.59 | -0.62 | -0.63 | -99% | -99% | -92% | -97% | -99% |
| 9 | 0.00 | -0.97 | -0.91 | -0.03 | -0.99 | 0% | -98% | -92% | -4% | -100% |
| 10 | -0.50 | -0.99 | -0.96 | -0.94 | 0.00 | -50% | -99% | -96% | -94% | 0% |
| 11 | -0.99 | -0.98 | -0.97 | -0.58 | -0.91 | -100% | -98% | -97% | -58% | -91% |
| 12 | -0.14 | -0.98 | -0.02 | -0.07 | -0.99 | -14% | -98% | -2% | -7% | -100% |
| 13 | -0.01 | -0.97 | -0.14 | -0.03 | -0.97 | -1% | -97% | -14% | -3% | -98% |
| 14 | -0.97 | -0.99 | -0.99 | -0.96 | -0.99 | -97% | -99% | -99% | -96% | -100% |
| 15 | 0.00 | -0.99 | -0.96 | -0.82 | -0.04 | 0% | -100% | -97% | -82% | -4% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16 | -0.97 | -0.96 | -0.96 | -0.96 | -0.97 | -100% | -99% | -99% | -99% | -100% |
| 17 | -0.01 | -0.99 | -0.13 | -0.03 | -0.99 | -1% | -99% | -13% | -3% | -99% |
| 18 | 0.00 | -0.99 | -0.02 | 0.00 | 0.00 | 0% | -99% | -2% | 0% | 0% |
| 19 | 0.00 | -0.99 | -0.17 | -0.97 | 0.00 | 0% | -99% | -17% | -97% | 0% |
| 20 | -0.99 | -0.32 | -0.11 | 0.00 | 0.00 | -99% | -32% | -11% | 0% | 0% |
| 21 | -0.97 | -0.97 | -0.97 | -0.97 | -0.97 | -100% | -100% | -100% | -99% | -100% |
| 22 | -0.97 | -0.85 | 0.00 | -0.01 | -0.02 | -97% | -86% | 0% | -1% | -2% |
| 23 | -0.99 | -0.04 | -0.93 | 0.00 | -0.94 | -99% | -4% | -94% | 0% | -94% |
| 24 | 0.00 | -0.94 | 0.00 | 0.00 | 0.00 | 0% | -94% | 0% | 0% | 0% |
| 25 | -0.99 | -0.97 | -0.15 | -0.02 | -0.99 | -99% | -98% | -15% | -2% | -100% |
| 26 | -0.97 | -0.92 | -0.98 | 0.00 | 0.00 | -97% | -92% | -98% | 0% | 0% |
| 27 | -0.97 | -0.92 | -0.01 | -0.93 | -0.97 | -99% | -94% | -1% | -95% | -99% |
| 28 | -0.03 | -0.93 | -0.04 | -0.01 | -0.92 | -3% | -93% | -4% | -1% | -92% |
| 29 | -0.92 | -0.98 | -0.03 | -0.02 | -0.88 | -92% | -99% | -3% | -2% | -88% |
| 30 | -0.99 | -0.97 | -0.99 | -0.19 | -0.99 | -100% | -99% | -100% | -19% | -100% |
| 31 | 0.00 | -0.98 | 0.00 | 0.00 | 0.00 | 0% | -98% | 0% | 0% | 0% |
| 32 | -0.99 | -0.98 | -0.98 | -0.98 | -0.91 | -100% | -99% | -99% | -99% | -92% |
| 33 | -0.99 | -0.98 | -0.95 | -0.74 | -0.09 | -100% | -99% | -96% | -74% | -9% |
| 34 | -0.99 | -0.95 | -0.99 | -0.65 | -0.98 | -99% | -95% | -99% | -65% | -99% |
| 35 | -0.02 | -0.92 | 0.00 | 0.00 | -0.69 | -2% | -93% | 0% | 0% | -69% |
| 36 | 0.00 | -0.97 | 0.00 | 0.00 | -0.46 | 0% | -98% | 0% | 0% | -46% |
| 37 | -0.74 | 0.00 | 0.00 | 0.00 | 0.00 | -75% | 0% | 0% | 0% | 0% |
| 38 | 0.00 | 0.00 | -0.99 | -0.88 | 0.00 | 0% | 0% | -99% | -88% | 0% |
| 39 | 0.00 | -0.98 | -0.96 | 0.00 | -0.99 | 0% | -98% | -97% | 0% | -100% |
| 40 | -0.98 | -0.98 | -0.97 | -0.96 | -0.98 | -100% | -100% | -99% | -98% | -100% |
| 41 | 0.00 | -0.97 | -0.98 | -0.14 | -0.97 | 0% | -97% | -99% | -14% | -97% |
| 42 | 0.00 | 0.00 | 0.00 | 0.00 | -0.99 | 0% | 0% | 0% | 0% | -99% |
| 43 | -1.00 | -0.97 | -0.99 | -0.03 | -1.00 | -100% | -97% | -100% | -3% | -100% |
| 44 | -0.97 | -0.13 | -0.04 | -0.02 | -0.01 | -97% | -13% | -4% | -2% | -1% |
| 45 | 0.00 | -0.01 | -0.01 | 0.00 | 0.00 | 0% | -1% | -1% | 0% | 0% |
| 46 | 0.00 | -0.96 | -0.99 | -0.01 | -0.89 | 0% | -96% | -100% | -1% | -89% |
| 47 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0% | -1% | 0% | 0% | 0% |
| 48 | 0.00 | 0.00 | -0.99 | 0.00 | -0.02 | 0% | 0% | -99% | 0% | -2% |
| 49 | -0.99 | -0.96 | -0.99 | 0.00 | -0.99 | -100% | -96% | -99% | 0% | -100% |
| 50 | -1.00 | -1.00 | -0.94 | 0.00 | -0.69 | -100% | -100% | -94% | 0% | -69% |
| 51 | -0.99 | -0.99 | -0.99 | -0.96 | -0.99 | -100% | -99% | -100% | -97% | -99% |
| 52 | -0.95 | -0.95 | 0.00 | -0.93 | 0.00 | -96% | -95% | 0% | -93% | 0% |
| 53 | -0.83 | -0.99 | -0.02 | -0.89 | -0.99 | -83% | -100% | -2% | -90% | -99% |
| 54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 55 | -0.01 | -0.97 | -0.09 | 0.00 | -0.99 | -1% | -97% | -9% | 0% | -100% |
| 56 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0% | -1% | 0% | 0% | 0% |
| 57 | -0.98 | -0.97 | -0.05 | -0.94 | -0.98 | -100% | -98% | -5% | -96% | -99% |
| 58 | 0.00 | -0.88 | -0.93 | 0.00 | -0.94 | 0% | -89% | -93% | 0% | -94% |
| 59 | 0.00 | -0.99 | -0.01 | -0.01 | -0.01 | 0% | -99% | -1% | -1% | -1% |
| 60 | -0.99 | -0.01 | -0.80 | -0.03 | -0.12 | -99% | -1% | -80% | -3% | -12% |
| 61 | -0.12 | -0.95 | -0.12 | 0.00 | 0.00 | -12% | -95% | -12% | 0% | 0% |
| 62 | -0.01 | -0.02 | 0.00 | 0.00 | -0.96 | -1% | -2% | 0% | 0% | -96% |
| 63 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0% | 0% | -1% | 0% | 0% |
| 64 | -0.98 | -0.99 | 0.00 | 0.00 | -0.87 | -98% | -99% | 0% | 0% | -87% |
| 65 | -0.96 | -0.90 | 0.00 | 0.00 | 0.00 | -96% | -90% | 0% | 0% | 0% |
| 66 | -0.97 | -0.97 | -0.96 | -0.88 | -0.97 | -99% | -100% | -98% | -90% | -99% |
| 67 | -0.69 | -0.96 | -0.21 | 0.00 | -0.99 | -69% | -96% | -21% | 0% | -99% |
| 68 | 0.00 | -0.03 | 0.00 | 0.00 | 0.00 | 0% | -3% | 0% | 0% | 0% |
| 69 | -1.00 | -0.97 | -0.08 | -0.01 | -0.99 | -100% | -97% | -8% | -1% | -100% |
| 70 | -0.98 | -0.96 | -0.94 | -0.03 | -0.01 | -99% | -96% | -94% | -3% | -1% |
| 71 | 0.00 | -0.02 | 0.00 | 0.00 | 0.00 | 0% | -2% | 0% | 0% | 0% |
| 72 | -0.46 | 0.00 | 0.00 | 0.00 | 0.00 | -46% | 0% | 0% | 0% | 0% |
| 73 | 0.00 | -0.98 | -0.96 | -0.94 | -0.99 | 0% | -99% | -97% | -94% | -100% |
| 74 | -0.99 | -0.88 | -0.83 | -0.95 | -0.99 | -100% | -89% | -84% | -95% | -100% |
| 75 | -0.01 | -0.33 | -0.86 | -0.01 | 0.00 | -1% | -33% | -86% | -1% | 0% |
| 76 | 0.00 | -0.97 | 0.00 | 0.00 | -0.09 | 0% | -97% | 0% | 0% | -9% |
| 77 | -0.02 | -0.17 | 0.00 | -0.01 | -0.03 | -2% | -17% | 0% | -1% | -3% |
| 78 | -0.99 | -0.97 | -0.08 | -0.01 | -0.01 | -99% | -98% | -8% | -1% | -1% |
| 79 | -0.98 | -0.97 | -0.96 | -0.89 | -0.98 | -100% | -99% | -98% | -91% | -100% |
| 80 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 81 | -0.22 | -0.99 | -0.96 | -0.99 | -0.99 | -22% | -100% | -96% | -99% | -99% |
| 82 | -0.98 | -0.97 | -0.96 | -0.06 | -0.02 | -100% | -98% | -97% | -6% | -2% |
| 83 | -0.98 | 0.00 | -0.01 | -0.03 | -0.94 | -98% | 0% | -1% | -3% | -94% |
| 84 | -0.99 | -0.99 | 0.00 | -0.01 | -0.99 | -100% | -99% | 0% | -1% | -100% |
| 85 | 0.00 | -0.93 | -0.01 | 0.00 | -0.01 | 0% | -93% | -1% | 0% | -1% |
| 86 | 0.00 | -0.90 | 0.00 | 0.00 | -0.01 | 0% | -90% | 0% | 0% | -1% |
| 87 | -0.11 | -0.99 | -0.99 | -0.99 | -0.99 | -11% | -100% | -99% | -100% | -100% |
| 88 | -0.99 | -0.97 | -0.98 | -0.64 | -0.99 | -100% | -97% | -99% | -64% | -99% |
| 89 | -0.64 | -0.25 | 0.00 | 0.00 | 0.00 | -64% | -25% | 0% | 0% | 0% |
| 90 | 0.00 | -0.99 | 0.00 | 0.00 | 0.00 | 0% | -99% | 0% | 0% | 0% |
| 91 | 0.00 | -0.99 | 0.00 | 0.00 | 0.00 | 0% | -99% | 0% | 0% | 0% |
| 92 | -0.98 | -0.98 | -0.92 | -0.97 | -0.99 | -100% | -99% | -93% | -99% | -100% |
| 93 | 0.00 | -0.99 | -0.99 | 0.00 | 0.00 | 0% | -99% | -99% | 0% | 0% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 94 | -0.92 | -0.01 | -0.02 | 0.00 | 0.00 | -92% | -1% | -2% | 0% | 0% |
| 95 | -0.97 | -0.98 | -0.97 | -0.98 | -0.98 | -99% | -100% | -99% | -100% | -100% |
| 96 | -0.99 | -0.97 | -0.97 | -0.68 | -0.99 | -100% | -98% | -97% | -68% | -100% |
| 97 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 98 | 0.00 | -0.99 | -0.02 | 0.00 | -0.02 | 0% | -100% | -2% | 0% | -2% |
| 99 | -0.98 | -0.93 | -0.92 | -0.70 | -0.98 | -100% | -95% | -94% | -71% | -100% |

Table A.2: Variation of prediction - Sentiment analysis

## Time of elaboration

| **ID** | **LIME (s)** | **T-EBAnO (s)** | **SHAP (s)** | **ID** | **LIME (s)** | **T-EBAnO(s)** | **SHAP (s)** |
|---|---|---|---|---|---|---|---|
| 0 | 304 | 55 | 548 | 50 | 305 | 12 | 106 |
| 1 | 308 | 42 | 842 | 51 | 303 | 21 | 159 |
| 2 | 310 | 64 | 1510 | 52 | 303 | 3 | 18 |
| 3 | 306 | 63 | 752 | 53 | 304 | 17 | 127 |
| 4 | 297 | 8 | 56 | 54 | 304 | 53 | 587 |
| 5 | 296 | 27 | 234 | 55 | 301 | 52 | 857 |
| 6 | 297 | 11 | 107 | 56 | 301 | 17 | 205 |
| 7 | 297 | 36 | 242 | 57 | 305 | 23 | 208 |
| 8 | 300 | 54 | 658 | 58 | 304 | 56 | 717 |
| 9 | 302 | 67 | 1062 | 59 | 304 | 55 | 650 |
| 10 | 300 | 70 | 1349 | 60 | 307 | 55 | 1154 |
| 11 | 297 | 18 | 227 | 61 | 305 | 22 | 239 |
| 12 | 297 | 49 | 511 | 62 | 305 | 24 | 246 |
| 13 | 297 | 91 | 503 | 63 | 306 | 61 | 808 |
| 14 | 299 | 47 | 548 | 64 | 312 | 74 | 3001 |
| 15 | 298 | 24 | 223 | 65 | 306 | 25 | 223 |
| 16 | 301 | 59 | 523 | 66 | 305 | 57 | 577 |
| 17 | 302 | 49 | 601 | 67 | 299 | 38 | 316 |
| 18 | 301 | 36 | 348 | 68 | 305 | 59 | 748 |
| 19 | 301 | 33 | 437 | 69 | 304 | 22 | 149 |
| 20 | 301 | 5 | 46 | 70 | 301 | 27 | 174 |
| 21 | 303 | 50 | 814 | 71 | 304 | 45 | 516 |
| 22 | 303 | 24 | 260 | 72 | 304 | 16 | 144 |
| 23 | 304 | 14 | 149 | 73 | 305 | 64 | 884 |
| 24 | 317 | 45 | 494 | 74 | 306 | 16 | 100 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 25 | 312 | 23 | 189 | 75 | 309 | 44 | 1038 |
| 26 | 309 | 19 | 158 | 76 | 308 | 30 | 524 |
| 27 | 306 | 50 | 763 | 77 | 307 | 33 | 498 |
| 28 | 298 | 33 | 378 | 78 | 307 | 11 | 83 |
| 29 | 305 | 48 | 1035 | 79 | 306 | 35 | 363 |
| 30 | 301 | 58 | 896 | 80 | 306 | 20 | 227 |
| 31 | 306 | 20 | 194 | 81 | 308 | 51 | 647 |
| 32 | 302 | 65 | 906 | 82 | 302 | 47 | 665 |
| 33 | 303 | 25 | 198 | 83 | 305 | 19 | 217 |
| 34 | 302 | 19 | 162 | 84 | 309 | 59 | 1004 |
| 35 | 304 | 4 | 29 | 85 | 308 | 34 | 488 |
| 36 | 302 | 57 | 608 | 86 | 306 | 31 | 394 |
| 37 | 300 | 56 | 865 | 87 | 306 | 49 | 714 |
| 38 | 303 | 64 | 1034 | 88 | 309 | 21 | 131 |
| 39 | 304 | 58 | 889 | 89 | 309 | 38 | 290 |
| 40 | 304 | 23 | 155 | 90 | 307 | 17 | 112 |
| 41 | 309 | 86 | 1767 | 91 | 310 | 40 | 350 |
| 42 | 310 | 50 | 607 | 92 | 309 | 53 | 434 |
| 43 | 311 | 13 | 122 | 93 | 309 | 44 | 406 |
| 44 | 304 | 29 | 223 | 94 | 307 | 52 | 498 |
| 45 | 302 | 15 | 192 | 95 | 311 | 54 | 424 |
| 46 | 301 | 67 | 923 | 96 | 309 | 24 | 120 |
| 47 | 302 | 19 | 147 | 97 | 308 | 22 | 126 |
| 48 | 304 | 46 | 538 | 98 | 305 | 21 | 108 |
| 49 | 305 | 54 | 489 | 99 | 305 | 28 | 174 |

Table A.3: Elaboration time - Sentiment analysis

**Score**

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |
| 0 | 0.06 | 0.89 | 0.08 | 0.09 | 0.95 |
| 1 | 0.66 | 0.72 | 0.01 | 0.00 | 0.23 |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.94 |
| 3 | 0.12 | 0.01 | 0.01 | 0.62 | 0.01 |
| 4 | 0.00 | 0.27 | 0.00 | 0.00 | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| 5 | 0.85 | 0.86 | 0.88 | 0.00 | 0.11 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 |
| 8 | 0.89 | 0.99 | 0.93 | 0.90 | 0.95 |
| 9 | 0.00 | 0.96 | 0.86 | 0.07 | 0.94 |
| 10 | 0.60 | 0.97 | 0.87 | 0.90 | 0.00 |
| 11 | 0.87 | 0.88 | 0.76 | 0.64 | 0.91 |
| 12 | 0.24 | 0.96 | 0.04 | 0.12 | 0.97 |
| 13 | 0.01 | 0.87 | 0.24 | 0.05 | 0.93 |
| 14 | 0.86 | 0.96 | 0.91 | 0.93 | 0.95 |
| 15 | 0.01 | 0.92 | 0.74 | 0.72 | 0.07 |
| 16 | 0.87 | 0.96 | 0.94 | 0.91 | 0.95 |
| 17 | 0.01 | 0.84 | 0.22 | 0.06 | 0.94 |
| 18 | 0.00 | 0.83 | 0.04 | 0.00 | 0.00 |
| 19 | 0.00 | 0.95 | 0.27 | 0.00 | 0.00 |
| 20 | 0.82 | 0.47 | 0.19 | 0.00 | 0.00 |
| 21 | 0.87 | 0.99 | 0.98 | 0.95 | 0.94 |
| 22 | 0.82 | 0.73 | 0.01 | 0.02 | 0.03 |
| 23 | 0.87 | 0.08 | 0.71 | 0.00 | 0.94 |
| 24 | 0.01 | 0.83 | 0.01 | 0.00 | 0.01 |
| 25 | 0.87 | 0.92 | 0.25 | 0.04 | 0.96 |
| 26 | 0.84 | 0.87 | 0.79 | 0.00 | 0.00 |
| 27 | 0.86 | 0.95 | 0.02 | 0.93 | 0.94 |
| 28 | 0.06 | 0.71 | 0.08 | 0.01 | 0.92 |
| 29 | 0.84 | 0.94 | 0.05 | 0.04 | 0.90 |
| 30 | 0.89 | 0.98 | 0.93 | 0.31 | 0.94 |
| 31 | 0.00 | 0.58 | 0.00 | 0.00 | 0.00 |
| 32 | 0.86 | 0.98 | 0.97 | 0.92 | 0.91 |
| 33 | 0.92 | 0.95 | 0.90 | 0.59 | 0.17 |
| 34 | 0.85 | 0.92 | 0.82 | 0.61 | 0.94 |
| 35 | 0.03 | 0.66 | 0.00 | 0.00 | 0.79 |
| 36 | 0.01 | 0.88 | 0.00 | 0.01 | 0.61 |
| 37 | 0.74 | 0.00 | 0.00 | 0.00 | 0.00 |
| 38 | 0.00 | 0.01 | 0.98 | 0.92 | 0.00 |
| 39 | 0.00 | 0.90 | 0.94 | 0.01 | 0.95 |
| 40 | 0.94 | 0.98 | 0.95 | 0.95 | 0.98 |

| | | | | | |
|---|---|---|---|---|---|
| 41 | 0.00 | 0.97 | 0.95 | 0.25 | 0.93 |
| 42 | 0.01 | 0.01 | 0.00 | 0.00 | 0.96 |
| 43 | 0.86 | 0.95 | 0.94 | 0.05 | 0.96 |
| 44 | 0.83 | 0.22 | 0.07 | 0.04 | 0.02 |
| 45 | 0.00 | 0.02 | 0.01 | 0.01 | 0.00 |
| 46 | 0.00 | 0.92 | 0.85 | 0.02 | 0.89 |
| 47 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 |
| 48 | 0.00 | 0.01 | 0.80 | 0.00 | 0.04 |
| 49 | 0.91 | 0.95 | 0.96 | 0.01 | 0.95 |
| 50 | 0.86 | 0.94 | 0.91 | 0.01 | 0.79 |
| 51 | 0.87 | 0.95 | 0.77 | 0.63 | 0.97 |
| 52 | 0.83 | 0.78 | 0.01 | 0.47 | 0.00 |
| 53 | 0.80 | 0.94 | 0.03 | 0.81 | 0.95 |
| 54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| 55 | 0.01 | 0.93 | 0.16 | 0.01 | 0.93 |
| 56 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| 57 | 0.85 | 0.93 | 0.10 | 0.74 | 0.95 |
| 58 | 0.00 | 0.88 | 0.87 | 0.01 | 0.92 |
| 59 | 0.00 | 0.91 | 0.02 | 0.02 | 0.02 |
| 60 | 0.85 | 0.01 | 0.85 | 0.05 | 0.21 |
| 61 | 0.21 | 0.80 | 0.21 | 0.00 | 0.00 |
| 62 | 0.01 | 0.03 | 0.00 | 0.00 | 0.94 |
| 63 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| 64 | 0.86 | 0.99 | 0.00 | 0.00 | 0.88 |
| 65 | 0.84 | 0.86 | 0.00 | 0.00 | 0.00 |
| 66 | 0.87 | 0.96 | 0.94 | 0.89 | 0.94 |
| 67 | 0.72 | 0.92 | 0.32 | 0.00 | 0.94 |
| 68 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 |
| 69 | 0.86 | 0.93 | 0.15 | 0.02 | 0.96 |
| 70 | 0.88 | 0.87 | 0.79 | 0.06 | 0.02 |
| 71 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 |
| 72 | 0.55 | 0.00 | 0.00 | 0.00 | 0.00 |
| 73 | 0.00 | 0.97 | 0.93 | 0.90 | 0.94 |
| 74 | 0.88 | 0.89 | 0.87 | 0.92 | 0.96 |
| 75 | 0.02 | 0.48 | 0.85 | 0.01 | 0.01 |
| 76 | 0.00 | 0.92 | 0.00 | 0.00 | 0.16 |

| | | | | | |
|---|---|---|---|---|---|
| 77 | 0.04 | 0.26 | 0.01 | 0.01 | 0.06 |
| 78 | 0.86 | 0.90 | 0.14 | 0.03 | 0.02 |
| 79 | 0.88 | 0.97 | 0.88 | 0.87 | 0.94 |
| 80 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| 81 | 0.34 | 0.93 | 0.89 | 0.85 | 0.96 |
| 82 | 0.89 | 0.93 | 0.92 | 0.12 | 0.05 |
| 83 | 0.88 | 0.01 | 0.02 | 0.05 | 0.94 |
| 84 | 0.86 | 0.95 | 0.00 | 0.02 | 0.94 |
| 85 | 0.00 | 0.87 | 0.01 | 0.00 | 0.02 |
| 86 | 0.00 | 0.78 | 0.00 | 0.00 | 0.02 |
| 87 | 0.20 | 0.93 | 0.91 | 0.80 | 0.95 |
| 88 | 0.84 | 0.87 | 0.87 | 0.72 | 0.95 |
| 89 | 0.70 | 0.38 | 0.01 | 0.01 | 0.01 |
| 90 | 0.01 | 0.70 | 0.00 | 0.00 | 0.00 |
| 91 | 0.01 | 0.60 | 0.01 | 0.01 | 0.00 |
| 92 | 0.84 | 0.96 | 0.82 | 0.90 | 0.96 |
| 93 | 0.00 | 0.97 | 0.96 | 0.00 | 0.00 |
| 94 | 0.83 | 0.02 | 0.05 | 0.00 | 0.00 |
| 95 | 0.88 | 0.98 | 0.86 | 0.97 | 0.95 |
| 96 | 0.88 | 0.92 | 0.86 | 0.71 | 0.96 |
| 97 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| 98 | 0.00 | 0.84 | 0.05 | 0.01 | 0.03 |
| 99 | 0.85 | 0.93 | 0.81 | 0.67 | 0.95 |

Table A.4: Score - Sentiment analysis

## A.2 Toxicity Detection on Civil Comment - Full Tables

**Percentage of highlighted text**

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |

| | | | | | |
|---|---|---|---|---|---|
| 0 | 40% | 60% | 40% | 100% | 20% |
| 1 | 32% | 3% | 38% | 43% | 8% |
| 2 | 18% | 11% | 11% | 100% | 7% |
| 3 | 28% | 30% | 20% | 38% | 11% |
| 4 | 22% | 6% | 17% | 6% | 6% |
| 5 | 34% | 60% | 25% | 21% | 9% |
| 6 | 33% | 10% | 38% | 44% | 11% |
| 7 | 28% | 17% | 28% | 48% | 14% |
| 8 | 36% | 15% | 43% | 8% | 10% |
| 9 | 25% | 49% | 24% | 59% | 11% |
| 10 | 26% | 34% | 26% | 11% | 11% |
| 11 | 39% | 48% | 39% | 100% | 13% |
| 12 | 22% | 56% | 33% | 100% | 11% |
| 13 | 33% | 8% | 40% | 36% | 9% |
| 14 | 29% | 48% | 12% | 100% | 7% |
| 15 | 44% | 56% | 33% | 100% | 22% |
| 16 | 36% | 27% | 36% | 100% | 9% |
| 17 | 41% | 6% | 29% | 100% | 12% |
| 18 | 23% | 13% | 26% | 18% | 10% |
| 19 | 24% | 13% | 19% | 42% | 7% |
| 20 | 11% | 4% | 11% | 44% | 11% |
| 21 | 33% | 44% | 41% | 33% | 11% |
| 22 | 33% | 17% | 33% | 100% | 17% |
| 23 | 33% | 42% | 25% | 100% | 8% |
| 24 | 21% | 47% | 32% | 47% | 16% |
| 25 | 30% | 71% | 47% | 18% | 9% |
| 26 | 27% | 29% | 29% | 33% | 10% |
| 27 | 30% | 2% | 10% | 7% | 11% |
| 28 | 24% | 9% | 21% | 64% | 3% |
| 29 | 25% | 2% | 25% | 17% | 6% |
| 30 | 23% | 11% | 27% | 28% | 11% |
| 31 | 40% | 33% | 35% | 19% | 9% |
| 32 | 8% | 37% | 44% | 32% | 8% |
| 33 | 38% | 31% | 25% | 33% | 10% |
| 34 | 22% | 7% | 41% | 30% | 11% |
| 35 | 35% | 46% | 35% | 62% | 13% |

| | | | | | |
|---|---|---|---|---|---|
| 36 | 26% | 33% | 23% | 33% | 7% |
| 37 | 33% | 33% | 33% | 100% | 22% |
| 38 | 20% | 10% | 45% | 100% | 15% |
| 39 | 27% | 6% | 27% | 12% | 6% |
| 40 | 26% | 68% | 10% | 38% | 9% |
| 41 | 44% | 4% | 4% | 8% | 12% |
| 42 | 31% | 8% | 15% | 100% | 8% |
| 43 | 36% | 43% | 34% | 33% | 10% |
| 44 | 26% | 21% | 21% | 37% | 16% |
| 45 | 27% | 30% | 30% | 80% | 13% |
| 46 | 25% | 33% | 52% | 15% | 8% |
| 47 | 17% | 17% | 50% | 100% | 17% |
| 48 | 25% | 70% | 24% | 18% | 7% |
| 49 | 19% | 15% | 38% | 36% | 10% |
| 50 | 26% | 6% | 19% | 46% | 11% |
| 51 | 35% | 51% | 24% | 45% | 11% |
| 52 | 35% | 58% | 26% | 55% | 10% |
| 53 | 37% | 55% | 37% | 8% | 11% |
| 54 | 41% | 2% | 33% | 26% | 8% |
| 55 | 25% | 22% | 31% | 52% | 11% |
| 56 | 20% | 20% | 20% | 40% | 20% |
| 57 | 39% | 17% | 17% | 89% | 11% |
| 58 | 29% | 57% | 29% | 100% | 29% |
| 59 | 27% | 9% | 9% | 55% | 9% |
| 60 | 42% | 33% | 17% | 100% | 17% |
| 61 | 22% | 17% | 28% | 44% | 11% |
| 62 | 28% | 4% | 11% | 35% | 11% |
| 63 | 31% | 24% | 13% | 9% | 6% |
| 64 | 22% | 17% | 11% | 11% | 11% |
| 65 | 25% | 8% | 33% | 100% | 8% |
| 66 | 29% | 11% | 45% | 11% | 11% |
| 67 | 30% | 21% | 18% | 15% | 15% |
| 68 | 21% | 29% | 36% | 50% | 21% |
| 69 | 33% | 12% | 24% | 76% | 10% |
| 70 | 37% | 31% | 25% | 41% | 12% |
| 71 | 42% | 47% | 37% | 100% | 16% |

| | | | | | |
|---|---|---|---|---|---|
| 72 | 38% | 28% | 41% | 28% | 7% |
| 73 | 33% | 36% | 33% | 67% | 6% |
| 74 | 29% | 14% | 14% | 100% | 14% |
| 75 | 36% | 32% | 32% | 39% | 14% |
| 76 | 14% | 14% | 29% | 86% | 14% |
| 77 | 40% | 20% | 60% | 100% | 20% |
| 78 | 17% | 33% | 67% | 100% | 0% |
| 79 | 32% | 6% | 19% | 65% | 10% |
| 80 | 37% | 47% | 28% | 30% | 12% |
| 81 | 26% | 23% | 26% | 38% | 10% |
| 82 | 25% | 4% | 8% | 33% | 8% |
| 83 | 38% | 38% | 25% | 100% | 25% |
| 84 | 36% | 36% | 45% | 100% | 18% |
| 85 | 22% | 19% | 22% | 26% | 9% |
| 86 | 18% | 18% | 28% | 73% | 15% |
| 87 | 27% | 18% | 27% | 50% | 18% |
| 88 | 25% | 88% | 50% | 100% | 13% |
| 89 | 40% | 10% | 30% | 90% | 10% |
| 90 | 18% | 36% | 36% | 100% | 9% |
| 91 | 38% | 50% | 44% | 100% | 19% |
| 92 | 34% | 55% | 41% | 48% | 7% |
| 93 | 37% | 19% | 30% | 100% | 15% |
| 94 | 38% | 41% | 32% | 32% | 12% |
| 95 | 33% | 33% | 33% | 100% | 33% |
| 96 | 20% | 60% | 60% | 100% | 20% |
| 97 | 26% | 94% | 49% | 37% | 9% |
| 98 | 30% | 20% | 50% | 100% | 20% |
| 99 | 28% | 54% | 30% | 60% | 10% |

Table A.5: Percentage of highlighted text - Toxicity detection

**Variation of prediction**

| | **LIME** | **T-EBAnO** | | | **SHAP** | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | | | **MLWE** | **POS** | **SEN** | |
| | **Absolut Variation** | | | | | **Relative Variation** | | | | |
| 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 1 | -0.98 | -0.98 | -0.98 | -0.98 | -0.98 | -100% | -100% | -100% | -100% | -100% |
| 2 | -1.00 | -1.00 | -0.99 | -1.00 | -1.00 | -100% | -100% | -99% | -100% | -100% |
| 3 | -0.01 | -0.01 | 0.00 | 0.00 | 0.00 | -1% | -1% | 0% | 0% | 0% |
| 4 | -0.92 | -0.92 | -0.79 | -0.92 | -0.92 | -100% | -100% | -86% | -100% | -100% |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 7 | -0.84 | -0.80 | -0.27 | -0.04 | 0.03 | -87% | -83% | -28% | -4% | 3% |
| 8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 9 | -0.01 | -0.03 | 0.00 | -0.01 | 0.00 | -1% | -3% | 0% | -1% | 0% |
| 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 16 | -1.00 | -1.00 | -1.00 | -1.00 | 0.00 | -100% | -100% | -100% | -100% | 0% |
| 17 | -0.99 | -0.99 | -0.99 | -0.99 | -0.99 | -100% | -100% | -100% | -100% | -100% |
| 18 | -0.12 | -0.06 | -0.01 | -0.01 | -0.01 | -12% | -6% | -1% | -1% | -1% |
| 19 | -0.17 | -0.06 | -0.05 | -0.14 | -0.04 | -18% | -6% | -5% | -14% | -4% |
| 20 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 22 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 24 | -0.01 | -0.93 | 0.00 | -0.93 | 0.00 | -1% | -93% | 0% | -93% | 0% |
| 25 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0% | -1% | 0% | 0% | 0% |
| 26 | 0.00 | 0.00 | 0.00 | -0.01 | 0.00 | 0% | 0% | 0% | -1% | 0% |
| 27 | -0.98 | -0.98 | -0.98 | -0.98 | -0.98 | -100% | -100% | -100% | -100% | -100% |
| 28 | -0.99 | -0.98 | -0.98 | -0.99 | -0.96 | -100% | -100% | -100% | -100% | -98% |
| 29 | -0.99 | -0.98 | -0.99 | -0.99 | -0.99 | -100% | -99% | -100% | -99% | -100% |
| 30 | -0.95 | -0.92 | -0.90 | -0.91 | -0.94 | -97% | -93% | -92% | -93% | -96% |
| 31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 32 | 0.06 | -0.60 | -0.83 | -0.35 | 0.02 | 7% | -67% | -92% | -39% | 3% |
| 33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 34 | -1.00 | -0.99 | -1.00 | -0.99 | -0.99 | -100% | -100% | -100% | -99% | -99% |
| 35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 36 | -0.01 | -0.01 | 0.00 | 0.00 | 0.00 | -1% | -1% | 0% | 0% | 0% |
| 37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 38 | 0.02 | -0.93 | -0.89 | 0.04 | -0.93 | 2% | -97% | -93% | 4% | -97% |
| 39 | -0.99 | -0.99 | -0.99 | -0.99 | -0.99 | -100% | -100% | -100% | -100% | -100% |
| 40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 41 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 42 | -1.00 | -1.00 | -1.00 | -1.00 | 0.00 | -100% | -100% | -100% | -100% | 0% |
| 43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 45 | -0.01 | -0.02 | 0.00 | -0.01 | 0.00 | -1% | -2% | 0% | -1% | 0% |
| 46 | -0.02 | -0.01 | -0.01 | 0.00 | -0.01 | -2% | -1% | -1% | 0% | -1% |
| 47 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 48 | -0.01 | -0.04 | -0.01 | 0.00 | 0.00 | -1% | -4% | -1% | 0% | 0% |
| 49 | -0.89 | -0.04 | -0.09 | -0.96 | -0.05 | -90% | -4% | -9% | -97% | -5% |
| 50 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | -1% | 0% | 0% | 0% | 0% |
| 51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 54 | -0.98 | -0.98 | -0.98 | -0.98 | -0.98 | -100% | -100% | -100% | -100% | -100% |
| 55 | -0.02 | -0.01 | 0.00 | 0.00 | -0.01 | -2% | -1% | 0% | 0% | -1% |
| 56 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 58 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0% | -1% | 0% | 0% | 0% |
| 59 | -1.00 | -0.98 | -0.98 | -1.00 | -0.98 | -100% | -98% | -98% | -100% | -98% |
| 60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 61 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 62 | -0.99 | -0.96 | -0.96 | -0.99 | -0.99 | -100% | -97% | -97% | -100% | -100% |
| 63 | -0.01 | -0.01 | 0.00 | 0.00 | 0.00 | -1% | -1% | 0% | 0% | 0% |
| 64 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 65 | -1.00 | -0.99 | -1.00 | -1.00 | -0.99 | -100% | -100% | -100% | -100% | -100% |
| 66 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 67 | -0.01 | 0.00 | 0.00 | 0.00 | -0.01 | -1% | 0% | 0% | 0% | -1% |
| 68 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 69 | -1.00 | -0.99 | -0.99 | -0.97 | -0.07 | -100% | -100% | -99% | -98% | -7% |
| 70 | -0.07 | -0.03 | -0.04 | -0.92 | -0.01 | -7% | -3% | -4% | -92% | -1% |
| 71 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 72 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 73 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 74 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 75 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 76 | -1.00 | -1.00 | -0.99 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 78 | 0.00 | -0.77 | 0.02 | 0.02 | 0.00 | 0% | -78% | 2% | 2% | 0% |
| 79 | -0.99 | -0.98 | -0.90 | -0.99 | -0.98 | -100% | -99% | -91% | -100% | -98% |
| 80 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | -1% | 0% | 0% | 0% | 0% |
| 81 | -0.01 | -0.01 | -0.01 | 0.00 | 0.00 | -1% | -1% | -1% | 0% | 0% |
| 82 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 83 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 84 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 85 | -0.88 | -0.80 | -0.83 | -0.83 | -0.02 | -97% | -89% | -92% | -91% | -3% |
| 86 | -0.94 | -0.92 | -0.88 | 0.02 | -0.92 | -97% | -95% | -91% | 2% | -95% |
| 87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 88 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 89 | -1.00 | -0.99 | -0.98 | -1.00 | -0.03 | -100% | -99% | -98% | -100% | -3% |
| 90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 91 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 92 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 95 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -100% | -100% | -100% | -100% | -100% |
| 96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 97 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0% | 0% | 0% | 0% | 0% |
| 98 | -0.01 | -0.01 | 0.00 | 0.00 | -0.01 | -1% | -1% | 0% | 0% | -1% |
| 99 | -0.93 | -0.95 | -0.01 | -0.05 | -0.05 | -95% | -96% | -1% | -5% | -5% |

Table A.6: Variation of prediction - Toxicity Detection

### Time of elaboration

| **ID** | **LIME (s)** | **T-EBAnO (s)** | **SHAP (s)** | **ID** | **LIME (s)** | **T-EBAnO(s)** | **SHAP (s)** |
|---|---|---|---|---|---|---|---|
| 0 | 42 | 1 | 4 | 50 | 280 | 13 | 137 |
| 1 | 198 | 4 | 53 | 51 | 281 | 14 | 163 |
| 2 | 137 | 5 | 16 | 52 | 177 | 2 | 29 |
| 3 | 281 | 39 | 281 | 53 | 283 | 10 | 138 |
| 4 | 83 | 2 | 8 | 54 | 284 | 28 | 201 |
| 5 | 276 | 16 | 174 | 55 | 282 | 15 | 151 |

| 6 | 277 | 8 | 95 | 56 | 15 | 0 | 3 |
|---|---|---|---|---|---|---|---|
| 7 | 159 | 3 | 30 | 57 | 82 | 1 | 9 |
| 8 | 277 | 40 | 287 | 58 | 21 | 0 | 3 |
| 9 | 275 | 13 | 131 | 59 | 35 | 0 | 4 |
| 10 | 278 | 42 | 284 | 60 | 41 | 1 | 3 |
| 11 | 120 | 2 | 9 | 61 | 43 | 1 | 9 |
| 12 | 21 | 1 | 4 | 62 | 280 | 6 | 72 |
| 13 | 281 | 41 | 223 | 63 | 280 | 13 | 140 |
| 14 | 218 | 4 | 53 | 64 | 74 | 1 | 11 |
| 15 | 30 | 1 | 3 | 65 | 38 | 1 | 3 |
| 16 | 30 | 1 | 3 | 66 | 195 | 5 | 45 |
| 17 | 76 | 1 | 6 | 67 | 238 | 4 | 55 |
| 18 | 209 | 5 | 60 | 68 | 48 | 1 | 5 |
| 19 | 277 | 24 | 186 | 69 | 343 | 8 | 88 |
| 20 | 149 | 3 | 13 | 70 | 337 | 11 | 126 |
| 21 | 134 | 3 | 12 | 71 | 132 | 1 | 6 |
| 22 | 17 | 0 | 2 | 72 | 192 | 3 | 30 |
| 23 | 32 | 1 | 3 | 73 | 202 | 3 | 40 |
| 24 | 96 | 2 | 11 | 74 | 23 | 0 | 2 |
| 25 | 277 | 48 | 323 | 75 | 191 | 3 | 16 |
| 26 | 274 | 5 | 83 | 76 | 17 | 0 | 3 |
| 27 | 276 | 72 | 324 | 77 | 38 | 1 | 4 |
| 28 | 189 | 4 | 86 | 78 | 20 | 0 | 2 |
| 29 | 238 | 7 | 82 | 79 | 216 | 3 | 29 |
| 30 | 280 | 19 | 164 | 80 | 299 | 4 | 68 |
| 31 | 275 | 9 | 111 | 81 | 336 | 14 | 161 |
| 32 | 278 | 28 | 224 | 82 | 331 | 6 | 86 |
| 33 | 275 | 5 | 86 | 83 | 32 | 0 | 2 |
| 34 | 136 | 5 | 14 | 84 | 42 | 0 | 3 |
| 35 | 276 | 7 | 96 | 85 | 315 | 16 | 188 |
| 36 | 276 | 8 | 104 | 86 | 271 | 4 | 72 |
| 37 | 27 | 0 | 3 | 87 | 113 | 2 | 12 |
| 38 | 99 | 2 | 9 | 88 | 20 | 0 | 1 |
| 39 | 185 | 5 | 53 | 89 | 104 | 2 | 7 |
| 40 | 281 | 21 | 194 | 90 | 36 | 1 | 4 |
| 41 | 132 | 3 | 12 | 91 | 71 | 1 | 5 |
| 42 | 42 | 1 | 3 | 92 | 304 | 11 | 141 |
| 43 | 282 | 15 | 161 | 93 | 157 | 2 | 29 |
| 44 | 75 | 2 | 28 | 94 | 205 | 3 | 48 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 45 | 283 | 6 | 84 | 95 | 12 | 0 | 1 |
| 46 | 283 | 12 | 148 | 96 | 17 | 0 | 1 |
| 47 | 20 | 0 | 2 | 97 | 316 | 39 | 319 |
| 48 | 286 | 33 | 272 | 98 | 38 | 1 | 3 |
| 49 | 287 | 32 | 244 | 99 | 306 | 5 | 85 |

Table A.7: Elaboration time - Toxicity Detection

### Score

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |
| 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.81 | 0.99 | 0.77 | 0.72 | 0.96 |
| 2 | 0.90 | 0.94 | 0.94 | 0.00 | 0.96 |
| 3 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 |
| 4 | 0.87 | 0.97 | 0.85 | 0.97 | 0.97 |
| 5 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7 | 0.79 | 0.83 | 0.41 | 0.08 | 0.00 |
| 8 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 9 | 0.03 | 0.05 | 0.00 | 0.03 | 0.01 |
| 10 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 16 | 0.78 | 0.84 | 0.78 | 0.00 | 0.00 |
| 17 | 0.74 | 0.97 | 0.83 | 0.00 | 0.94 |
| 18 | 0.21 | 0.11 | 0.03 | 0.01 | 0.02 |
| 19 | 0.29 | 0.11 | 0.10 | 0.23 | 0.08 |
| 20 | 0.94 | 0.98 | 0.94 | 0.71 | 0.94 |
| 21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 22 | 0.80 | 0.91 | 0.80 | 0.00 | 0.91 |

| | | | | | |
|---|---|---|---|---|---|
| 23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 24 | 0.01 | 0.67 | 0.00 | 0.67 | 0.00 |
| 25 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| 26 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 |
| 27 | 0.83 | 0.99 | 0.95 | 0.96 | 0.94 |
| 28 | 0.86 | 0.95 | 0.88 | 0.53 | 0.97 |
| 29 | 0.86 | 0.99 | 0.86 | 0.91 | 0.97 |
| 30 | 0.86 | 0.91 | 0.81 | 0.81 | 0.93 |
| 31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 32 | 0.00 | 0.65 | 0.70 | 0.49 | 0.00 |
| 33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 34 | 0.87 | 0.96 | 0.74 | 0.82 | 0.94 |
| 35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 36 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 |
| 37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 38 | 0.00 | 0.93 | 0.69 | 0.00 | 0.90 |
| 39 | 0.84 | 0.97 | 0.84 | 0.94 | 0.97 |
| 40 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| 41 | 0.72 | 0.98 | 0.98 | 0.96 | 0.94 |
| 42 | 0.82 | 0.96 | 0.92 | 0.00 | 0.00 |
| 43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 45 | 0.03 | 0.04 | 0.00 | 0.01 | 0.01 |
| 46 | 0.04 | 0.01 | 0.02 | 0.00 | 0.01 |
| 47 | 0.91 | 0.91 | 0.67 | 0.00 | 0.91 |
| 48 | 0.02 | 0.07 | 0.02 | 0.01 | 0.01 |
| 49 | 0.86 | 0.07 | 0.16 | 0.78 | 0.10 |
| 50 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 |
| 51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 54 | 0.74 | 0.99 | 0.80 | 0.85 | 0.96 |
| 55 | 0.03 | 0.02 | 0.01 | 0.01 | 0.02 |
| 56 | 0.89 | 0.89 | 0.89 | 0.75 | 0.89 |
| 57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 58 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| 59 | 0.84 | 0.94 | 0.94 | 0.62 | 0.94 |
| 60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 61 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 62 | 0.84 | 0.96 | 0.93 | 0.79 | 0.94 |
| 63 | 0.02 | 0.01 | 0.00 | 0.00 | 0.01 |
| 64 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 65 | 0.86 | 0.95 | 0.80 | 0.00 | 0.95 |
| 66 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 67 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 |
| 68 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 69 | 0.80 | 0.94 | 0.86 | 0.38 | 0.12 |
| 70 | 0.13 | 0.06 | 0.07 | 0.72 | 0.02 |
| 71 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 72 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 73 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 74 | 0.83 | 0.92 | 0.92 | 0.00 | 0.92 |
| 75 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 76 | 0.92 | 0.92 | 0.83 | 0.25 | 0.92 |
| 77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 78 | 0.00 | 0.72 | 0.00 | 0.00 | 0.00 |
| 79 | 0.81 | 0.96 | 0.86 | 0.52 | 0.94 |
| 80 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| 81 | 0.03 | 0.01 | 0.01 | 0.00 | 0.01 |
| 82 | 0.85 | 0.98 | 0.96 | 0.80 | 0.96 |
| 83 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 84 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 85 | 0.86 | 0.85 | 0.84 | 0.82 | 0.05 |
| 86 | 0.89 | 0.88 | 0.81 | 0.00 | 0.90 |
| 87 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 88 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 89 | 0.75 | 0.94 | 0.82 | 0.18 | 0.06 |
| 90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 91 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 92 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 93 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

| | | | | | |
|---|---|---|---|---|---|
| 95 | 0.80 | 0.80 | 0.80 | 0.00 | 0.80 |
| 96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 97 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| 98 | 0.02 | 0.02 | 0.01 | 0.00 | 0.02 |
| 99 | 0.82 | 0.62 | 0.02 | 0.09 | 0.10 |

Table A.8: Score - Toxicity Detection

## A.3 Topic Classification on AG News - Full Tables

**Percentage of highlighted text**

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |
| 0 | 19.0% | 38.1% | 42.9% | 100.0% | 9.5% |
| 1 | 13.0% | 54.3% | 45.7% | 100.0% | 6.5% |
| 2 | 20.0% | 17.5% | 37.5% | 100.0% | 10.0% |
| 3 | 17.5% | 7.0% | 17.5% | 24.6% | 5.3% |
| 4 | 17.1% | 45.7% | 51.4% | 100.0% | 5.7% |
| 5 | 17.8% | 40.1% | 34.2% | 53.3% | 5.9% |
| 6 | 10.3% | 53.8% | 44.9% | 48.7% | 5.1% |
| 7 | 14.4% | 77.1% | 45.8% | 45.1% | 5.2% |
| 8 | 13.3% | 33.3% | 66.7% | 100.0% | 6.7% |
| 9 | 18.8% | 12.5% | 25.0% | 100.0% | 9.4% |
| 10 | 11.1% | 15.3% | 47.2% | 100.0% | 8.3% |
| 11 | 17.0% | 56.6% | 62.3% | 100.0% | 11.3% |
| 12 | 14.0% | 30.2% | 53.5% | 100.0% | 11.6% |
| 13 | 19.0% | 28.6% | 38.1% | 100.0% | 9.5% |
| 14 | 15.0% | 30.0% | 50.0% | 100.0% | 10.0% |
| 15 | 23.5% | 11.8% | 29.4% | 100.0% | 5.9% |
| 16 | 17.8% | 48.9% | 53.3% | 51.1% | 8.9% |
| 17 | 17.0% | 43.4% | 50.9% | 64.2% | 11.3% |

| | | | | | |
|---|---|---|---|---|---|
| 18 | 19.6% | 28.3% | 41.3% | 89.1% | 10.9% |
| 19 | 18.2% | 18.2% | 36.4% | 100.0% | 4.5% |
| 20 | 21.1% | 15.8% | 10.5% | 100.0% | 10.5% |
| 21 | 20.0% | 40.0% | 26.7% | 100.0% | 6.7% |
| 22 | 16.7% | 25.0% | 50.0% | 100.0% | 8.3% |
| 23 | 11.9% | 43.3% | 29.9% | 88.1% | 7.5% |
| 24 | 15.2% | 10.9% | 10.9% | 100.0% | 8.7% |
| 25 | 21.1% | 68.4% | 52.6% | 100.0% | 10.5% |
| 26 | 22.2% | 18.5% | 22.2% | 100.0% | 7.4% |
| 27 | 15.8% | 31.6% | 26.3% | 57.9% | 7.9% |
| 28 | 17.9% | 23.9% | 35.8% | 50.7% | 10.4% |
| 29 | 15.8% | 48.7% | 43.4% | 60.5% | 5.3% |
| 30 | 14.6% | 29.3% | 48.8% | 100.0% | 4.9% |
| 31 | 17.6% | 52.9% | 55.9% | 100.0% | 8.8% |
| 32 | 16.7% | 27.8% | 41.7% | 100.0% | 11.1% |
| 33 | 12.8% | 10.6% | 31.9% | 17.0% | 12.8% |
| 34 | 19.7% | 34.4% | 37.7% | 100.0% | 8.2% |
| 35 | 18.3% | 18.3% | 33.8% | 70.4% | 9.9% |
| 36 | 18.2% | 36.4% | 63.6% | 100.0% | 9.1% |
| 37 | 16.7% | 19.4% | 50.0% | 52.8% | 11.1% |
| 38 | 17.9% | 28.2% | 43.6% | 100.0% | 10.3% |
| 39 | 15.7% | 43.1% | 49.0% | 100.0% | 5.9% |
| 40 | 15.4% | 59.0% | 41.0% | 100.0% | 10.3% |
| 41 | 11.5% | 9.6% | 34.6% | 100.0% | 5.8% |
| 42 | 17.6% | 19.6% | 37.3% | 100.0% | 9.8% |
| 43 | 13.0% | 9.3% | 40.7% | 100.0% | 5.6% |
| 44 | 18.9% | 56.6% | 45.3% | 62.3% | 7.5% |
| 45 | 17.2% | 44.8% | 55.2% | 75.9% | 10.3% |
| 46 | 18.4% | 28.9% | 44.7% | 100.0% | 10.5% |
| 47 | 9.8% | 14.8% | 14.8% | 100.0% | 6.6% |
| 48 | 21.1% | 31.6% | 42.1% | 100.0% | 10.5% |
| 49 | 18.9% | 48.6% | 35.1% | 32.4% | 5.4% |
| 50 | 17.9% | 25.6% | 46.2% | 100.0% | 7.7% |
| 51 | 18.8% | 60.4% | 52.1% | 100.0% | 8.3% |
| 52 | 20.0% | 42.9% | 40.0% | 100.0% | 8.6% |
| 53 | 14.8% | 22.2% | 40.7% | 100.0% | 7.4% |

| | | | | | |
|---|---|---|---|---|---|
| 54 | 15.1% | 34.0% | 34.0% | 100.0% | 5.7% |
| 55 | 19.2% | 23.1% | 19.2% | 100.0% | 11.5% |
| 56 | 5.0% | 15.0% | 45.0% | 100.0% | 7.5% |
| 57 | 15.2% | 45.7% | 47.8% | 100.0% | 8.7% |
| 58 | 14.7% | 29.4% | 41.2% | 100.0% | 8.8% |
| 59 | 18.2% | 40.9% | 38.6% | 100.0% | 9.1% |
| 60 | 18.4% | 21.1% | 57.9% | 100.0% | 7.9% |
| 61 | 16.1% | 29.0% | 41.9% | 100.0% | 9.7% |
| 62 | 15.2% | 72.2% | 40.5% | 54.4% | 8.9% |
| 63 | 11.1% | 13.9% | 52.8% | 100.0% | 8.3% |
| 64 | 19.5% | 43.9% | 53.7% | 100.0% | 9.8% |
| 65 | 20.0% | 10.0% | 15.0% | 100.0% | 10.0% |
| 66 | 14.6% | 26.8% | 56.1% | 100.0% | 9.8% |
| 67 | 20.4% | 14.3% | 40.8% | 100.0% | 8.2% |
| 68 | 23.1% | 50.0% | 53.8% | 100.0% | 11.5% |
| 69 | 14.6% | 31.7% | 46.3% | 22.0% | 9.8% |
| 70 | 15.8% | 31.6% | 44.7% | 100.0% | 10.5% |
| 71 | 15.1% | 45.2% | 56.2% | 100.0% | 4.1% |
| 72 | 16.0% | 32.0% | 56.0% | 100.0% | 8.0% |
| 73 | 7.1% | 14.3% | 35.7% | 100.0% | 7.1% |
| 74 | 11.4% | 20.5% | 43.2% | 100.0% | 9.1% |
| 75 | 14.6% | 17.1% | 43.9% | 100.0% | 9.8% |
| 76 | 18.2% | 18.2% | 45.5% | 100.0% | 6.8% |
| 77 | 14.3% | 32.1% | 42.9% | 100.0% | 10.7% |
| 78 | 20.5% | 35.9% | 38.5% | 100.0% | 7.7% |
| 79 | 22.2% | 27.8% | 61.1% | 100.0% | 11.1% |
| 80 | 17.1% | 25.7% | 54.3% | 100.0% | 2.9% |
| 81 | 12.5% | 20.8% | 41.7% | 100.0% | 4.2% |
| 82 | 17.4% | 23.9% | 41.3% | 100.0% | 10.9% |
| 83 | 17.2% | 22.4% | 37.9% | 100.0% | 8.6% |
| 84 | 15.0% | 12.5% | 50.0% | 75.0% | 7.5% |
| 85 | 22.2% | 19.4% | 52.8% | 100.0% | 8.3% |
| 86 | 15.6% | 31.1% | 53.3% | 100.0% | 11.1% |
| 87 | 18.2% | 2.3% | 45.5% | 100.0% | 4.5% |
| 88 | 18.9% | 12.2% | 35.1% | 45.9% | 6.8% |
| 89 | 12.8% | 18.1% | 46.8% | 47.9% | 5.3% |

| | | | | | |
|---|---|---|---|---|---|
| 90 | 20.0% | 22.5% | 45.0% | 100.0% | 10.0% |
| 91 | 21.1% | 15.8% | 26.3% | 100.0% | 10.5% |
| 92 | 14.3% | 20.0% | 31.4% | 100.0% | 8.6% |
| 93 | 16.2% | 48.6% | 54.1% | 100.0% | 5.4% |
| 94 | 11.8% | 14.7% | 41.2% | 85.3% | 11.8% |
| 95 | 17.6% | 41.2% | 50.0% | 100.0% | 11.8% |
| 96 | 14.3% | 40.0% | 48.6% | 100.0% | 8.6% |
| 97 | 10.6% | 17.6% | 48.2% | 100.0% | 8.2% |
| 98 | 21.3% | 34.0% | 46.8% | 100.0% | 6.4% |
| 99 | 18.4% | 34.7% | 36.7% | 100.0% | 10.2% |

Table A.9: Percentage of highlighted text - Topic Classification

### Variation of prediction

| | **LIME** | **T-EBAnO** | | | **SHAP** | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | | | **MLWE** | **POS** | **SEN** | |
| | **Absolut Variation** | | | | | **Relative Variation** | | | | |
| 0 | -0.06 | -0.98 | -0.94 | -0.92 | 0.00 | -6% | -100% | -96% | -93% | 0% |
| 1 | 0.00 | -0.83 | -0.89 | -0.74 | 0.00 | 0% | -83% | -89% | -74% | 0% |
| 2 | -0.10 | -0.94 | -0.89 | -0.73 | -0.02 | -10% | -95% | -89% | -74% | -2% |
| 3 | -0.93 | -0.91 | -0.93 | -0.93 | -0.91 | -99% | -98% | -99% | -99% | -97% |
| 4 | -0.91 | -0.94 | -0.61 | -0.69 | -0.94 | -95% | -98% | -64% | -72% | -99% |
| 5 | -0.04 | -0.82 | -0.04 | -0.11 | -0.03 | -4% | -84% | -4% | -11% | -3% |
| 6 | 0.00 | -0.99 | -0.01 | 0.00 | 0.00 | 0% | -99% | -1% | 0% | 0% |
| 7 | 0.00 | -0.53 | 0.00 | 0.00 | 0.00 | 0% | -53% | 0% | 0% | 0% |
| 8 | -0.05 | -0.69 | -0.58 | -0.36 | -0.06 | -5% | -69% | -59% | -36% | -6% |
| 9 | -0.96 | -0.95 | -0.93 | -0.70 | -0.86 | -99% | -99% | -96% | -73% | -90% |
| 10 | -0.96 | -0.96 | -0.89 | -0.73 | -0.89 | -97% | -97% | -90% | -73% | -90% |
| 11 | -0.01 | -0.92 | -0.35 | -0.73 | 0.00 | -1% | -92% | -35% | -74% | 0% |
| 12 | -0.08 | -0.93 | -0.69 | -0.73 | -0.03 | -8% | -93% | -69% | -74% | -3% |
| 13 | -0.31 | -0.90 | -0.44 | -0.72 | -0.02 | -31% | -92% | -44% | -73% | -2% |
| 14 | -0.38 | -0.88 | -0.61 | -0.66 | -0.90 | -41% | -96% | -66% | -71% | -98% |
| 15 | -0.74 | -0.90 | -0.67 | -0.69 | -0.90 | -77% | -94% | -70% | -72% | -94% |
| 16 | -0.06 | -0.95 | -0.26 | -0.23 | -0.01 | -6% | -96% | -26% | -23% | -1% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 17 | -0.01 | -0.99 | -0.95 | 0.00 | 0.00 | -1% | -99% | -95% | 0% | 0% |
| 18 | -0.18 | -0.95 | -0.30 | -0.91 | -0.04 | -18% | -96% | -31% | -92% | -4% |
| 19 | -0.90 | -0.92 | -0.92 | -0.88 | 0.03 | -96% | -97% | -97% | -93% | 4% |
| 20 | -0.24 | -0.84 | -0.84 | -0.86 | -0.84 | -26% | -91% | -91% | -93% | -91% |
| 21 | -0.09 | -0.76 | -0.92 | -0.73 | -0.01 | -9% | -76% | -93% | -74% | -1% |
| 22 | -0.49 | -0.80 | -0.78 | -0.71 | -0.02 | -51% | -82% | -80% | -73% | -2% |
| 23 | -0.85 | -0.92 | -0.91 | -0.65 | -0.81 | -92% | -99% | -98% | -70% | -88% |
| 24 | -0.50 | -0.50 | -0.44 | -0.44 | -0.50 | -100% | -100% | -86% | -86% | -99% |
| 25 | -0.01 | -0.95 | -0.78 | -0.73 | 0.00 | -1% | -95% | -78% | -74% | 0% |
| 26 | 0.02 | -0.52 | -0.50 | -0.92 | 0.00 | 2% | -54% | -52% | -96% | 0% |
| 27 | 0.00 | -0.01 | -0.01 | 0.00 | -1.00 | 0% | -1% | -1% | 0% | -100% |
| 28 | 0.00 | -0.07 | -0.78 | 0.00 | -0.05 | 0% | -7% | -78% | 0% | -5% |
| 29 | -0.29 | -0.28 | -0.02 | 0.00 | 0.00 | -29% | -28% | -2% | 0% | 0% |
| 30 | -0.02 | -0.01 | -0.06 | -0.96 | -0.03 | -2% | -1% | -6% | -96% | -3% |
| 31 | 0.00 | -0.01 | -0.21 | -0.96 | 0.00 | 0% | -1% | -21% | -96% | 0% |
| 32 | 0.00 | -0.12 | -0.04 | -0.37 | -0.01 | 0% | -12% | -4% | -37% | -1% |
| 33 | -0.98 | -0.99 | -0.01 | -0.99 | -0.99 | -98% | -100% | -1% | -100% | -99% |
| 34 | 0.00 | -0.18 | -0.98 | -0.37 | 0.00 | 0% | -18% | -98% | -37% | 0% |
| 35 | -0.95 | -0.69 | -0.95 | -0.02 | -0.10 | -96% | -69% | -95% | -2% | -10% |
| 36 | -0.71 | -0.93 | -0.66 | -0.89 | -0.74 | -74% | -97% | -68% | -93% | -77% |
| 37 | -0.99 | -0.99 | -0.58 | -0.98 | -0.99 | -100% | -99% | -58% | -99% | -99% |
| 38 | -0.99 | -1.00 | -0.91 | -0.37 | 0.00 | -99% | -100% | -91% | -37% | 0% |
| 39 | 0.00 | -0.15 | -0.50 | -0.37 | 0.00 | 0% | -15% | -50% | -37% | 0% |
| 40 | 0.00 | -0.23 | -0.01 | -0.96 | 0.00 | 0% | -23% | -1% | -96% | 0% |
| 41 | -0.95 | -0.90 | -0.89 | -0.92 | -0.03 | -96% | -91% | -90% | -93% | -3% |
| 42 | -0.97 | -0.18 | -0.96 | -0.37 | -0.17 | -97% | -18% | -96% | -37% | -17% |
| 43 | -0.92 | -0.85 | -0.94 | -0.68 | -0.85 | -98% | -90% | -100% | -72% | -91% |
| 44 | 0.00 | -1.00 | 0.00 | 0.00 | 0.00 | 0% | -100% | 0% | 0% | 0% |
| 45 | 0.00 | -1.00 | -0.93 | 0.00 | 0.00 | 0% | -100% | -93% | 0% | 0% |
| 46 | -0.54 | -0.93 | -0.59 | -0.37 | -0.05 | -54% | -93% | -59% | -37% | -5% |
| 47 | -0.61 | -0.65 | -0.54 | -0.67 | -0.68 | -82% | -88% | -73% | -91% | -92% |
| 48 | -0.99 | -0.70 | -0.96 | -0.37 | -0.02 | -99% | -71% | -97% | -37% | -2% |
| 49 | -0.01 | -0.92 | -0.06 | -0.01 | 0.00 | -1% | -92% | -6% | -1% | 0% |
| 50 | -0.99 | -0.99 | -0.95 | -0.36 | -0.97 | -100% | -99% | -96% | -37% | -98% |
| 51 | 0.00 | -0.96 | -0.01 | -0.96 | 0.00 | 0% | -96% | -1% | -96% | 0% |
| 52 | -0.03 | -0.85 | -0.05 | -0.37 | -0.02 | -3% | -85% | -5% | -37% | -2% |
| 53 | -0.97 | -0.97 | -0.89 | -0.71 | -0.32 | -99% | -99% | -92% | -73% | -33% |
| 54 | 0.00 | -0.99 | -0.92 | -0.93 | 0.00 | 0% | -99% | -92% | -93% | 0% |
| 55 | -0.37 | -0.42 | -0.32 | -0.41 | -0.25 | -81% | -94% | -70% | -92% | -56% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 56 | -0.75 | -0.77 | -0.75 | -0.14 | -0.75 | -98% | -100% | -97% | -18% | -98% |
| 57 | 0.00 | -0.38 | -0.96 | -0.37 | 0.00 | 0% | -38% | -96% | -37% | 0% |
| 58 | 0.00 | -0.99 | -0.99 | -0.93 | 0.00 | 0% | -99% | -99% | -93% | 0% |
| 59 | 0.00 | -0.99 | -0.92 | -0.93 | 0.00 | 0% | -99% | -92% | -93% | 0% |
| 60 | 0.00 | -0.03 | -0.02 | -0.96 | 0.00 | 0% | -3% | -2% | -96% | 0% |
| 61 | -0.75 | -0.99 | -0.79 | -0.96 | -0.04 | -76% | -100% | -79% | -96% | -4% |
| 62 | 0.00 | -0.98 | 0.00 | 0.00 | 0.00 | 0% | -98% | 0% | 0% | 0% |
| 63 | -0.94 | -0.93 | -0.99 | -0.73 | -0.04 | -95% | -94% | -99% | -74% | -4% |
| 64 | -0.47 | -0.78 | -0.11 | -0.37 | -0.97 | -47% | -78% | -11% | -37% | -97% |
| 65 | -0.95 | -0.95 | -0.95 | -0.69 | -0.95 | -100% | -100% | -99% | -72% | -100% |
| 66 | 0.00 | -0.88 | -0.01 | -0.37 | -0.26 | 0% | -88% | -1% | -37% | -26% |
| 67 | 0.00 | -0.92 | -0.88 | -0.37 | -0.79 | 0% | -92% | -88% | -37% | -79% |
| 68 | 0.00 | -1.00 | -0.97 | -0.96 | 0.00 | 0% | -100% | -97% | -96% | 0% |
| 69 | -0.89 | -0.59 | -0.07 | -0.06 | -0.96 | -89% | -59% | -7% | -6% | -96% |
| 70 | -0.67 | -0.98 | -0.88 | -0.73 | -0.05 | -67% | -98% | -89% | -74% | -5% |
| 71 | -0.01 | -0.96 | -0.87 | -0.73 | 0.00 | -1% | -96% | -87% | -74% | 0% |
| 72 | -0.02 | -0.98 | -0.92 | -0.73 | 0.00 | -2% | -99% | -93% | -74% | 0% |
| 73 | -0.05 | -0.15 | -0.89 | -0.36 | -0.05 | -5% | -15% | -90% | -36% | -5% |
| 74 | -0.73 | -0.93 | -0.81 | -0.72 | -0.75 | -74% | -94% | -82% | -73% | -76% |
| 75 | -0.98 | -1.00 | -0.99 | -0.96 | -0.99 | -98% | -100% | -100% | -96% | -99% |
| 76 | -0.99 | -0.04 | -0.86 | -0.37 | 0.00 | -99% | -4% | -86% | -37% | 0% |
| 77 | -0.13 | -0.86 | -0.98 | -0.96 | 0.00 | -13% | -86% | -98% | -96% | 0% |
| 78 | -0.02 | -0.04 | -0.02 | -0.37 | 0.00 | -2% | -4% | -2% | -37% | 0% |
| 79 | -0.83 | -0.08 | -0.26 | -0.95 | -0.24 | -84% | -8% | -27% | -96% | -24% |
| 80 | -0.56 | -0.92 | -0.48 | -0.37 | -0.04 | -56% | -92% | -48% | -37% | -4% |
| 81 | -0.47 | -0.97 | -0.30 | -0.96 | -0.83 | -47% | -98% | -31% | -96% | -84% |
| 82 | -0.70 | -0.99 | -0.93 | -0.93 | -0.01 | -70% | -99% | -94% | -93% | -1% |
| 83 | -0.93 | -0.95 | -0.89 | -0.71 | -0.04 | -96% | -98% | -91% | -73% | -4% |
| 84 | -0.90 | -0.65 | -0.86 | 0.07 | -0.88 | -100% | -73% | -96% | 7% | -98% |
| 85 | -0.01 | 0.00 | -0.67 | -0.37 | 0.00 | -1% | 0% | -67% | -37% | 0% |
| 86 | -0.86 | -0.97 | -0.07 | -0.93 | -0.01 | -86% | -97% | -7% | -93% | -1% |
| 87 | -0.99 | -0.99 | -0.10 | -0.37 | -1.00 | -100% | -99% | -10% | -37% | -100% |
| 88 | 0.02 | -0.98 | -0.98 | -0.98 | -0.98 | 2% | -100% | -100% | -100% | -100% |
| 89 | -0.94 | -0.91 | -0.98 | -0.91 | -0.02 | -95% | -92% | -99% | -92% | -2% |
| 90 | -1.00 | -0.99 | -0.95 | -0.37 | -0.01 | -100% | -99% | -95% | -37% | -1% |
| 91 | -0.84 | -0.61 | -0.61 | -0.92 | -0.02 | -85% | -62% | -62% | -93% | -2% |
| 92 | -0.08 | -0.87 | -0.92 | -0.69 | -0.81 | -9% | -91% | -96% | -72% | -86% |
| 93 | 0.00 | -0.96 | -0.96 | -0.93 | 0.00 | 0% | -96% | -96% | -93% | 0% |
| 94 | -1.00 | -0.99 | -0.94 | -0.48 | -1.00 | -100% | -100% | -95% | -49% | -100% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 95 | 0.00 | -0.08 | -0.84 | -0.96 | 0.00 | 0% | -8% | -84% | -96% | 0% |
| 96 | 0.00 | 0.00 | 0.00 | -0.96 | 0.00 | 0% | 0% | 0% | -96% | 0% |
| 97 | 0.00 | -0.26 | -0.01 | -0.93 | 0.00 | 0% | -26% | -1% | -93% | 0% |
| 98 | 0.00 | -0.99 | -0.67 | -0.37 | 0.00 | 0% | -99% | -67% | -37% | 0% |
| 99 | -0.04 | -0.28 | -0.99 | -0.93 | 0.00 | -4% | -28% | -99% | -93% | 0% |

Table A.10: Variation of prediction - Topic Classification

### Time of elaboration

| ID | LIME (s) | T-EBAnO (s) | SHAP (s) | ID | LIME (s) | T-EBAnO(s) | SHAP (s) |
|---|---|---|---|---|---|---|---|
| 0 | 144 | 2 | 5 | 50 | 255 | 3 | 5 |
| 1 | 299 | 4 | 9 | 51 | 317 | 3 | 7 |
| 2 | 297 | 3 | 5 | 52 | 217 | 2 | 4 |
| 3 | 351 | 4 | 8 | 53 | 136 | 1 | 3 |
| 4 | 241 | 2 | 4 | 54 | 318 | 4 | 7 |
| 5 | 344 | 21 | 132 | 55 | 158 | 2 | 3 |
| 6 | 346 | 22 | 146 | 56 | 251 | 3 | 6 |
| 7 | 346 | 22 | 159 | 57 | 281 | 4 | 9 |
| 8 | 57 | 1 | 2 | 58 | 184 | 2 | 4 |
| 9 | 179 | 2 | 3 | 59 | 285 | 3 | 6 |
| 10 | 316 | 10 | 31 | 60 | 225 | 3 | 5 |
| 11 | 316 | 4 | 9 | 61 | 185 | 1 | 3 |
| 12 | 261 | 4 | 6 | 62 | 319 | 6 | 31 |
| 13 | 108 | 1 | 2 | 63 | 187 | 3 | 8 |
| 14 | 77 | 1 | 2 | 64 | 257 | 2 | 10 |
| 15 | 87 | 1 | 2 | 65 | 116 | 1 | 4 |
| 16 | 262 | 3 | 7 | 66 | 250 | 2 | 6 |
| 17 | 316 | 4 | 9 | 67 | 319 | 4 | 9 |
| 18 | 301 | 3 | 6 | 68 | 157 | 1 | 3 |
| 19 | 115 | 1 | 2 | 69 | 228 | 2 | 5 |
| 20 | 98 | 1 | 2 | 70 | 226 | 3 | 4 |
| 21 | 54 | 1 | 1 | 71 | 319 | 6 | 48 |
| 22 | 46 | 1 | 1 | 72 | 131 | 2 | 3 |
| 23 | 317 | 5 | 35 | 73 | 58 | 1 | 1 |
| 24 | 249 | 3 | 14 | 74 | 296 | 3 | 6 |
| 25 | 98 | 1 | 2 | 75 | 259 | 2 | 5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 26 | 165 | 2 | 3 | 76 | 277 | 3 | 6 |
| 27 | 319 | 9 | 29 | 77 | 154 | 1 | 2 |
| 28 | 319 | 6 | 22 | 78 | 257 | 3 | 5 |
| 29 | 318 | 6 | 21 | 79 | 100 | 1 | 2 |
| 30 | 241 | 4 | 8 | 80 | 211 | 2 | 4 |
| 31 | 210 | 2 | 3 | 81 | 102 | 1 | 3 |
| 32 | 213 | 3 | 5 | 82 | 292 | 3 | 8 |
| 33 | 317 | 3 | 8 | 83 | 320 | 5 | 26 |
| 34 | 317 | 6 | 18 | 84 | 266 | 2 | 14 |
| 35 | 316 | 6 | 35 | 85 | 239 | 3 | 6 |
| 36 | 117 | 1 | 4 | 86 | 317 | 2 | 7 |
| 37 | 221 | 2 | 11 | 87 | 310 | 3 | 6 |
| 38 | 249 | 3 | 12 | 88 | 319 | 6 | 28 |
| 39 | 319 | 4 | 18 | 89 | 320 | 9 | 73 |
| 40 | 241 | 2 | 11 | 90 | 243 | 3 | 8 |
| 41 | 318 | 4 | 15 | 91 | 94 | 1 | 2 |
| 42 | 319 | 4 | 8 | 92 | 221 | 2 | 10 |
| 43 | 320 | 5 | 9 | 93 | 217 | 2 | 11 |
| 44 | 320 | 4 | 7 | 94 | 193 | 2 | 15 |
| 45 | 154 | 2 | 3 | 95 | 198 | 2 | 7 |
| 46 | 232 | 3 | 5 | 96 | 207 | 2 | 6 |
| 47 | 318 | 6 | 18 | 97 | 320 | 11 | 72 |
| 48 | 231 | 2 | 4 | 98 | 318 | 4 | 8 |
| 49 | 224 | 2 | 4 | 99 | 320 | 3 | 8 |

Table A.11: Elaboration time - Topic Classification

**Score**

| | **LIME** | **T-EBAnO** | | | **SHAP** |
|---|---|---|---|---|---|
| **ID** | | **MLWE** | **POS** | **SEN** | |
| 0 | 0.11 | 0.76 | 0.72 | 0.00 | 0.00 |
| 1 | 0.01 | 0.59 | 0.67 | 0.00 | 0.00 |
| 2 | 0.18 | 0.88 | 0.73 | 0.00 | 0.04 |
| 3 | 0.90 | 0.95 | 0.90 | 0.86 | 0.96 |
| 4 | 0.89 | 0.70 | 0.55 | 0.00 | 0.96 |
| 5 | 0.07 | 0.70 | 0.07 | 0.18 | 0.06 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | 0.00 | 0.63 | 0.03 | 0.00 | 0.00 |
| 7 | 0.00 | 0.32 | 0.00 | 0.00 | 0.00 |
| 8 | 0.10 | 0.68 | 0.42 | 0.00 | 0.11 |
| 9 | 0.89 | 0.93 | 0.84 | 0.00 | 0.90 |
| 10 | 0.93 | 0.91 | 0.66 | 0.00 | 0.91 |
| 11 | 0.01 | 0.59 | 0.36 | 0.00 | 0.00 |
| 12 | 0.14 | 0.80 | 0.56 | 0.00 | 0.05 |
| 13 | 0.45 | 0.80 | 0.52 | 0.00 | 0.03 |
| 14 | 0.56 | 0.81 | 0.57 | 0.00 | 0.94 |
| 15 | 0.77 | 0.91 | 0.71 | 0.00 | 0.94 |
| 16 | 0.12 | 0.67 | 0.33 | 0.32 | 0.01 |
| 17 | 0.02 | 0.72 | 0.65 | 0.01 | 0.00 |
| 18 | 0.30 | 0.82 | 0.40 | 0.19 | 0.07 |
| 19 | 0.88 | 0.89 | 0.77 | 0.00 | 0.00 |
| 20 | 0.39 | 0.87 | 0.90 | 0.00 | 0.90 |
| 21 | 0.16 | 0.67 | 0.82 | 0.00 | 0.02 |
| 22 | 0.63 | 0.79 | 0.61 | 0.00 | 0.03 |
| 23 | 0.90 | 0.72 | 0.82 | 0.20 | 0.90 |
| 24 | 0.92 | 0.94 | 0.88 | 0.00 | 0.95 |
| 25 | 0.02 | 0.47 | 0.59 | 0.00 | 0.00 |
| 26 | 0.00 | 0.65 | 0.63 | 0.00 | 0.00 |
| 27 | 0.00 | 0.02 | 0.02 | 0.01 | 0.96 |
| 28 | 0.01 | 0.13 | 0.71 | 0.00 | 0.09 |
| 29 | 0.43 | 0.36 | 0.04 | 0.01 | 0.01 |
| 30 | 0.05 | 0.02 | 0.11 | 0.00 | 0.06 |
| 31 | 0.00 | 0.02 | 0.28 | 0.00 | 0.00 |
| 32 | 0.01 | 0.20 | 0.08 | 0.00 | 0.02 |
| 33 | 0.92 | 0.94 | 0.02 | 0.91 | 0.93 |
| 34 | 0.01 | 0.28 | 0.76 | 0.00 | 0.00 |
| 35 | 0.88 | 0.75 | 0.78 | 0.04 | 0.18 |
| 36 | 0.78 | 0.77 | 0.47 | 0.00 | 0.84 |
| 37 | 0.91 | 0.89 | 0.54 | 0.64 | 0.94 |
| 38 | 0.90 | 0.84 | 0.70 | 0.00 | 0.00 |
| 39 | 0.01 | 0.24 | 0.51 | 0.00 | 0.00 |
| 40 | 0.00 | 0.30 | 0.02 | 0.00 | 0.00 |
| 41 | 0.92 | 0.91 | 0.76 | 0.00 | 0.05 |

| | | | | | |
|---|---|---|---|---|---|
| 42 | 0.89 | 0.30 | 0.76 | 0.00 | 0.29 |
| 43 | 0.92 | 0.90 | 0.74 | 0.00 | 0.92 |
| 44 | 0.00 | 0.60 | 0.00 | 0.00 | 0.00 |
| 45 | 0.00 | 0.71 | 0.61 | 0.00 | 0.00 |
| 46 | 0.65 | 0.80 | 0.57 | 0.00 | 0.09 |
| 47 | 0.86 | 0.86 | 0.79 | 0.00 | 0.93 |
| 48 | 0.88 | 0.69 | 0.72 | 0.00 | 0.05 |
| 49 | 0.02 | 0.66 | 0.12 | 0.02 | 0.00 |
| 50 | 0.90 | 0.85 | 0.69 | 0.00 | 0.95 |
| 51 | 0.00 | 0.56 | 0.01 | 0.00 | 0.00 |
| 52 | 0.06 | 0.68 | 0.10 | 0.00 | 0.03 |
| 53 | 0.92 | 0.87 | 0.72 | 0.00 | 0.48 |
| 54 | 0.00 | 0.79 | 0.77 | 0.00 | 0.00 |
| 55 | 0.81 | 0.85 | 0.75 | 0.00 | 0.69 |
| 56 | 0.96 | 0.92 | 0.70 | 0.00 | 0.95 |
| 57 | 0.00 | 0.45 | 0.68 | 0.00 | 0.00 |
| 58 | 0.00 | 0.82 | 0.74 | 0.00 | 0.00 |
| 59 | 0.00 | 0.74 | 0.74 | 0.00 | 0.00 |
| 60 | 0.00 | 0.07 | 0.04 | 0.00 | 0.00 |
| 61 | 0.80 | 0.83 | 0.67 | 0.00 | 0.08 |
| 62 | 0.00 | 0.43 | 0.00 | 0.00 | 0.00 |
| 63 | 0.92 | 0.90 | 0.64 | 0.00 | 0.08 |
| 64 | 0.59 | 0.65 | 0.18 | 0.00 | 0.94 |
| 65 | 0.89 | 0.95 | 0.92 | 0.00 | 0.95 |
| 66 | 0.00 | 0.80 | 0.02 | 0.00 | 0.40 |
| 67 | 0.00 | 0.89 | 0.71 | 0.00 | 0.85 |
| 68 | 0.00 | 0.67 | 0.63 | 0.00 | 0.00 |
| 69 | 0.87 | 0.63 | 0.13 | 0.12 | 0.93 |
| 70 | 0.75 | 0.81 | 0.68 | 0.00 | 0.09 |
| 71 | 0.01 | 0.70 | 0.58 | 0.00 | 0.00 |
| 72 | 0.04 | 0.81 | 0.60 | 0.00 | 0.00 |
| 73 | 0.10 | 0.26 | 0.75 | 0.00 | 0.10 |
| 74 | 0.81 | 0.86 | 0.67 | 0.00 | 0.83 |
| 75 | 0.91 | 0.91 | 0.72 | 0.00 | 0.94 |
| 76 | 0.90 | 0.08 | 0.67 | 0.00 | 0.01 |
| 77 | 0.23 | 0.76 | 0.72 | 0.00 | 0.01 |

| | | | | | |
|---|---|---|---|---|---|
| 78 | 0.03 | 0.08 | 0.05 | 0.00 | 0.00 |
| 79 | 0.81 | 0.14 | 0.32 | 0.00 | 0.38 |
| 80 | 0.67 | 0.82 | 0.47 | 0.00 | 0.08 |
| 81 | 0.61 | 0.88 | 0.40 | 0.00 | 0.89 |
| 82 | 0.76 | 0.86 | 0.72 | 0.00 | 0.01 |
| 83 | 0.89 | 0.87 | 0.74 | 0.00 | 0.08 |
| 84 | 0.92 | 0.79 | 0.66 | 0.00 | 0.95 |
| 85 | 0.02 | 0.01 | 0.55 | 0.00 | 0.00 |
| 86 | 0.85 | 0.81 | 0.12 | 0.00 | 0.03 |
| 87 | 0.90 | 0.98 | 0.17 | 0.00 | 0.98 |
| 88 | 0.00 | 0.94 | 0.79 | 0.70 | 0.96 |
| 89 | 0.91 | 0.87 | 0.69 | 0.67 | 0.04 |
| 90 | 0.89 | 0.87 | 0.70 | 0.00 | 0.01 |
| 91 | 0.82 | 0.72 | 0.68 | 0.00 | 0.05 |
| 92 | 0.16 | 0.85 | 0.80 | 0.00 | 0.88 |
| 93 | 0.00 | 0.67 | 0.62 | 0.00 | 0.00 |
| 94 | 0.94 | 0.92 | 0.73 | 0.23 | 0.94 |
| 95 | 0.00 | 0.14 | 0.63 | 0.00 | 0.00 |
| 96 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| 97 | 0.01 | 0.40 | 0.01 | 0.00 | 0.00 |
| 98 | 0.01 | 0.79 | 0.59 | 0.00 | 0.00 |
| 99 | 0.07 | 0.39 | 0.77 | 0.00 | 0.01 |

Table A.12: Score - Topic Classification

# Bibliography

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL http://arxiv.org/abs/1903.04561.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

Daniel James Fuchs. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1):1, 2018.

Milo Honegger. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*, 2018.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021. ISSN 0004-3702.

doi: https://doi.org/10.1016/j.artint.2021.103473. URL https://www.sciencedirect.com/science/article/pii/S0004370221000242.

Zachary C. Lipton. The mythos of model interpretability, 2017.

ZC Lipton. The mythos of model interpretability. arxiv 2016. *arXiv preprint arXiv:1606.03490*, 2019.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

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

Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2017.10.011. URL https://www.sciencedirect.com/science/article/pii/S1051200417302385.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1097. URL https://aclanthology.org/N18-1097.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016a. URL http://arxiv.org/abs/1602.04938.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016b.

Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and machine learning*, pages 159–175. Springer, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, volume 40, pages 543–567. Wiley Online Library, 2021.

Francesco Ventura, Salvatore Greco, Daniele Apiletti, and Tania Cerquitelli. Explaining the deep natural language processing by mining textual interpretable features. *CoRR*, abs/2106.06697, 2021. URL https://arxiv.org/abs/2106.06697.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021. ISSN 2079-9292. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593.