Interpretability for Deep Learning Text Classifiers

by

Diana Lucaci

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
MCS degree in
Computer Science

School of Electrical Engineering and Computer Science (EECS)
Faculty of Engineering
University of Ottawa

Abstract

The ubiquitous presence of automated decision-making systems that have a performance comparable to humans brought attention towards the necessity of interpretability for the generated predictions. Whether the goal is predicting the system's behavior when the input changes, building user trust, or expert assistance in improving the machine learning methods, interpretability is paramount when the problem is not sufficiently validated in real applications, and when unacceptable results lead to significant consequences.

While for humans, there are no standard interpretations for the decisions they make, the complexity of the systems with advanced information-processing capacities conceals the detailed explanations for individual predictions, encapsulating them under layers of abstractions and complex mathematical operations. Interpretability for deep learning classifiers becomes, thus, a challenging research topic where the ambiguity of the problem statement allows for multiple exploratory paths.

Our work focuses on generating natural language interpretations for individual predictions of deep learning text classifiers. We propose a framework for extracting and identifying the phrases of the training corpus that influence the prediction confidence the most through unsupervised key phrase extraction and neural predictions. We assess the contribution margin that the added justification has when the deep learning model predicts the class probability of a text instance, by introducing and defining a contribution metric that allows one to quantify the fidelity of the explanation to the model. We assess both the performance impact of the proposed approach on the classification task as quantitative analysis and the quality of the generated justifications through extensive qualitative and error analysis.

This methodology manages to capture the most influencing phrases of the training corpus as explanations that reveal the linguistic features used for individual test predictions, allowing humans to predict the behavior of the deep learning classifier.

Acknowledgements

I would like to start by expressing my deep and sincere gratitude to my research supervisor, **Dr. Diana Inkpen**, School of Electrical Engineering and Computer Science, University of Ottawa for her extraordinary insights, valuable feedback, and patient guidance. This work would not have been materialized in the present form without her detailed observations and intellectual directions in the course of completion. It was a great privilege and honor to work and study under your supervision.

This research has been possible thanks to the generous funding support from the Natural Sciences and Engineering Research Council of Canada (NSERC) for which I am also truly grateful.

For the financial support without which it would not have been possible to start this AI Master program, and for the numerous career and networking opportunities, I would like to acknowledge the **Vector Institute** that helped me expand my career horizons.

This international adventure and countless other career opportunities would have not been possible if it were not for my dearest professor and friend **Dr. Corina Forascu**, professor at the Computer Science Faculty of "Alexandru Ioan Cuza" University of Iasi.

I would also like to express my thankfulness to all my professors and mentors from the University of Ottawa (MSc) and "Alexandru Ioan Cuza" University of Iasi (BSc) who paved my way to career success.

I am extremely grateful for my parents and grandparents, for their prayers, love, and sacrifices for my education. I am very much thankful for my friends from Iasi (Radu, Stefan, Valeriu, and Robert) and from Ottawa (Sophie, Mozghan, Andryi, and Vasileios) for their continuous support and cheers. Special thanks to Norbert for his patience, optimism, and valuable advice.

This thesis is dedicated, with love and appreciation, to my mother, who believes in me (sometimes more than I do) and encourages me at every step of my life.

Table of Contents

List of Tables

List of Figures

Acronyms

AGI Artificial General Intelligence. 8

AI Artificial Intelligence. 1, 8, 16, 32

ANN Artificial Neural Network. 12, 14

BLEU Bilingual Evaluation Understudy. 37, 40, 46

CBOW Continuous Bag of Words. 22

CNN Convolutional Neural Network. 13, 19, 33

CV Computer Vision. 9, 17

CVAE Conditional Variational Autoencoder. 46

DL Deep Learning. 1, 2, 14, 16, 50, 53

EF Explanation Factor. 46

GEF Generative Explanation Framework. 46

GloVe Global Vectors for Word Representation. 23

GRU Gated Recurrent Unit. 13, 15

HTML HyperText Markup Language. 49, 50

IMDB Internet Movie Database. 39, 49, 59

kNN k-Nearest Neighbors. 20, 44

LIME Local Interpretable Model-Agnostic Explanations. 36, 37, 42

LSTM Long Short-Term Memory. 13, 15, 33, 52, 64

ML Machine Learning. 1, 3, 4, 16, 53, 58

MLP Multilayer Perceptron. 4, 5, 12, 14, 19, 53

MT Machine Translation. 13

NLI Natural Language Inference. 36, 40

NLP Natural Language Processing. 2, 8, 13, 14, 17, 19, 21, 25, 33, 38, 45, 97

POS Part Of Speech. 27, 30

RAKE Rapid Automatic Keyword Extraction. 29

RNN Recurrent Neural Network. 13, 15, 19, 33

SNLI Stanford Natural Language Inference. 40

SP-LIME Submodular Pick - Local Interpretable Model-Agnostic Explanations. 44

TF-IDF Term Frequency–Inverse Document Frequency. 22, 51

VADER Valence Aware Dictionary and sEntiment Reasoner. 31, 78

VAE Variational Autoencoder. 13

XAI Explainable Artificial Intelligence. vi, 2, 8, 16–18, 32, 96, 98

Chapter 1

Introduction

In this chapter, we highlight the importance and the need for explainability, presenting our source of inspiration from human behavior and the motivation to achieve model interpretability through our proposed framework. We also include the research questions that we answer, the contributions brought in this thesis, and the outline of the next chapters.

1.1 Problem

State-of-the-art Machine Learning (ML) and Deep Learning (DL) techniques outperform humans on a number of tasks (Liu et al., 2019b; Yu et al., 2020; Badia et al., 2020), bringing the advantage of both better results (in terms of performance measures such as accuracy, precision, recall, F-score, or others, depending on the objective of the task) and fast predictions, which leads to speeding up the decision-making process. These techniques are thus present in many predictive systems in a wide variety of domains such as legal (Chhatwal et al., 2018), medical diagnoses (Vásquez Morales et al., 2019), financial, telecommunications, car driving, and others. The ubiquitous presence of automated decision-making systems drew the attention of the Artificial Intelligence (AI) research community and of the legislators [1] to the need of interpretability of the models.

[1] Regulation of the European Parliament and of the Council of 27 April 2016 (General data protection regulation)

Previous work on this topic emphasizes the necessity of explaining the decisions made by an autonomous system "in understandable terms to a human" (**?**), for reasons such as building user trust and expert assistance in improving the models, for scenarios when the problem is not sufficiently validated in real applications, and when unacceptable results lead to significant consequences.

Model explainability has emerged as a subtask of XAI and has been actively researched for purposes such as gathering insights about the model behavior, gaining the ability to make informed decisions when choosing between different models, building user trust, debugging, improving the performance through parameter tuning or architecture design with the help of the explanations, and expert assistance. While explainability refers to the ability to convey information about how the model behaves, interpretability focuses on the ability to correctly predict the model's result. A high level of interpretability can lead to model explainability.

Depending on the end-users of the generated explanations, there are different goals that the current research in XAI aims towards.

On one hand, the explanations that try to assess user trust in the model or in a certain prediction need to be easily understandable by humans and faithful to the model. For Natural Language Processing (NLP) applications, extractive methods for explainability accomplish this through attention-based approaches (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Pruthi et al., 2020), gradient-based approaches (Ancona et al., 2017), or feature contributions that are compared to reference data points (Shrikumar et al., 2017) in order to determine the correlation between a feature subset of the instance under classification and the predicted output. The extractive methods aim to achieve interpretable explanations by highlighting the part of the input that is sufficient for correct classification.

On the other hand, the explanations aimed for machine learning experts need to be more compelling and detailed, conveying insights into how the model operates and how it can be improved. Other challenges of the DL systems that model interpretability could solve are ensuring fairness, detecting the presence of bias, and allowing for further opti-

mizations for obtaining better performance in the desired direction (for example, avoiding discrimination). These goals have become particularly challenging also due to an incomplete problem formalization (Doshi-Velez and Kim, 2017).

Regardless of the end-user, the common goal of the explainable models is to present justifications for predictions and reveal insights about the model's behavior.

1.2 Motivation

For humans, the decision-making process is most of the time based on their past experiences and acquired knowledge. Whether people remembering similar contexts or situations, have some prior information or adapt what they know to a new situation, they always include their past experiences in the decisions

ML emerged as a field of Computer Science that provides systems the ability to learn and improve from experience in order to make predictions without being explicitly programmed, aiming to act like humans. In the learning process, the training algorithms use a labeled dataset in order to identify patterns of the data and their associated information or classes. Similarly to humans, at testing time, when no ground truth is present, the designed model makes predictions based on the *experience* that it accumulated.

The natural language explanations that humans employ to justify their behavior often refer to past experiences from which we learn and adapt. Inspired from this human habit, we are translating it to ML algorithms for text classifiers and we analyze what information of the training set is influencing a prediction for a machine learning model the most.

Popular previous approaches focus on extractive methods that highlight the part of the instance under classification that led the model towards a certain prediction, while our approach diverges by providing natural language explanations (keyphrases) generated from the training set rather than extracted from the instance that is being classified. Different optimization techniques or proxy models employed to accomplish this goal are discussed in detail in Chapter 3. This category of explanations is able to justify a prediction without making use of the training set, thus, missing information that has been used by

the model in the classification process. To alleviate this shortcoming, another category of explanations, most useful for image or text data, is through data points. This category of methods is easily interpretable and consists of extracting training instances that are either the most similar to the classified example or extracted as a prototype or criticism point of the considered classes (categories). Such methods, also used for data analysis and visualization, provide more insights into the training data.

Distancing from all these previous perspectives, we researched the architecture of an explanation generator model that determines the keyphrases of the training set that have the most influence on the prediction of an instance. Our proposed approach aims to make use of the training data when generating explanations, in an attempt to gather more insights into the most influential features (words and phrases) of the training set. Our motivation is to enhance the interpretability of a deep learning text classifier through individual prediction's explanations, exploiting both numerical metrics (most insightful for the architects of the ML models) and natural language explanations (for lay humans).

Thus, we propose an unsupervised framework for generating corpus-based explanations that enhance the prediction confidence of deep learning classifiers for a given instance under test. Analyzing the explanations that enhance the prediction provides insights on the feature that the model has learned during training, allowing the users to also identify the reasoning of the model. The predilection of the model towards a small set of explanations can reveal its incapacity of generalization or imbalance of the training data from a semantic point of view.

Proxy models are model-agnostic approaches that aim to reproduce the behavior of the original model (that is being explained) and also to provide interpretable explanations, often these are simple methods that loose performance (decision trees, rule-based, linear regression). To explain the decisions of a complex deep learning model, we aim to find a compromise between the elaborate architectures of models that achieve state of the art results and the simple interpretable algorithms. Thus, our proposed approach focuses on capturing the patterns learned by the model through Multilayer Perceptron (MLP) models (one of the most interpretable model category out of the class of neural algorithms) for

generating natural language justifications for individual decisions.

Furthermore, since the acquisition of gold-truth explanations is expensive, the applicability and feasibility of the unsupervised methods motivate us to focus in this direction. In our research work, we focus on developing a generative model for explaining individual predictions using natural language explanations. We showcase its effectiveness through both automatic evaluation and human evaluation when comparing two models with the goal of choosing which one is better.

1.3 Hypothesis

In this thesis, we answer the following research questions, supporting our conclusions with experimental results and analysis.

RQ1 How can one identify the most influential training phrases in an unsupervised manner?

Identifying the most influential phrases has multiple applications in interpretability, allowing humans to understand what is the basis of the decisions made by a classifier. We propose an unsupervised framework guided by desiderata for the generated justifications to identify the keyphrases of the training set that bring high contributions to the classification confidence for individual predictions.

RQ2 Is a MLP architecture able to generate explanations that improve individual predictions?

The motivation behind this research question is that MLPs are the building blocks of neural approaches, being the simplest architecture, and one of the most interpretable out of the class of deep neural networks. Our experiments began with a MLP architecture for a generator of explanations, that aims to decode the features learned by the main classifier. We have performed numerous experiments to determine the optimal number of layers, parameters, and hyperparameters using manual tuning

and we have found a model that ensures desiderata such as preserving the original classifier's performance and improving individual predictions.

RQ3 What evaluation methods can be employed to evaluate the quality of the generated explanations?

The usefulness of the proposed approach needs to be evaluated in the context of interpretability and its applications. To evaluate our explanation generation framework, we defined two desiderata for which we design evaluation measures and methods to highlight the results of the proposed method, employing both automatic evaluation and human analysis of the generated results.

RQ4 How can the proposed approach be used for interpretability?

To assess the practicability of the proposed approach, our extensive analysis proposes different applications for the interpretability task. The generated justifications are informative for semantic-based model comparison, semantic outlier detection for the training set, and model interpretability.

1.4 Contributions

This thesis brings the following contributions:

1. Introduces a new unsupervised framework for explanation generation;
2. Proposes an explanation generation architecture guided only by desiderata for the explanations (and not by gold-truth explanations);
3. Defines a new contribution metric to assess the quality of an explanation relative to the impact on the original classifier's predictions;
4. Presents and discusses the results of the experiments for explanation generation through both quantitative and qualitative analysis, aiming to explain the text classifier's behavior through model interpretability.

1.5 Outline

The rest of this thesis is structured as follows:

Chapter 2: This chapter presents the main concepts related to our research work, starting with a high-level introduction of the key concepts of the field and continuing with a more detailed description of the algorithms and approaches that have been used as part of our methodology.

Chapter 3: We then highlight the state-of-the art results and methods for the interpretability and explainability tasks, familiarizing the reader with the terminology specific to this subfield. The literature review summarizes the evaluation methods, the popular datasets used for these applications, and the most popular interpretability approaches.

Chapter 4: Next, we describe the proposed framework: starting with the description of the dataset, text preprocessing steps, continuing with the text mining stage for dictionary acquisition, and with the explanation generation model's architecture and ending with a description of the proposed evaluation methods.

Chapter 5: In this chapter, we motivate all the architectural and parameter choices, reporting the results of the hyperparameter tuning and describing how we achieve the proposed desiderata for the resulting explanations. This chapter also includes the results of different experiments for the employed dictionaries and alternative model architectures.

Chapter 6: The evaluation metrics for the interpretability task are reported in this chapter as quantitative analysis, followed by a qualitative analysis that includes human evaluation of the generated explanations.

Chapter 7: The last chapter presents the conclusions of our research work, summarizing the main findings and contributions, and proposing future directions of research.

Chapter 2

Background

This chapter presents a brief introduction of the main topics that our research work focuses on, placing our proposed methodology under the umbrella of XAI methods for Text Classification. Presenting a brief introduction of AI and its subfields, we also describe the text mining approaches adopted in previous work that we have used in our experiments for keyword and key phrase extraction.

2.1 AI

The high computational power that the computers brought have been exploited for decision-making purposes, with the goal of creating expert systems. AI emerged as a Computer Science field that aims to simulate human behavior and intelligence into machines, creating algorithms that are capable to learn, think, and act humanly. This field branches in two main categories: **Narrow AI** and **Artificial General Intelligence (AGI)**.

2.1.1 Narrow AI

To achieve the end goal of designing algorithms that are capable of making accurate and rational predictions, narrow tasks have been defined so that the focus of the algorithm is to optimize a certain objective. The tasks are from a number of fields such as NLP,

Computer Vision (CV), Robotics, Knowledge Representation, and numerous others, depending also on the underlying application from the real world that can benefit from the machine's computational power. Examples of such tasks are text classification, generation, summarization, image classification, object detection, robotic arms, and industrial robots.

2.1.2 Artificial General Intelligence

While tackling narrow tasks and obtaining high performance in particular contexts has always been the focus of the research community, a more ambitious goal is to create systems that are able to solve a wide range of problems, combining the acquired knowledge to solve complex tasks and make rational decisions. Alan Turing asked the question "Can machines think?" in 1950 (Turing, 1950), proposing the Turing Test. According to this test, the incapacity of a human to distinguish between two agents that respond to written questions, one of which is a human and the other an intelligent system, leads to establishing that the agent reached human-level intelligence. While the lay humans are highly influenced by the popular Science-Fiction scenarios, the current intelligent systems still lack the ability to reason in a human-like manner.

2.1.3 Machine Learning

Inspired by human behavior, the learning process has been transferred to Computer Science through Machine Learning – a field that researches the algorithms that can "automatically improve with experience" (Mitchell, 1997).

The learning process of such algorithms is defined as the ability of a system to improve its performance on a task, given some training experience. Different approaches to obtain incremental improvements make use of notions from fields such as Mathematics, Probability and Statistics, Information Theory, and Computational Optimization, drawing insights from large amounts of data in order to make predictions on unseen data. The algorithms vary in complexity and obtain different performances on a wide variety of tasks, ranging from simple rule-based, decision trees or instance-based learning, to statistical approaches

such as Bayesian inference and Maximum Likelihood Modeling.

Learning can occur in different settings: supervised, semi-supervised, or unsupervised. In this context, we will be using the terminology of training data – referring to labelled data that is used for training purposes only, validation data – which is labelled data used for evaluating the performance of the algorithm during training, and test data – a set of unseen labelled data points that are used to measure the generalization power of the algorithm.

2.1.3.1 Supervised Learning

In supervised learning, a collection of labelled data from which the algorithm draws statistical insights is being used. Common tasks of supervised learning are classification and regression, having applications in many fields, and being particularly useful when a considerable amount of labeled data exists.

2.1.3.2 Unsupervised Learning

When labeled data is unavailable or labelling data is expensive, time-consuming, or requiring domain experts, unsupervised algorithms perform analysis for clustering the data points, drawing insights in the absence of ground truth about the data.

2.1.3.3 Semi-Supervised Learning

When little labelled data is available, machine learning algorithms can also make use of unlabelled data to improve their performance, or use the labelled data for training and then for making predictions for the unlabeled instances.

2.1.4 Neural Networks

Inspired from the biological neurons of the human brain, the perceptron (artificial neuron) is the fundamental unit of a neural network. The original perceptron introduced in 1943

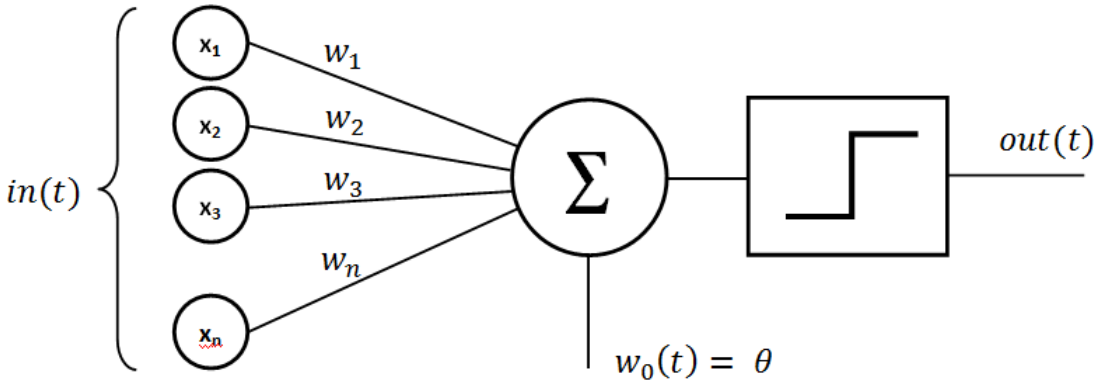


Figure 2.1: Graphical representation of a perceptron

(and later republished) by McCulloch and Pitts (1988) is also known as linear threshold gate and consists of a set of inputs $x_1, x_2, ..., x_n$, weighted by the values $w_1, w_2, ..., w_n$, a constant threshold value $\theta = w_0$, and an output y. The output is a binary value computed as a weighted sum of the input values (Equation 2.1) that is passed through a linear step function at threshold θ (Equation 2.2) that allows the neuron to fire (output 1) or not (output 0). The graphical representation can be observed in Figure 2.1.

$$Sum = \sum_{i=1}^{n} w_i \cdot x_i \tag{2.1}$$

$$y = f(Sum) = \begin{cases} 1, \text{if } Sum \geq \theta \\ 0, \text{otherwise} \end{cases} \tag{2.2}$$

The learning algorithm of the perceptron starts with random initialization of the weights, updating those values based on the training labelled input instances. Although this model has a high computing potential, there are two main limitations which led to the progress towards more complex architectures of multiple perceptronts: neural networks. Firstly, the output of the perceptron is binary, which limits the predictive power of the model. Secondly, the perceptron can only correctly classify linear separable input vectors.

The neural networks consist of layers of perceptrons that aggregate the input from the previous layer or the input layer and pass the net input (Equation 2.1) to an activation function that can differ from the classical step threshold function. Examples of commonly

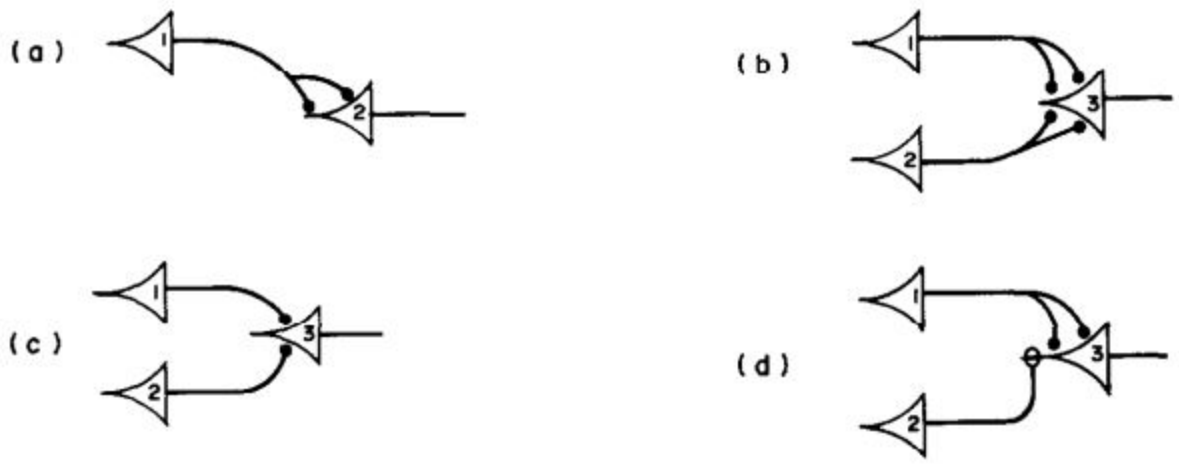


Figure 2.2: Graphical representation of Neural Networks from the original paper (McCulloch and Pitts, 1988)

used activation functions are $Softmax$, $Sigmoid$, $tanh$, $relu$. The representations of different chained perceptrons architectures are depicted in Figure 2.2. A network with more than one layer is called MLP (or Feed Forward Neural Network). The computation performed in order to get the output of a neural network is called a forward pass and consists of propagating the input information through the entire network up until the final layer. Having the gold truth of a training instance, the error of classification can be computed and propagated back in order to update the weights of the formed computational graph. This is achieved through the backpropagation algorithm that allows the backward pass of information from the output layer towards the input using the gradient of each involved parameter.

The advantage of these networks is their capability to learn non-linear patterns, once the hyperparameters such as the number of layers and the number of perceptrons for each layer are established.

2.1.5 Deep Learning

The increase of performance brought by the increase of complexity of the MLP models led to the rapid development of different improved architectures of the Artificial Neural

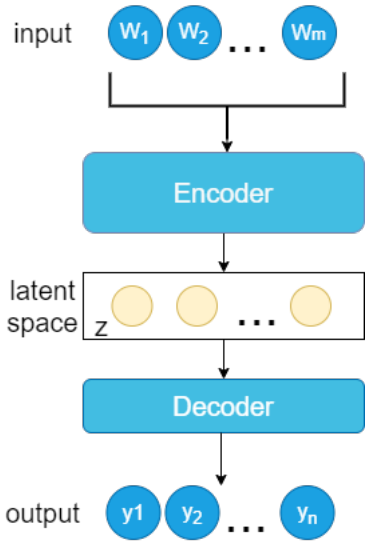


Figure 2.3: Encoder-Decoder Neural Network architecture

Network (ANN) model. The most popular models that achieved state of the art results for NLP applications at different points in time are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) – and their flavours: Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) –, Transformer, BERT, Variational Autoencoder (VAE).

2.1.5.1 Encoder-Decoder Architectures

In seq2seq (sequence-to-sequence) NLP applications such as Machine Translation (MT), the challenge is to capture the information of a variable-sized input (the sequence of words) and output another sequence of variable size output. The solution to such problems is to generate a fixed-sized internal representation that can be used for the task in question (translation, classification, text generation). This idea is commonly known as Encoder-Decoder models.

The general design pattern of the Encoder-Decoder Architectures is depicted in Figure 2.3. The variable-size input is encoded in a latent variable (inner representation) that is used as input for the decoder model to produce the sequence of output. The architecture of the Encoders and Decoders is usually a RNN.

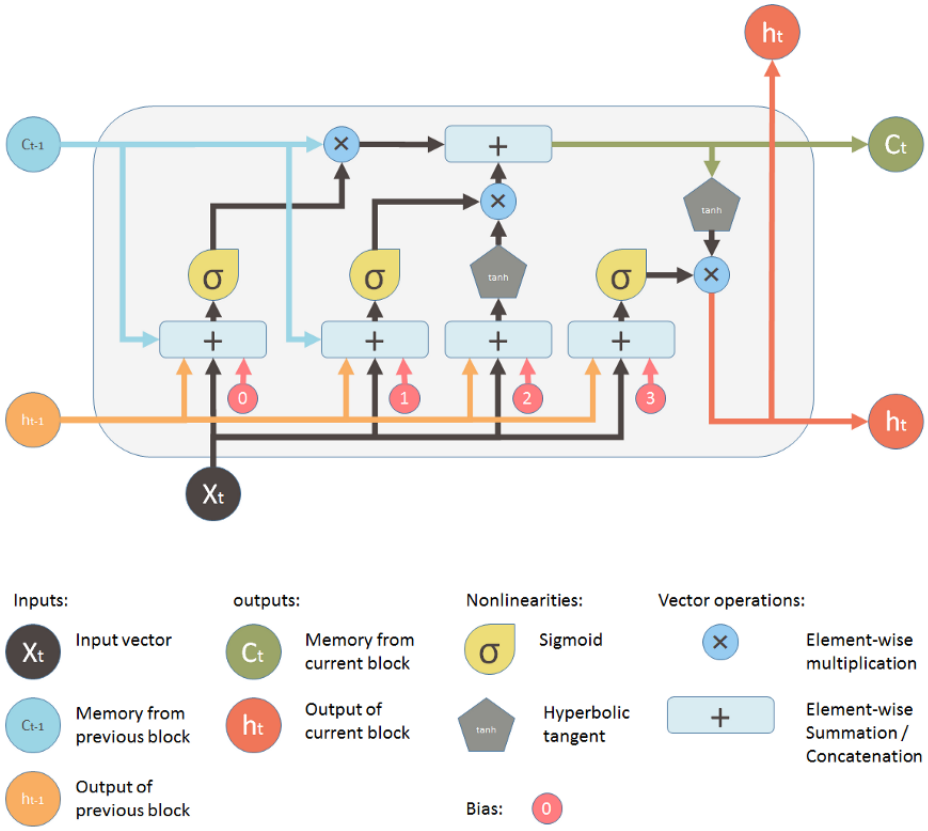


Figure 2.4: Long Short-Term Memory (LSTM) Network architecture. Image taken from Shi (2016, accessed August 3, 2020)

2.1.5.2 Recurrent Neural Networks

NLP applications of DL require sequence of varying size manipulation: sentences with different number of words, paragraphs with multiple sentences, or documents. For handling such ordered sequences, apart from padding the input sequences to a fixed maximum length, variations of the MLP models process at each step one token (word unit).

Recurrent Neural Networks are a special type of ANNs, where the output of each step is used as input for the next step. The backpropagation algorithm briefly described above is also used here for the learning process to update the weights of the model. As the depth of the network increases with the number of input signals that are recurring at each timestamp, the gradient that propagates the error decreases in magnitude and leads to low propagated values – a phenomenon also known as the vanishing gradient problem. To

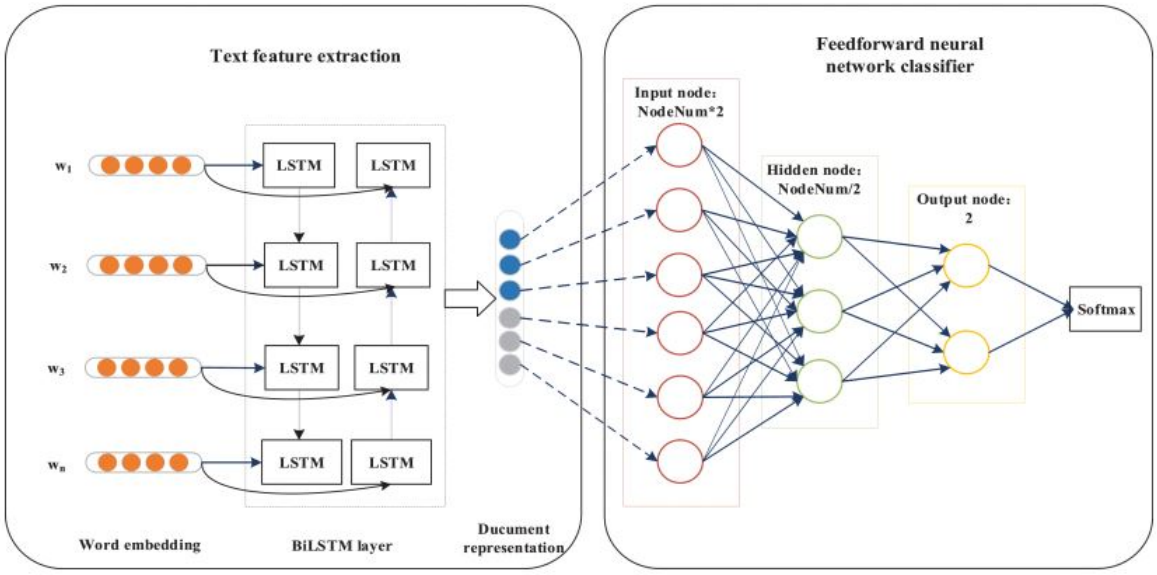


Figure 2.5: Bidirectional LSTM architecture for Sentiment Analysis (Xu et al., 2019)

overcome this problem, gated variations of RNNs have been proposed in order to control whether the previous recurrent state should be further used in the computation or not. The most popular gated RNN are GRUs (Rana, 2016) and LSTMs (Hochreiter and Schmidhuber, 1997). The LSTM architecture shown in Figure 2.4 contains, apart from the gated cell that ensures the short-term memory, another cell state that preserves the information between the time steps in the long run, by preserving an internal representation.

Bidirectional RNN The RNNs and their variations are networks that accumulate information as the input is being processed at each timestamp in only one direction: the input x_t is dependent on $x_0, x_1, ..., x_{t-1}$. The natural language also contains sentences or a larger context when the words refer to the context that follows. To accommodate this shortcoming of the RNNs of processing the sequences by using the information from the entire context (to the left and right of the current input token), the bidirectional recurrent networks have been introduced, where the information flows in both directions through the network.

In Sentiment Analysis and Text Classification applications, the bidirectional LSTM architecture has obtained state-of-the-art results. An example of a biLSTM network ar-

chitecture adapted for classification can be seen in Figure 2.5.

2.2 XAI

XAI is a rapidly growing field of AI that emerged once the highly complex DL algorithms became more and more present in different real-life applications. When the black-box algorithms started to have better performance than the simpler interpretable methods such as Decision Trees or Rule-Based Approaches, their adoption grew in many industries, nowadays being present in software applications for multiple domains such as health, law, automotive, retail, marketing, education, and many others, assisting domain experts and sometimes even fully automating the decision-making process.

While evaluation methods for a held-out dataset can create a certain level of user trust, the rapidly evolving environments of the real-world applications, the abundance, diversity, and unstructuredness of the data pose a challenge for the trained models when they are used on new unseen data, especially when there is a high chance that it does not fit the data distribution of the training or testing set.

When the decision made by an automated system can have a significant impact on people's lives, a justification for a decision or the ability to predict the behavior of an automated system becomes paramount. Furthermore, regulations for data protections have started to be put in place[1] in order to protect the users and to allow them to understand how their data is being used. This also requires a certain level of explainability for the algorithms that use the users' data for training purposes.

The research community has, thus, started to analyze methods of explainability and interpretability for the existing state-of-the-art DL algorithms, providing a formalization for this problem and aiming to identify approaches suitable for different purposes such as building user trust or helping ML experts to make an informed decision in terms of the chosen architecture, hyperparameters, or data used for training a model.

[1]Regulation of the European Parliament and of the Council of 27 April 2016 (General data protection regulation)

Although interpretability and explainability are terms that are often used interchangeably, we highlight the subtle difference in perspective, stated in the literature in the next two sections, followed by a review of the methods grouped in five categories.

2.2.1 Interpretability

Interpretability refers to the extent to which humans can derive insights from the behavior of an automated system and not from the architecture or employed method. This consists of the analysis of the output based on the input and perturbations of the input and drawing insights about how the model works. This approach of determining how a decision is being made, while it has the advantage of offering predictive power with respect to the model (the users are able to predict the behavior of the automated system for a particular test instance), it is subject to human biases for model interpretability.

2.2.2 Explainability

The interpretability level of a given algorithm relies on the capability of it being easily understood by the users. While for methods such as linear regression, one can explain the approach using simple linear mathematical concepts, the more complex algorithms become more difficult to explain using simple concepts. Presenting a motivation behind different architectural hyperparameters empirically determined becomes challenging and leads to decreased user trust. For this reason, explainability methods come in to provide justifications and explanations about how the model works. Explainability is, thus, the field that tries to explain internal mechanisms using human terms.

2.2.3 A review of XAI

The approaches that have previously been used to either interpret or explain a deep learning model in applications such as NLP or CV can be classified in the categories described below. While this is not an exclusive list and the categories described below overlap for

certain methods available in the literature, we try to capture the most common approaches, highlighting their advantages, disadvantages, and common use-cases and applications.

2.2.3.1 Proxy Models

These approaches employ highly interpretable model architectures in order to emulate the behavior of the model that is being explained. The performance and generalization power of the interpretable models is generally lower compared to the models explained that have high complexity and a large number of parameters. The performance of the proxy models does not match the performance of the model that they stand for, whereas locally interpretable models aim to solve that problem by focusing on subsets of data points, thus offering explanations that are locally accurate. A popular approach in the field of XAI using local surrogate models is LIME (Ribeiro et al., 2016), which explains a decision boundary through linear regression models that are locally faithful to the original model.

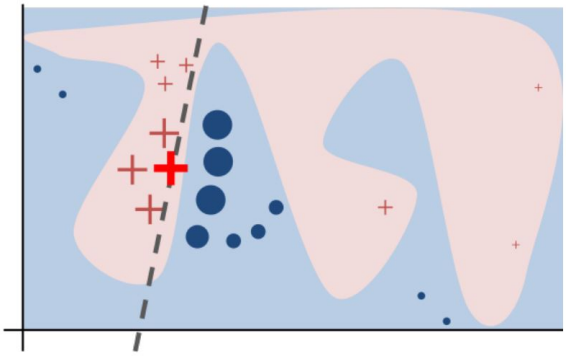


Figure 2.6: Toy-example for LIME. The image is taken from the original paper (Ribeiro et al., 2016)

2.2.3.2 Introspective Models

The neural introspective models make use of a second deep learning model to generate explanations. Whether the approach is supervised (Barratt, 2017) or unsupervised (our proposed approach), the goal is to generate natural language explanations based on the internal representation of the explained classifier. The general methodology consists of

using the latent state of the classifier as input for a natural language generation model that outputs natural language phrases. The usage of the internal state of classifier varies depending on the architecture of the model and it can use a concatenation of the neural layers of a MLP model or hidden states of a RNN.

Because the explained classifier has its parameters frozen, there is no loss of performance on the classification task, but the interpretability level might also be affected by the lack of explainability: a black-box is explained with another black-box.

2.2.3.3 Correlative Methods and Saliency Maps

This category of explanation methods focuses on visualization techniques for the relations between the input and the output without altering the model that is being explained. Perturbing the input and visualizing the impact different altered data points have on the prediction is key for inferring insights about the decision-making process of the classifier. The saliency maps were initially proposed as a visualization technique of Convolutional Neural Networks for Computer Vision applications. Aiming to gather insights about the most influential pixels for the predicted class, the saliency maps are image-specific visualizations obtained through a back-propagation pass of a given CNN model (Simonyan et al., 2013). They have also been adapted for NLP tasks (Li et al., 2016), where the tokens of the classified text are shown in a heatmap.

While this category's explainable methods are intuitive, they are susceptible to a number of pitfalls such as pointing to local minima or presenting the difficulty to translate the findings into reliable insights due to the fact that numerous inputs could lead to the same pattern.

2.2.3.4 Post-Hoc Explanations

The interpretable predictive methods provide explanations through their simple structure that is easily understandable by lay humans. These intrinsic explanations focus on explaining how the model works and the interpretability level is inversely proportional to its

complexity. While certain analyses can be performed on the data before training a model (pre-modelling explainability), or during training (architectural adjustments, jointly training for both prediction and explanation, or through explainable models), the post hoc explanations derive meaning after the model has been trained, providing explanations for particular classes of classifiers (**model-specific**) or independent of the model's architecture (**model-agnostic**).

Although the posthoc analysis can be performed by iterative perturbations, choosing an explanation is susceptible to human biases and it is cumbersome to find an interpretable explanation.

2.2.3.5 Example-Based Explanations

Some of the interpretable models such as k-Nearest Neighbors (kNN) or Decision Trees rely on this method of selecting the most similar already classified data point to make a new prediction. While this approach is self-explainable, the level of interpretability is dependent on the feature set of the data points and on the ability to represent them in humanly understandable ways.

The example-based explanations rely on explaining decisions based on the similarity of the classifier's output for a given data point to the prediction of a training instance or of the most representative instance of the predicted class. Variations of this approach comprise counterfactual, adversarial, and influential examples explanations that aim to sketch the decision boundary of the classifier. While this approach shows when the model fails to produce a correct prediction, allowing lay humans to determine the strengths and weaknesses of a classifier, the method becomes infeasible when there is a large number of possible explanations or when working with large datasets.

2.2.3.6 Conclusions

The progress of the field managed to advance in different objectives from presenting complex, mathematical explanations, to natural language, easy to understand and insightful

interpretations, optimizing the trade-offs between performance and interpretability. It is generally accepted that high-quality explanations give the user predictive power, allowing them to understand how the model makes predictions and building user trust. While this is still an ambitious goal to attain by a single model-agnostic method, the presented advantages and disadvantages of different categories of approaches presented above represent a guideline of which technique should be employed depending on the desired outcome.

2.3 Text Classification

Among the important NLP tasks present in many real-world applications such as Natural Language Understanding, Text Generation, Machine Translation, Summarization, Text simplification, Grammar correction, Question Answering, Speech to Text, and others, the task of Text Classification emerges as one of the tasks with multiple applications in automated systems. Whether it is document, email, article, sentence, review, post, or other types of text classification, this supervised learning problem occurs in many applications where the amount of data that needs to be processed becomes infeasible for a manual job. Furthermore, applications such as (multi-aspect) sentiment analysis, topic labeling, or language detection can also be defined as classification problems.

2.3.1 Sentiment Analysis

Classifying opinions expressed in natural language is an example of a text classification application. Sentiment analysis focuses on the analysis of the emotions (positive, neutral, and negative) of the textual data, bringing value to business services such as customer support and feedback, marketing, brand monitoring, market research, or social media analysis.

Sentiment Analysis has primarily been applied to short text in the form of posts, comments, or reviews that are frequently used on social media platforms and web applications for different services or product reviews such as traveling (airlines), movies, or paper reviews.

2.3.2 Text Preprocessing

Natural Language textual data requires a number of preprocessing steps that clean the text (by removing special characters, punctuation, stop words), split it into separate tokens, and map those to the corresponding root form (stemming) or to semantically complete stems (lemmatization). Determining the boundaries of each word goes beyond simple rules such as splitting by white spaces or punctuation due to the word contractions, abbreviations, or compound words. This process is called **tokenization** and it is a preprocessing step that plays a crucial role in encoding the text in the following stages of text processing. Apart from custom rule-based implementations, there are also a number of open-source tools and libraries such as SpaCy (Honnibal and Montani, 2017) or NLTK (Loper and Bird, 2002) that can perform tokenization.

2.3.3 Word Embeddings

The automatic processing of natural language requires transforming the words into numeric features that can preserve the underlying semantic meaning. The initial approaches of vectorization of the text under classification relied on word frequencies, Term Frequency–Inverse Document Frequency (TF-IDF) scores, or bag of word approaches such as one-hot encoding. These statistical methods have a number of limitations that include the inability to capture the sequential nature of the language and the contextual meaning of the words, especially of the polysemous words. Moreover, the high dimensionality of the one-hot embeddings becomes infeasible for a real-world vocabulary, as the vectors have the same number of elements as the total number of words that are considered.

Later on, neural embedding algorithms were trained on large corpora, attempting to generate lower dimension vectors that capture more meaningful information. Models such as Continuous Bag of Words (CBOW) and Skip-Gram were trained to predict the target word from the context words and, respectively, the context of a given word.

Embedding vectors obtained by the Word2Vec algorithm (Mikolov et al., 2013) preserve the synonym relations between words by exploiting the word co-occurrences. Another pop-

ular pretrained word embeddings that capture semantic and syntactic meaning are Global Vectors for Word Representation (GloVe) (Pennington et al., 2014). Using pretrained embeddings decreases the required computational power of the training process, having the advantage that the vectors already capture information about the co-occurrences of the words from the large corpora that they have been trained on. On the other hand, training randomly initialized vectors directly on the desired NLP task is another popular choice that appears in the literature, as it tackles the out of vocabulary words problem, resulting in vectors that capture the context information of the training data.

2.3.4 Evaluation Metrics

Text classification is a supervised learning problem for which a dataset of labeled instances exists. The dataset is usually split into a training subset of data points and a testing set that is being used only for evaluation, to assess the generalization power of the classification algorithm on unseen data. When performing data acquisition, the data points should be equally distributed according to the set of classes that they will be divided into, such that the resulting dataset is balanced (containing a similar number of instances for each predicted class). For an accurate evaluation, when the labeled data set is being split into training and test set, the test set should, ideally, preserve the same data distribution as the training set and, more importantly, as the real distribution of the data, so that the algorithm will perform well on new real-world datasets. When the available datasets are not balanced, approaches such as undersampling and oversampling are used in order to balance it. Another challenge related to the data is that the distribution of the labeled data points available for training can differ from the distribution of the real-world data.

Metrics. To be able to compare the performance and measure the quality of different models and classification algorithms, for text classification, the metrics that are commonly reported in the literature are accuracy, precision, recall, and F1-measure. For the binary classification, we refer to the positive correctly classified instances as *True Positives* and to the incorrectly classified positive samples as *False Positives*, and as *True Negatives* and

False Negatives for the negative instances, respectively. These values are usually reported in a 2x2 matrix that is commonly referred to as the **Confusion Matrix** (Table 2.1).

		True class	
		Positive	Negative
Predicted Class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Table 2.1: Confusion Matrix

Classification Accuracy represents the ratio between the correctly classified instances and the total number of input samples (Equation 2.3). Accuracy is a classification evaluation metric suitable for balanced datasets. Its values are usually reported as a percentage.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total number of predictions}} \tag{2.3}$$

Precision (Positive Predicted Values) measures the proportion of predicted positives out of the total true positive instances (Equation 2.4). This measure is suitable especially for use cases when a high confidence in the positive prediction is required and the cost of fall positives is high. An example of applications that require high precision are medical diagnoses.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2.4}$$

Recall reports the proportion of actual positive instances correctly identified (Equation 2.5). This evaluation metric is particularly useful for applications that aim to detect fraud or unusual behavior of certain systems, where the cost of false negatives is high.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.5}$$

When there is a trade-off between minimizing the false positives and false negatives, choosing the appropriate evaluation metric between precision and recall is recommended.

F1-score. To assess if a model manages to obtain both a high precision and recall value, the F1-score (or F1-measure) is chosen since it is a combination of both precision and recall in the form of their harmonic mean (Equation 2.6).

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{2.6}$$

For binary classification, the precision and recall are usually reported for the positive class, and the F1-score is thus also reported for the positive class. Variations of the F1 score include the micro F1 and macro F1 score, along with the weighted (Average) F1 score. The **weighted F1 score** computes the average between the metrics for each class weighted by support (the number of true instances for each label).

2.4 Text Mining

The ever-increasing amount of information available in the form of natural language, especially on the social media platforms, makes it infeasible for humans to process it. For this reason, automatic methods of extracting and organizing the information become indispensable. **Text mining** approaches analyze such large collections of documents through different NLP tools and methods, transforming the raw textual data into valuable insights about it. One of the most important tasks in text mining is keyword extraction.

2.4.1 Keyword and Key phrase Generation

The NLP task of keyword extraction is important for fields such as information retrieval, text summarization, text classification, topic detection, or as a measure of similarity for text clustering. The keywords refer to single words, while a key phrase is a multi-word lexeme that is "clumped as a unit" (Manning and Schütze, 1999). They have the capability of summarizing a context, allowing one to organize and find documents by content. They are both be referred to as *key terms*.

Keyword generation is being performed by keyword assignment and keyword generation (Siddiqi and Sharan, 2015). While the first one refers to building a predefined vocabulary of key terms, an approach that has the advantage of simplicity and consistency, the second one brings more flexibility, selecting the candidate terms from the given corpus. The later determines the important keywords and keyphrases present in the document, independent of a previously built vocabulary. Although there are advantages of using keyword extraction, it is a difficult task that lacks in consistency, presenting difficulties in accurately detecting and ranking the most relevant keyphrases in a given semantic context. We are going to further discuss the extractive methods and the popular solutions from the community of NLP researchers.

There are a number of approaches for keyword extraction task such as statistical methods, linguistic ones, machine learning, each being analyzed based on the width of their applicability or depth of the semantic understanding capabilities (Shamsfard and Abdollahzadeh Barforoush, 2003).

Depending on whether they are using an annotated corpus for training or not, they can also be classified as supervised or unsupervised approaches. Usually, annotated datasets for this task is expensive, requiring domain experts for selecting the key terms from a document. To overcome this, researchers automatically created datasets using the published scientific articles, along with the associated keywords (Caragea et al., 2014).

Generally, keyword extraction relies on three main stages: candidate selection, feature extraction (scoring), and keyword selection (Kaur and Gupta, 2010).

In the following subsections, we will highlight the most relevant publications, briefly explaining the proposed methods for this task.

2.4.2 Linguistic Approaches

The English language presents a number of patters in terms of phrases, especially when it comes to technical terms (Justeson and Katz, 1995). Exploiting the grammatical structures and identifying common patterns allows one to reduce the search space when identifying

the candidate phrases. Common Part Of Speech (POS) constituents for phrases include nouns, prepositions, and adjectives. These approaches require extensive text analysis (both syntactical and lexical analysis), domain knowledge, and language-specific insights.

Another path of exploiting the semantics of the words, especially for domain-specific documents is through symbolic learning. While this improves the depth of the semantic understanding of a given text, the tools used for semantically rich representations of words restrict the width of their coverage.

2.4.3 Statistical Methods

If the linguistic approaches have a deep semantic coverage, the statistical methods have the advantage of being language-independent, not requiring domain specific knowledge, and being applicable on a wide variety of text styles.

The numerical features that provide relevant insights for the task of keyword extraction are based on term frequency, position (Herrera and Pury, 2008), and co-occurrences. Many approaches have been proposed during the years, exploiting different claims that discriminate the key terms based on the aforementioned numerical features.

Key terms are words or phrases that capture the semantics of a context, summarizing, and representing a content. Thus, one can derive properties based on location in the document or frequency that reduce the search space for keywords. Since 1975, a discrimination value analysis ranked the words according to how well they separate the documents based on their content (Salton et al., 1975). Later on, theories stating that key-words tend to cluster together, while the other words are scattered along a document have been proposed (Ortuño et al., 2007; Carpena et al., 2009). The co-occurrence of words and their position in the document has also been used together with linguistic features (POS) in previous work (Hu and Wu, 2006), in an attempt to also use grammatical insights.

Based on numerical features, a number of algorithms are well-known and widely used for keyword extraction. In the following subsections, we will briefly explain how they work.

2.4.3.1 TF-IDF

One of the most common approach used in the literature is computing the numerical statistic Term Frequency–Inverse Document Frequency. This method overcomes the problem faced by the simple frequency based approach, where very frequent words are often prepositions, pronouns, or conjunctions. Such words often do not carry too much semantic meaning when taken out of context. The inverse document frequency ensures that the words with higher scores have a high frequency in a given document, but a low number of appearances in the rest of the corpus. This way, a word with a high score carries a lot of semantic meaning that distinguishes a document from the corpus it is taken out of. This characteristic is also shared by the keywords, which makes the TF-IDF method a good baseline for the key-word extraction task.

2.4.3.2 TextRank

TextRank is an unsupervised graph-based method used for both keyword extraction and text summarization. Proposed by Mihalcea and Tarau (2004), this method builds a graph where each node represents a lexical unit (one word), as a keyword candidate. The edges between the nodes are added when the words are situated in a window of a maximum N words in the original document, where N is a hyperparameter. Additional information such as lexical information (part of speech - an edge between the vertices having the same part of speech), semantic similarity, and weights for each of these features can be encapsulated in the graph. Depending on the number of edges and on the type of the graph (directed or undirected), a formula is applied to compute the score for each node. The scores for all the nodes are then sorted in decreasing order and the top k keywords are extracted. The hyperparameter k is established based on the size of the original text. For example, for the dataset used in the original paper that uses short text (abstracts of scientific papers), a third of the number of the total words is chosen as the number of extracted keywords. In a post-processing phase, the selected keywords that are adjacent in the original document are grouped to form multi-word expressions.

The main drawback of this method (in its original form) is that it does not take into

account the semantic similarities between texts. This can be addressed by adding new edges weighted by similarity scores such as cosine similarity. The work has also been extended with enhanced variations that use Word2Vec (Zuo et al., 2017) and Doc2Vec (LI et al., 2019).

2.4.3.3 RAKE

Rapid Automatic Keyword Extraction (RAKE) is a statistical domain and language-independent approach that uses both the frequencies and the co-occurrences of the words (Rose et al., 2010) . After tokenization and stop-word removal, the candidate keywords and keyphrases are extracted (using as delimiters the stop-words) and the matrix of co-occurrences is built out of each candidate word (token). Next, the authors assign a score to each word that is computed as the ratio between its frequency and the degree of the word in the matrix of co-occurrences (the sum of the number of co-occurrences the word has with any other candidate word). The score for the candidate keyphrases is obtained from the sum of the scores of each word of the keyphrase. Furthermore, if two key terms appear in the original text in the same order twice or multiple times, they will be merged in a new keyphrase. The ranking of the scores in decreasing order is followed by thresholding based on the number of keywords to be extracted.

This is a statistical method that allows one to extract multi-word expressions in a fast and straightforward approach. Its limitation is that the more words are in an extracted key term, the less the cohesion of the phrase will be. Furthermore, when applied on short text, the key terms end up having scores that are very close to each other, creating a challenge for the ranking and selection stages.

2.4.3.4 Yake!

A more recent unsupervised statistical approach for single-documents key term extraction called *Yake!* was proposed in 2018. This approach captures the context information using 5 local statistical features for each term individually:

1. Casing (W_{Case}) - refers to the casing of the word

2. Word Positional ($W_{Position}$) - refers to the position of the word, the early occurrences weighting more

3. Word Frequency (W_{Freq})

4. Word Relatedness to the Context (W_{Rel}) - computes the number of different terms that occur too the left and right side of the candidate word

5. Word DifSentence ($W_{DifSentence}$) - represents the number of times the term appears in different sentences.

$$S(w) = \frac{W_{Rel} \cdot W_{Position}}{W_{Case} + \frac{W_{Freq}}{W_{Rel}} + \frac{W_{DifSentence}}{W_{Rel}}} \tag{2.7}$$

The individual word score function from equation 2.7 generates low scores for important keywords. For keyphrases (kp) with multiple words, the sum of each frequency (Tf) and the individual scores are used in the equation 2.8, using a sliding window of 3-grams.

$$S(kp) = \frac{\prod_{k \in kp} S(w)}{Tf(kp) \cdot (1 + \sum_{w \in kp} S(w))} \tag{2.8}$$

2.4.4 Rule-Based Approaches

Rule-based approaches aim to mine rules using different linguistic and syntactic features such as the part of speech or the sentence's structure. The disadvantage of these methods is that they require domain knowledge and language insights, also being computationally expensive.

While the linguistic approaches do not have width and the statistical methods do not comprise semantic information, Hu and Wu proposed a compromise solution, presenting a POS hierarchy that overcomes both aforementioned shortcomings (Khoury et al., 2008). The lexical items are categorized according to the 46 parts-of-speech of the Penn Treebank (Santorini, 1990) and grouped together on multiple layers of generalization as we go up in the hierarchy. The rules are generated using this hierarchy to cover the training set of American English texts extracted from the Brown corpus.

Previous work on such methods also makes use of the frequencies of words to compute the support and the confidence of the generated rules for ranking and filtering the generated keywords (Jayakumar, 2012).

For the sentiment analysis task, rules containing the polarity words of the candidate phrases can be used to filter out neutral terms. This can be accomplished using rule-based or lexicon-based sentiment analysis tools such as Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert, 2015).

Chapter 3

Related Work

This chapter reviews the state-of-the-art approaches for XAI, providing details about the methods that have been employed so far to explain or interpret deep learning classifiers and their predictions. We also mention the challenges faced by the researchers in the field of Explainability and Interpretability for AI for data acquisition, formally defining the task, and determining the appropriate evaluation methods.

3.1 Explainability and Interpretability

In the literature, the terms explainability and interpretability are often used interchangeably. Although they both refer to the capability of lay-humans to express in easy-to-understand terms how black-box models make predictions, for purposes such as building user trust or having the ability to predict and adjust the behavior of the algorithms in real-time, there is a subtle difference between those concepts. Explainability refers to the user's ability to understand the inner workings of an algorithm, allowing them to answer the question "How does it work?". While the explainability research area is preoccupied with deciphering why a certain prediction has been made based on the architecture and employed methodology, the interpretability term refers to the ability to consistently predict the behavior of an intelligent system without making assumptions or without knowing implementation and architectural details (Kim et al., 2016).

The interpretability of Machine Learning models is traded off for the higher performance, more complex classifiers. The class of interpretable models (Decision Trees, Rule-based, Linear Regression, K-Nearest Neighbors, Naive Bayes – because of the independence assumption of the features) often underperform compared to more complex models from the deep learning category such as LSTM (Hochreiter and Schmidhuber, 1997), RNN, CNN, Transformer (Vaswani et al., 2017; Devlin et al., 2019). Aspiring for achieving both state-of-the-art results on the given task and interpretable decisions, the efforts of the research community started to focus on explainability for such complex algorithms. Whether the proposed methods focus on interpretability for deep neural networks (such as the Gradient-Based or Network Dissection methods) or if they can be applied to any classifier (model-agnostic), there is great applicability for tasks handling different data types: textual, visual, and numeric.

As highlighted in the book "Interpretable machine learning. A Guide for Making Black Box Models Explainable" by Molnar (2019), the advantage of easily automating "interpretability when it is decoupled from the underlying machine learning model" favors the future of model-agnostic interpretability research. In the following subsections, we will highlight the progress of research for both model-dependent tools and of those designed for black-box models, but we will mainly focus on model-agnostic methods for NLP tasks.

3.2 Recipient - Who is the explanation intended for?

The explainability methods offer different degrees of interpretability depending on the recipient. Moreover, their focus is on different aspects: model understanding, instance's prediction explainability, feature contributions, representation understanding, providing representative examples. While lay-humans require easy-to-understand explanations for particular decisions, AI researchers need a deeper understanding of the model and complex data insights to be able to prevent or improve a model that is malfunctioning. For example, in image processing with medical applications, highlighting the regions of the picture is informative enough for both the medical personnel and the patients, while for

the ML experts very detailed information (pixel-level) would be more informative (Samek and Müller, 2019). For natural language processing, self-explanatory, coherent phrases in natural language are the most insightful explanations for people without any background in algorithm design or computer science. The focus of the AI researchers community has been on understanding the model behavior, offering them the tools needed to fix and improve their systems. In these scenarios, the explanations become more complex and not necessarily easy to understand, containing information about the model architecture (van Aken et al., 2019; Guan et al., 2019), the used representations (Koç et al., 2018; Jang et al., 2018), statistical metrics about the training set (corpus analysis), feature importance or performance metrics when different perturbations take place (Wiegreffe and Pinter, 2019).

3.3 Evaluation Methods

The evaluation approaches for explainability are diverse and depend on the level of interpretability that is aimed for, the intended user, and the task in question. There are a number of approaches that involve lay-humans, experts, no humans, or a hybrid of those methods. We briefly review, in the following paragraphs, the evaluation proposed in the literature.

3.3.1 Human Evaluation

This evaluation method is an expensive process since it involves manual interpretation of the results, requiring multiple annotators that have to reach a certain level of agreement on the decisions they make. This approach raises a number of challenges such as defining the task (comparing classifiers, scoring the generated explanation, ranking explanations), finding and training the subjects (especially when domain experts are required), or assessing the agreement between the annotators. Doshi-Velez and Kim (2017) describe two evaluation levels depending on whether the task requires domain experts for evaluation or lay-humans.

Firstly, the class of **Application-grounded evaluation** methods refers to humans working with the real task. Multiple applications that require domain experts evaluating the explanations of an AI system refer to the medical domain.

Secondly, the **Human-grounded evaluation** methods are simplified proxy-tasks that are designed such that lay-humans can evaluate the quality of the explanations through a subset of (representative) explained instances. Depending on the goal of the generated explanation, different human evaluation tasks have been proposed.

Explanations that justify the prediction. For this evaluation, the humans are asked to provide their own prediction based on only the explanations of an instance that the original classifier. A variation of this method is to only select the instances for which the classifier has a high confidence prediction. The humans can also be asked to mention how confident they are in the decision that they made (Lertvittayakumjorn and Toni, 2019).

Investigate uncertain predictions. This task is used to analyze the uncertain predictions. The humans are provided with the input instance, the prediction, and a set of evidence and counter-evidence texts. They are asked to select the most likely class of the input text and specify if they are confident or not (Lertvittayakumjorn and Toni, 2019). The confidence score assesses the usefulness of the explanations.

Counterfactual simulation. For this category, humans are presented with an instance, its explanation and its predicted label and are asked what needs to be changed in order for the model to make the correct prediction (Doshi-Velez and Kim, 2017; Ribeiro et al., 2016).

Assessing the agreement between annotators. When human evaluation with multiple annotators is employed, the agreement between any two of them should also be assessed using statistical methods. For categorical values, the simplest metric that can be used is the percentage of cases when the annotators produced the same output. Another statistical approach is Cohen's kappa coefficient (McHugh, 2012), which is a more robust measure

that takes into account the possibility of the agreement occurring by chance. According to the value of κ, the agreement can be classified as "no agreement" ($\kappa < 0$), "slight" ($\kappa \in [0.01, 0.20]$), "fair" ($\kappa \in [0.21, 0.40]$), "moderate" ($\kappa \in [0.41, 0.60]$), "substantial" ($\kappa \in [0.61, 0.80]$), or "almost perfect agreement" ($\kappa \in [0.81, 1.00]$).

3.3.2 Automatic Evaluation

Functionally grounded. According to the classification proposed by Doshi-Velez and Kim (2017), the functionally grounded methods require a formal definition of interpretability, allowing one to design experiments that assess the quality of the explanations. To accomplish this, quantitative metrics that show the performance improvements of a given model are employed, along with formalizations of assessing the sparsity or the complexity of the explanations. For example, the authors of the Local Interpretable Model-Agnostic Explanations (LIME) algorithm propose a formula (3.1) that ensures both the **local fidelity** ($\mathcal{L}$) and the **interpretability** (Ω) of the proxy model (g). Minimizing the complexity measure of the explanations is used to ensure that the explanations will be comprehensible (Ribeiro et al., 2016).

$$\xi(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

Other desiderata such as **unambiguity** and **interactivity** have been addressed by the researchers (Lakkaraju et al., 2017), quantified using the coverage and overlapping of the explanations. Nevertheless, additional architectures added to a base model that aims to provide natural language explanations have the main goal of improving the performance on the main task (classification, natural language inference) or at least minimizing the performance loss. Camburu et al. (2018) show that both architecture-incorporated or post-hoc explanation generation is a more challenging task than the main task of the model that is being explained. Moreover, the proposed baseline on the Natural Language Inference (NLI) task that also provides explanations leads to a small drop in performance (1-3%).

Supervised learning. The methods that benefit from having ground truth explanations try to generate new explanations as similar to the ground truth as possible. Such approaches are evaluated using Bilingual Evaluation Understudy (BLEU) scores (Liu et al., 2019a) and the impact that adding them to the classified instance has on the final prediction on the classification task.

Defining baselines Other common evaluation metrics that allow one to draw insights from automatically generated data consist of defining baseline approaches. Similarly to the classification's task baseline of predicting the majority class, for the sentiment analysis task, one can make use of the sentiment analyzer to determine the polarity of the explanations and assess the correlation between the generated explanation's polarity, the predicted label, and the ground truth.

Explanation-label consistency Chen and Ji (2019a) define the coherence score to measure if the explanation's polarity is consistent with the prediction and report it for different methods: the proposed approach a generation model trained using Data Augmentation with External Knowledge, LIME, and the considering as explanations the words of the instance that have the highest cosine similarity of the corresponding embedding with the instance under classification.

Post-hoc accuracy Extractive methods of the subset of features that are highly correlated with the predicted label can be evaluated by computing the accuracy rate when only the selected features are being used for prediction. An example of such an evaluation method has been employed by Chen et al. (2018).

3.3.3 Hybrid Methods

Extensive experiments have been performed to assess the quality of the explanations using both human input and automatic measures such as the performance of a classifier on a

different dataset. These approaches are suitable for extractive methods. An example of such an approach is described in the work of Ribeiro et al. (2016).

Choosing the better classifier. The authors of the paper generate explanations for the models trained on the 20 newsgroup dataset and show to the human evaluator instances along with the predictions of two predictors, and the corresponding explanations (the importance given to the most relevant words from the instance under classification). Based on this, the lay-humans are asked to decide which algorithm will perform best in the real world. The decision of the humans is compared to the performance of the algorithms on a new dataset (on the same classification task). The authors also describe an iterative process of improving the model by cleaning the dataset and retraining the models based on the feedback from the human evaluators and analyze the performance changes. A similar approach (without the iterative process) has also been described by Lertvittayakumjorn and Toni (2019).

3.4 Datasets for the Explanation Generation Task

We are reviewing in this section the most popular datasets that have been used in NLP classification tasks for explanation generation. In the supervised category, there are datasets that, apart from the instances under classification, also contain ground truth explanation. For the unsupervised methods, the research community focused on defining criteria based on the classification task. This includes generating explanations that improve the prediction confidence of the classifier, finding the minimal subsets of the input features that preserve the prediction, or optimizing the maximal feature set that can be eliminated before the prediction changes.

3.4.1 Unsupervised Learning

20 newsgroup dataset. Previous work focused on a subset of this dataset[1] on the task of classifying Christianity versus Atheism texts. The research of Ribeiro et al. (2016) brought into light that the models that perform very well on the held-out set, perform very poorly on the same task on new test instances. This is due to the fact that the models learn some patterns specific to this dataset, the decision being highly influenced by words that have no connection with the content of the dataset ("Posting", "Host", "Re"). The explanations proved to be informative after extensive human evaluation on different tasks such as identifying the better classifier out of two, identifying the tokens that should be eliminated from the subsequent training (as a cleaning preprocessing step), or deciding whether a subset of explained instances are able to reveal the model's behavior.

IMDB. The local explanations extracted using LIME (Ribeiro et al., 2016) has also been evaluated for sentiment analysis classifiers on the movie review Internet Movie Database (IMDB) dataset (Maas et al., 2011) in previous work. Chen and Ji (2019a) assesses the impact of training data augmentation on the extracted explanations through two methods. Firstly, an external knowledge approach is employed for removing keywords based on a predefined list, such that the augmented instance is similar to the original text, but has no sentiment polarity. Secondly, the adversarial approach that substitutes a minimal number of words with their corresponding synonyms such that the semantic similarity is maintained. These two data augmentation approaches lead to improved local explanations for the text classifiers.

3.4.2 Supervised Learning

We note that for the supervised task of explanation generation, there are a few datasets that are out of the scope of this research work since they refer to specific tasks that are challenging to adapt to such as visual question answering (Park et al., 2018), text

[1] http://qwone.com/ jason/20Newsgroups/

justification for math problem solutions (Ling et al., 2017), or science question answering (Jansen et al., 2018).

e-SNLI. On the task of NLI, the Stanford Natural Language Inference (SNLI) corpus Bowman et al. (2015) has been annotated with crowd-sourced **free-form** explanations, leading to the augmented dataset called e-SNLI (Camburu et al., 2018) that provides three explanations for each instance in the dataset. The inter-annotator agreement is measured using BLEU scores in an attempt to determine if this measure would be suitable for evaluating the explanation generation models. Since the inter-annotator BLEU score of the thrid explanation with respect to the first two is 22.51, the authors conclude that human evaluation is needed for this dataset.

PCMag Review Dataset. This dataset[2] contains a review text, an overall score ({1.0, 1.5, 2.0,..., 5.0}), and three short comments: positive, negative and neutral. Previous approaches (Liu et al., 2019a) use these short comments in a generative explanation framework to provide fine-grained explanations for product reviews, while also improving the performance of a base classifier on the categorical classification task.

Skytrax User Reviews Dataset. This dataset[3] contains numerical scores for five attributes of the flights as fine-grained information. These scores are used in the Generative Explanation Framework (Liu et al., 2019a) to provide predictions for the test review text. This is an example of numerical explanations for a set of predefined features (possible explanations) that the training set is annotated with.

3.5 Extractive Explainability Methods

For the extractive approaches of explaining deep learning text classifiers, the explanations consist of keywords or keyphrases extracted from the instance that is being classified.

[2]https://github.com/LayneIns/Generative-Explanation-Framework-for-Text-Classification

[3]https://github.com/quankiquanki/skytrax-reviews-dataset

Whether extracting natural language explanations is being performed in a supervised or unsupervised manner, they rely on the criterion of determining the words or phrases that suffice for a correct prediction (Lei et al., 2016), determine the most correlated features by maximizing the mutual information (Chen et al., 2018), or analyze the propagation activation differences. Previous work has referred to this strategy by jointly training models with multiple goals (multi-task learning) or by optimizing the process of finding the minimal set of input features, without altering the classifier. In the following subsections, we will present an overview of the extractive methods.

3.5.1 Feature Selection

DeepLIFT (Shrikumar et al., 2017) is a model interpretability approach that evaluates the contribution of an input feature compared to a baseline data point. The contribution is being computed in the backpropagation stage, using the chain rule, where the gradient multiplier is replaced by the slope of the input with respect to the reference data point.

Chen et al. (2018) propose a CNN architecture for an extractive explainer (L2X) that uses the theoretical background of information theory to approximate the mutual information between the selected subset of features and the model's predicted label. The explainer is trained on the classification task using as ground truth the labels predicted by the model that is being explained, with the goal of maximizing the mutual information between the predicted label (of the original input instance) and the selected subset of k features. For the output distribution of the feature weights, the authors use Gumbel-Softmax continuous reparametrization. To assess the performance of the explainer, the extracted features are used for the classification task of the original model.

3.5.2 Multi-task Learning

The ensembles of models that are jointly trained for both classification and explanation generationAntognini et al. (2019), along with the supervised models for which leverage datasets with granular possible justifications (such as multi-aspect sentiment analysis), fall

under this category of explainable approaches. While this is a user-oriented approach to generate justifications, the interpretability problem of the model is overlooked since the task becomes teaching the model what justification each classified data point should have, rather than interpreting what a model learns from labelled data points on the classification task only.

3.5.3 Surrogate Models

Surrogate models are interpretable model-agnostic methods that mimic the behaviour of a non-interpretable, complex model. Depending on how faithful the interpretable models are, compared to the model that is being explained, we can divide them into local and global surrogates.

Local surrogate These approaches consist of locally faithful interpretable models. After defining the neighborhood of a given instance, the proxy-model is trained on a smaller set of data points such that it produces predictions similar to the classifier that is being explained.

An example of such model is LIME (Ribeiro et al., 2016). This algorithm provides explanations for a given classified data point, treating the explained classifier as a black-box and approximating its behavior using an interpretable surrogate such as linear regression (used in the original publication). Firstly, the neighborhood is being defined. For NLP applications with interpretable binary vectors representations (denoting the presence or absence of the words), the neighborhood of a given instance under classification is defined by (randomly) flipping the bits of some of the input words. The interpretable surrogate classifier is trained on this small subset of perturbed instances, using the output of the original classifier as ground truth. Since this computation happens for each instance that is being explained, this method's speed depends on the training time of the chosen surrogate model. One can understand the local behaviour of the complex model that is being explained by interpreting the simpler surrogate model.

Global surrogate models Unlike the local proxy-models that use small subsets of data points, these interpretable models are trained on the entire dataset. The main disadvantage of such approaches is that the models may be very close to the explained classifier for subsets of the dataset, but widely divergent for other subsets. Thus, the reliability of the explanations is difficult to assess even though on average, the metrics (for example, R-squared measure) show that the surrogate model is able to accurately approximate the predictions of the black box.

Examples of global surrogate models have been proposed by Liu et al. (2018); Craven and Shavlik (1995); Kuttichira et al. (2019) using CNNs (Knowledge Distillation) and Decision Trees.

3.5.4 Perturbation-Based Approaches

One of the most popular methods of identifying the most important feature considered by a black-box model is analyzing the impact of the classifier's confidence on the prediction when perturbing the input features.

Shapley Values. One example of such a method is inspired by the game theory. The Shapley value is the average marginal contribution of a feature value across all possible combinations of features. In other words, the Shapley value of a feature is the average change in the prediction that the instance under classification receives when the feature is added to the input set of features.

Attention-based models. Another research direction for deep learning explainability examines the correlation between the parameters used by a model and their influence on the prediction. Since the attention mechanism (Bahdanau et al., 2014) emerged as an improvement over the encoder-decoder architectures, applied especially in NLP applications such as machine translation (Vaswani et al., 2017) or text classification (Sinha et al., 2018; Yang et al., 2016), as a way to alleviate the long-range dependency problem of recurrent sequence-to-sequence architectures. The work on this approach was extended to

analyzing the correlation of the attention weights with gradient-based feature importance measures (Jain and Wallace, 2019), as an attempt to explain the decisions of a deep neural network. Various comparisons of the original architecture have been performed with perturbed weights, random weights, or adversarial weights (Wiegreffe and Pinter, 2019), analyzing the impact on the predicted class Serrano and Smith (2019). While this mechanism improves the performance across different NLP tasks, some experiments showcased the high attention weights having a low impact on the predicted label and on the adversarial distributions that preserve the model's output. This led to the conclusion that attention weights do not necessarily correspond to the learned features' importance.

3.6 Generative Explainability Methods

The class of algorithms that provide explanations that are not a part of the classified instance focus on the training data to justify a certain decision. This category includes methods such as example-based explanations (nearest neighbor, prototypes), a subset of representative instances - Submodular Pick - Local Interpretable Model-Agnostic Explanations (SP-LIME) (Ribeiro et al., 2016), or the semantic closest phrase out of a dictionary of possible explanations (Liu et al., 2019a). The generative methods are more suitable for applications for which the feature set is humanly understandable and carry satisfactory meaning that can be interpreted out of a given context. Thus, these methods is most suitable for natural language and image processing tasks.

Chapter 6 of Molnar (2019)'s book covers counterfactual explanations, adversarial explanations, prototypes, and influential instances, emphasizing some of the advantages and disadvantages that we are summarizing in the following paragraphs.

k-nearest neighbors. One of the most popular classification algorithms that is also used as a stand-alone algorithm for classification, kNN is also used as an explanation method for other models, with the goal of identifying the training instances that produces the closest output to the explained data point, when another classifier is being used.

Prototypes. A representative point for a collection of data is called prototype. Data visualization and exploration methods use prototypes to understand the data distribution through summarized descriptions. This approach is also used as a model-agnostic explanation method of justifying the decisions through instances extracted from the dataset. The challenge of this approach is to determine the optimal number of prototypes that describe the data such that they are representative of the data distribution that they stand for.

Counterfactual explanations. Counterfactual explanations generation is a model-agnostic approach that focuses on identifying the causality of a prediction. In other words, it tries to identify the changes of some of the input features that lead to a change in prediction (to a predefined output). While the prototypes generate justifications from the training set, the counterfactual explanations are in the form of new values of the features that are not necessarily from the training set. While this objective of determining the input values that lead to a particular output is straightforward to understand, there are usually numerous counterfactual explanations that can be taken into consideration. For NLP text classification, this approach has been previously used to explain why documents have or not been classified as with a particular label.

Adversarial explanations. Recent publications propose systems that determine whether an explanation generative model provides inconsistent explanations. Camburu et al. (2020) designed a framework that determines the adversarial explanations that can be generated from a subset of features of the instance under classification. This work highlights the weaknesses of previous state-of-the-art explanation generation models (Camburu et al., 2018).

Influential Instances. Determining the training set's influential instances is another approach of attempting to derive insights about the model's learned parameters and change of predictions when such points are removed from the training set. While this is a model-agnostic method of debugging machine learning algorithms that use differentiable parameters (such as neural networks), it is a computationally expensive approach since comparing

the models requires different training stages depending on the dataset.

Other frameworks. Other generative frameworks make use of gold explanations.

Liu et al. (2019a) proposes a Generative Explanation Framework (GEF) based on Conditional Variational Autoencoder (CVAE) that generates fine-grained explanations using the short comments associated with each product review that is being classified. The authors propose a loss function that comprises both a classification loss and an explanation generation loss. The role of the explanation loss is to minimize the difference between the generated explanation and both the key terms provided by humans (from the dataset) and the input text. For this, another classifier is being trained on the classification task using the gold explanations. An Explanation Factor (EF) computes the discrepancy between the classifier's output for gold explanations and generated explanations and, respectively, between the generated explanation and the class distribution predicted by the predictor that is being explained. At testing time, the gold explanations are removed from the generative process and evaluated using BLEU-1 scores (40.1, 35.9, 33.2) and the classification accuracy (44.04%). Thus, this framework generates fine-grained explanations for each class that are similar to the explanations provided by the humans as short comments of their main review text.

Chapter 4

Methodology

This chapter describes the proposed framework providing a high-level overview of each component of the framework and describing the approaches for each process involved. We also discuss the dataset, and the evaluation metrics and methods employed to assess the quality of the generated explanations.

4.1 General Architecture

The proposed framework architecture (Figure 4.1) consists of two deep learning models: the text classifier to be explained, C, and the explanation generator model, G, that generates explanations for the instances under classification, and an unsupervised keyphrase extraction algorithm for the dictionary of explanation acquisition.

After preprocessing the training set of text instances, keyphrase candidates are extracted as explanation candidates for which the explanation generator outputs a probability distribution.

The explanation model uses as input the internal learned representation of the LSTM classifier and generates a probability distribution over the possible explanations from the dictionary. The explanation generation is an unsupervised task, guided by two desiderata: preserving the prediction performance when the explanation is appended to the input text

for classification and improving the confidence of the classifier in the correct prediction. The first criterion is evaluated using the accuracy rate on the test set for both the classifier (as a baseline) and the pre-trained classifier when the concatenation between the explanation and the original text are considered as input.

The following subsections describe the processes employed in obtaining explanations for the text classifier's individual predictions and details about the models' architecture and parameters.

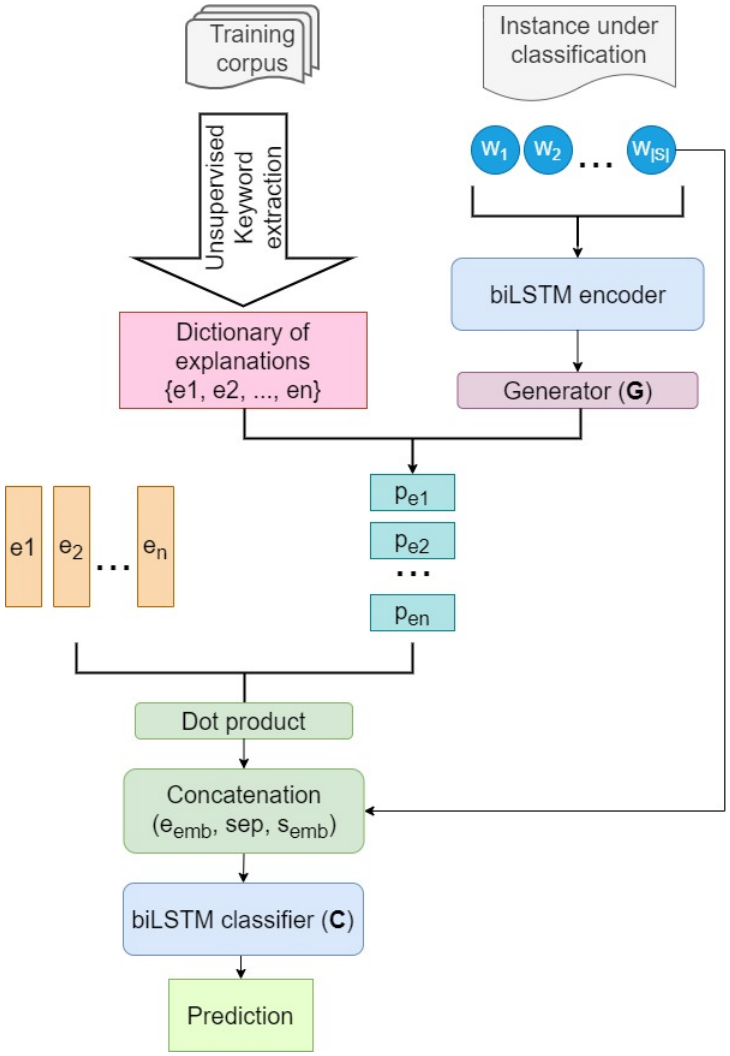


Figure 4.1: Framework architecture

4.2 Dataset Description

The IMDB movie review dataset was proposed for the sentiment analysis binary text classification task. It is a popular choice in the research community as it consists of substantially more data than previous benchmarks on sentiment analysis, making it suitable for deep learning approaches.

The dataset contains 25,000 training instances and 25,000 test instances. Having a separate test set is another advantage of this dataset that ensures a consistent comparison between different approaches reported in previous work.

Dataset acquisition. The dataset has been built such that for each movie there are no more than 30 reviews and the test and training set contain disjoint sets of movies. The author's choice of limiting the number of reviews per movie is motivated by the existence of correlations between the reviews and the associated ratings. Furthermore, the neutral rated movies have been removed from the resulting dataset that only contains positive reviews with a rating above 7 (out of 10) and negative reviews with ratings below 4.

Explainability Task Apart from the task that this dataset has been collected for, it has also been used in previous work for explainability, since the ability to automatically obtain the polarity of words and phrases opens up the possibility to automatically evaluate the explanations of the classifier's decisions. This dataset has been employed for extractive methods for explainability (Chen and Ji, 2019b).

4.3 Text Preprocessing

Structure of the data. Review texts contain similar challenges to social media textual data, where informal language usage prevails. The texts, whether they are reviews, opinions, or recommendations are often difficult to understand due to lack of proper punctuation, misspellings, word ambiguity, abbreviations, or the use of colloquial terms. The unstructured data also contain numbers, special characters, emojis, or even HyperText

Markup Language (HTML) tags, which can limit the readability of the text, even though the intent is to enrich the text with emotions through slang or graphical features. While some of the aforementioned items are intended to convey emotions, embedding them, and drawing meaningful insights is a challenge in Deep Learning.

Data cleaning. This preprocessing step is important, especially for social media data, as it removes the characters and words that do not convey insightful semantic information, preserving the aspects that are essential for conveying the transmitted message. This first cleaning process significantly improves the performance of DL classifiers. Thus, due to the noise of the unstructured text data that we are using, before extracting the explanation candidates, we are removing the punctuation, non-alphanumeric, and HTML tag characters. The remaining fundamental units of the reviews are the result of data standardization.

Tokenization. This is the process of splitting the textual data into words or phrases, which are called tokens. For our experiments, we are using the Python library Spacy (Honnibal and Montani, 2017).

4.4 Dictionary Acquisition

A model's prediction could be explained through any phrase or subset of words from the training set, but determining which phrase contributes the most when an instance is classified out of a large corpus is computationally infeasible, especially when extensive training data is being used. To alleviate this problem and still obtain the desired results, we propose extracting a subset of phrases that are acceptable candidates for explaining the model. We call this step of the framework dictionary acquisition and we describe below the applied preprocessing methods and the employed keyphrase algorithms.

Candidate extraction The most suitable candidates for the natural language explanations of the predictions are phrases that capture the semantics of a text instance's context.

Thus, we explore the impact of a few unsupervised keyword extraction algorithms, also assessing their capacity to convey insightful explanations in the classification context that they are used. We focus our experiments on the statistical approaches of keyword extraction as they are language and topic-independent methods that allow us to extend the framework to other classification tasks for different languages (in future work).

4.4.1 Keyword Extraction

TF-IDF Using the TF-IDF scores to extract keywords leads to 1-word explanations that are not always very insightful for interpretability. We have performed experiments to output multiple 1-word explanations, but this approach not only loses the contextual meaning of the explanation but it also difficult to interpret and to draw useful insights from it.

4.4.2 Keyphrase Extraction

In order to obtain more compelling explanations, we considered using YAKE, TextRank, and RAKE. For implementation, we have been using the following Python libraries: yake[1], pytextrank [2], and rake-nltk [3], respectively. While the average number of words for the phrases extracted with YAKE is 2.7 and for TextRank 2.2, we opt for using RAKE, for which we managed to obtain 4 and 5-word phrases for the considered corpora. The preference for multiple-word phrases as explanations for instance predictions is due to the semantic richness and to the highest probability of increasing the confidence of the classifier when long input texts are being classified.

RAKE-instance For this variation, we apply RAKE on individual reviews to generate the score for each phrase and we filter out the neutral-sentiment phrases using VADER. Even after applying this first filtering criterion, there are numerous phrases having the

[1]https://pypi.org/project/yake/
[2]https://pypi.org/project/pytextrank/
[3]https://pypi.org/project/rake-nltk/

exact score, since this statistical approach is applied to texts with similar lengths and frequencies of words for each review. This makes it difficult to select a diverse subset of phrases.

To alleviate this shortcoming, we create bins of phrases with the same score and we generate an exponential distribution representing the numbers of phrases to be selected for each bin, k_{bin}. Within each bin, we sort the phrases by the second criterion, the frequency in the corpus and, select only the top k_{bin} phrases. This method allows us to select the top most frequent phrases for each score value generated by RAKE, using the exponential distribution to threshold the number of phrases selected within each bin, proportional to the value of each RAKE score.

RAKE-corpus For this variation, we apply RAKE on the corpus consisting of all the instances from each class. Similarly to the previous approach, to increase the impact of the chosen explanation on the prediction rate, we further filter out the extracted keyphrases that have a neutral polarity, using VADER.

Compared to RAKE-instance, this method extracts phrases that are more specific to the topic of the corpus, rather than general adjectives with high polarity scores. The phrases contain details such as proper nouns, movie categories, aspects related to the plot, actors, performance, or to the director of the movie.

4.5 Detailed Model Architecture

4.5.1 Baseline Classifier

The model architecture used on the text classification task that we are aiming to explain through the proposed interpretability approach is a bidirectional LSTM network. BiLSTMs are a popular choice in the research community for text representation, achieving state-of-the-art results on text classification tasks. We perform multiple experiments for the sentiment analysis task using biLSTM models with different parameter configurations and

we choose a model that achieves satisfactory results on this task. We will refer to this model as VLSTM (Vanilla LSTM).

The goal of the next component of the architecture (the explanation generator model) is to preserve the performance of the classifier and to generate explanations based on the chosen classifier.

4.5.2 Explainer

To determine what features learned by the classifier from the training data influence the most the final prediction, we use a second module: the explanation generator.

For ML models, there is a trade-off between interpretability and performance. The models that achieve results comparable to humans are complex and less interpretable, while the models that are easier to interpret such as decision trees, linear regression, Naive Bayes, or k-Nearest Neighbor, are more transparent when it comes to prediction, but they underperform compared to DL models. **Our approach tries to balance this trade-off between interpretability and performance by using a standard deep MLP architecture trained to learn the patterns that a LSTM model identifies in the training set and to generate natural language explanations for individual predictions.** The MLP models have the simplest neural architecture, which makes them more interpretable than the more complex recurrent or convolutional networks. Gradient-based approaches such as feature visualization (Olah et al., 2018) or network dissection (Bau et al., 2020) allow one to visualize the information flow through the network.

Our motivation is also based on previous work on interpretability for biLSTMs (Camburu et al., 2018), where the authors try to interpret the biLSTM generated representation before the MLP classifier uses it to output the task-specific prediction.

To reach the best results, we have conducted multiple experiments using a MLP explanation generator. We have performed manual tuning to establish the model's number of layers, activation functions of the dense layers, dropout rate, regularization rate, word embedding type, loss function, and the hyperparameter of the loss function.

Unsupervised explainer. The explanation generation model is trained on the classification task, using only desiderata for the generated explanations and lacking gold-truth labels for the interpretability task. Thus, the labels for the sentiment analysis task are used primarily for the classifier, the explainer drawing insights from the learned patterns by identifying the phrases or words that improve the baseline model's confidence in the correct prediction.

Input. To explain the learned patterns that the biLSTM classifier uses to make predictions, the MLP classifier associates possible explanations to different configurations of the last layer of the classifier through neural predictions. The input of the explainer is, thus, the concatenation of the last hidden layers: forward and backward, obtained after a forward pass of the test instance.

Output. The MLP explainer generates a probability distribution for the dictionary of explanation candidates acquired in an unsupervised manner (described in Section 4.4). This probability distribution reveals the features that contribute the most to the correct classification.

Desiderata. The MLP explainer is trained in an unsupervised manner on the interpretability task, using as training guidelines, enforced through the loss function described below, the following desiderata:

1. to preserve the overall performance of the classifier when adding the explanations to the text instance;

2. to improve the confidence of the classifier for a given instance on the text classification task.

Loss function. To achieve these desiderata, we propose a loss function that has two parts. The first part (Equation 4.1) is the binary cross entropy applied between the target label y and the predicted label after the explanation is added for classification $f(x \cdot e_x)$. The

classified instance is denoted by x, its corresponding explanation is e_x, and the $\cdot$ operation denotes their concatenation.

$$\mathcal{L}_1(x) = BCE(f(x \cdot e_x), y). \tag{4.1}$$

The second part consists of a custom loss function (Equation 4.2) that aims to maximize the contribution C of the explanation e_x (further detailed in Section 4.6.2.1).

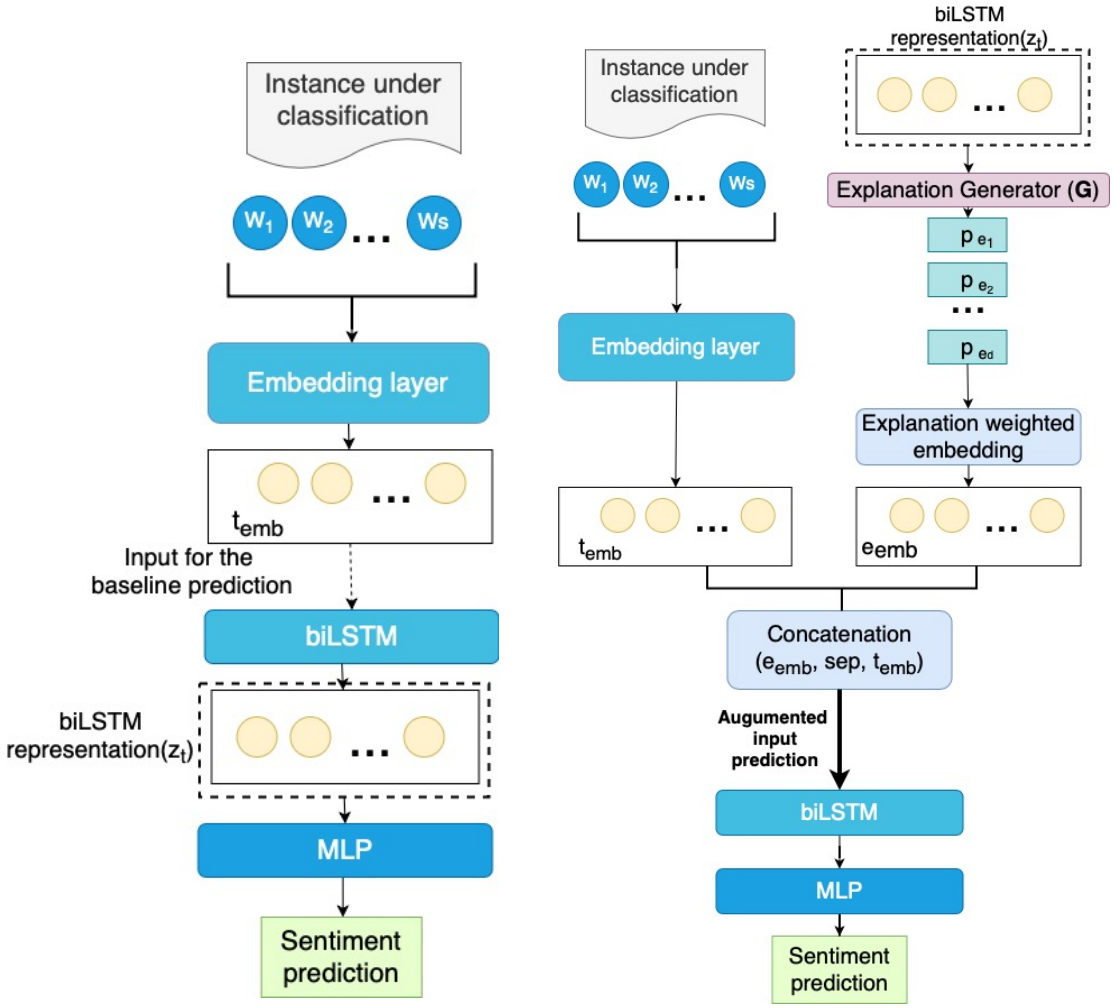
$$\mathcal{L}_2 = 1 - C(e_x) \tag{4.2}$$

The final loss function is parametrized by a hyperparameter, α, that weights the two desiderata:

$$\mathcal{L} = \alpha\mathcal{L}_1 + (1 - \alpha)\mathcal{L}_2 \tag{4.3}$$

While the classifier can be evaluated on the text classification task that it has been trained for, the explanation generator does not have gold-truth labels. Its training is only guided through the desiderata enforced by the custom loss function in Equation 4.3.

Additional experiments with variations of the loss function. In addition to the aforementioned two-term loss function from equation 4.3, we have also performed experiments with solely the $\mathcal{L}_2$ loss function (equation 4.2) and with a three-term loss function that aims to also maximize the cosine similarity between the generated explanation and the explained text. The next chapter reports the results for those experiments. We note that using solely the loss function from equation 4.2 leads to a slow learning process with a less steep learning curve since it aims to maximize only the contribution value, thus only indirectly preserving the prediction, while the similarity enforcement represents a constraint for the generator that negatively impacts the contribution values. In our comparative analyses, described in chapter 6, we found that the models can rely on phrases of the training set that might not be related to the context of the classified text, which translates into the model making an uninformed decision. We, thus, empirically concluded that the loss function in equation 4.3 obtains the best results, compared with those variations, so we focus our attention on exploring the results of this approach.



(a) Prediction of the original input text (b) Prediction of the augmented input with the generated explanation

Figure 4.2: A figure with two subfigures

Figure 4.3: Explainer - Architecture diagram. $w_1, w_2, ..., w_n$ – the words of a classified instance with n words, embedded into the text embedding t_{emb}. biLSTM – the explained text classifier. G – the explanation generator outputting the probability distribution $p_{e_1}, p_{e_2}, ..., p_{e_d}$, where d is the size of the dictionary of explanation candidates. e_{emb} – the weighted embedding of the explanation (according to the generated probability distribution). *sep* is the 0-vector separator between the original text and the newly added phrase used in order to minimize the impact of the contextual disruption. The biLSTM representation is the concatenation between the last hidden states: of the forward and backward LSTMs.

4.6 Evaluation Metrics

There is little consensus and no benchmark evaluation method for the explanation generation task due to the wide variety of approaches and of their end goals, therefore fair comparisons between explainability methods is very challenging. Our proposed methodology falls into the category of generative approaches, where the explanation is not part of the input review text, aiming for achieving a self-explained model that cannot be compared against human labels or other generative approaches such as prototype, counterfactual explanations, or adversarial explanations. Thus, we report the performance of the model through automatic evaluation methods and we analyze the usefulness of the proposed framework in three real-life applications through human evaluation, to assess if this method provides sufficient insights for making informed decisions about the model's behavior.

To assess the performance of the proposed approach, we evaluate it on both the classification task, with the goal of preserving the performance of the base classifier (Section 4.6.1) and on the explanation generation task (in Section 4.6.2). In the following subsections, we elaborate on the metrics used for each task, describing the proposed approach for evaluating the explanation generation model and the quality of the generated explanations.

4.6.1 Classification

For the sentiment analysis binary classification task, the goal of the proposed architecture is to preserve the baseline performance of the unaltered model that is being explained (VL-STM). We, thus, report the accuracy rate and average F1-score as **quantitative analysis**. The goal of this evaluation is to assess whether the proposed architecture manages to preserve the same performance on the classification task, or, ideally, to improve it, when the classified instances are replaced by the concatenation between the generated explanation and the initial text.

4.6.2 Explanation Generation

While the first desideratum of the framework is evaluated on the classification task (Section 4.6.1), for the second desideratum, we will evaluate both the performance of the explanation generation model by designing a performance metric for the contribution and the quality of the resulting explanations. We, thus, employ both automatic and manual methods of **qualitative analysis**, described in the following subsections.

4.6.2.1 Automatic Evaluation

The automatic evaluation presents a number of advantages. Firstly, it is a fast and accessible method that allows us to compute metrics that can be suitable for both lay humans and ML experts. Secondly, it does not require specialists or lay humans to be trained for evaluation, being less expensive, and requiring fewer resources. We further describe two methods of assessing the quality of the generated explanation through the confidence and the coherence score defined below.

Confidence score. We propose the contribution metric (defined by Equation 4.4) as a way of measuring how much impact the concatenation of the explanation and the instance to be classified has.

The contribution of an explanation generated for an instance x represents the added value in the correct prediction, a similar approach to the Shapley values from game theory (Molnar, 2019). For our use case, we propose the formula in equation 4.4, where f is the baseline classifier and e_x is the generated explanation for the instance x using the explanation generator g and $\cdot$ represents the concatenation operation. The positive and negative labels are represented by 0 and 1, respectively. The higher the contribution coefficient, the more impact the explanation has in **improving the confidence** of the classifier **towards the correct prediction**.

$$C(e_x) = \begin{cases} f(e_x \cdot x) - f(x), & \text{label} = 1 \\ f(x) - f(e_x \cdot x), & \text{label} = 0 \end{cases} \tag{4.4}$$

The proposed method compares the prediction confidence on a given test instance with the prediction for the instance concatenated with the corresponding generated explanation. For example, if the classifier predicts with a confidence of 0.6 that a given test instance is 1 ($f(x) = 0.6$) and the instance together with the generated explanation leads to a prediction of 0.8 ($f(e_x \cdot x) = 0.8$), then the explanation's contribution is 0.2. Conversely. if $f(e_x \cdot x) = 0.3$, then the contribution will be -0.3. Similarly when the label is 0: if $f(x) = 0.4$ and $f(e_x \cdot x) = 0.1$, then $C(e_x) = 0.3$, and if $f(x) = 0.1$ and $f(e_x \cdot x) = 0.3$, then $C(e_x) = -0.2$.

The coherence score has been proposed in previous work (Chen and Ji, 2019c) and reported for two extractive methods of explanations for the IMDB dataset. In Chen and Ji (2019c)'s work, the score represents the percentage of the explanations that have the same polarity as the prediction of the classifier or, when the explanation is neutral if the prediction is different than the label. This approach is suitable for extractive methods, where the explanation's polarity should match the prediction. For the generative methods, where the goal is to improve the model and increase the confidence in the correct prediction, it is important for the explanation to match the gold-truth label. We are, thus, using a variation of this coherence score.

Therefore, we will assess the coherence of the generated explanations with respect to the label, since the contribution of the explanation may not suffice for changing a wrong prediction. For example, if a positive review is misclassified with the predicted probability of 0.2, and a positive explanation can improve the prediction by 0.2, the explanation generator succeeded in choosing a positive contribution explanation, but the value-added does not suffice for a correct prediction. Another difference that we have in the coherence score is the absence of neutral explanations since the dictionary of the explanations has been preprocessed to only contain strong sentiments (either positive or negative). The

evaluation method in this case reflects if the explanation generator managed to correctly choose an explanation for which the polarity matches the prediction.

4.6.2.2 Human Evaluation

To further confirm the results reported using the coherence score, we define experiments for human evaluation, using a sample of 100 examples, that are used to also confirm the automatic evaluation using the coherence score. We employ these experiments for both assessing the interpretability of the explanations for one model only and for comparing the performance of two different models.

Interpretability for individual models. For this evaluation, we designed an experiment where the humans are presented with the input review, the generated explanation, and the confidence score of the explanations, and they are asked to estimate the model's prediction (positive or negative). This evaluation method assesses the interpretability of the generated explanations.

We have also performed a human evaluation on a similar task, where the humans were presented with the text review, the explanation, and the label, and were asked to estimate the model's prediction. For this experiment, we compared the coherence between the human's estimations and the sentiment polarity of the explanation. The results and conclusions for these experiments are further detailed in Section 6.2.1.

The sample of examples contains 50 instances correctly classified and 50 instances incorrectly classified, which have the highest contribution rates.

Interpretability for model comparison. Another evaluation method for model interpretability focuses on determining if the explanations are compelling enough to determine, out of two models that have a similar prediction rate, which one performs best on unseen data. Our experiment is designed such that the humans are presented with an explanation for each of the two considered models (M_1 and M_2) and are asked to decide which one out of the two correctly classifies the review.

Similar to the previous experiment, we select the instances with the highest contribution values for the explanations, 50 of them being correctly classified and 50 incorrectly classified. Thus, for each review text, one model correctly classifies it and one does not. To ensure that the sample represents the test set, for each of those 2 subsets, half of them have a contribution value of the correct class higher than that of the incorrect class.

4.7 Chapter Summary

This chapter described the general architecture of the proposed explanation generation approach, starting from the early text preprocessing steps, to describing the dictionary acquisition methods used for building the collection of the possible explanations, the motivation for the proposed neural model, the desiderata that guide the training process, and the evaluation methods employed for both the classification and for the interpretability task. While we presented here the high-level architecture of the proposed framework, in the next chapter, we further detail the architectural decisions, the choice of the hyperparameters, and the model selection process, presenting the impact of these choices on the classification performance.

Chapter 5

Experiments and model selection

Our methodology, model architecture, and parameter setting are supported by numerous experiments that are described in this chapter. We discuss here the results that we report, motivating our choices and processes that led to the final parameter setting that obtains the best results for the interpretability task. This chapter describes the experiments that we have performed to determine the best parameters for the baseline classifier (VLSTM) and for the explanation generator G. We are elaborating here on the choice of the word embeddings, architectural decisions, hyperparameter optimization, and motivation for the performed experiments. The model that performed best on the classification task, thus fulfilling the first desideratum of the explainer, is further evaluated on the interpretability task.

5.1 Dictionary of Explanation Candidates

In our reported results, we use dictionaries with different numbers of entries, using as extraction methods RAKE-corpus and RAKE-instance described in Section 4.4.2.

Small dictionary. Firstly, we use RAKE-corpus to extract a small dictionary of 60 phrases of maximum 5 words: 24 extracted from the positive corpus and 36 from the negative corpus. Since RAKE merges adjacent keywords to form phrases, when the keyphrases

are extracted from the corpus of the concatenated reviews in a single document for each of the two labels, the resulting phrases contain phrases such as "funny funny funny funny" or "bad bad bad bad". Recall also that the stop words are filtered out before the extraction of candidate words. Despite this shortcoming of the method, we still chose this dictionary since after filtering out the neutral-sentiment phrases, they still contain the general opinion conveyed in the reviews and are able to influence the prediction of the classifier, even if they do not contain granular aspects about the reviewer's point of view. The entries of the dictionary are listed in Appendix E.

Large dictionary. For comparison, we also employ a few large dictionaries of 589 entries obtained using RAKE-instance. **Firstly**, we experiment with a 589-entry dictionary that contains 4-word phrases: 298 extracted from the positive reviews and 291 from the negative ones to determine the architecture and hyper-parameters of the explanation generator model G. This dictionary is extracted after the extracted phrases have been sorted reversely by the RAKE outputted score and alphabetically. The polarity of the extracted phrases is not always consistent with the polarity of the text of the review from which they have been extracted, since the reviews usually contain both appreciations and critics. Thus, for the positive reviews, 71.14% of the selected keyphrases are positive, while for the negative reviews, 60.48% of the extracted phrases are negative. Overall, for the entire training corpus, the dictionary contains 327 positive-sentiment and 262 negative-sentiment keyphrases. We will refer to this dictionary as **RAKE-instance-for-hyperparameter-tuning**. **Secondly**, we refine the list of candidates and select the highest score, most frequent phrases by preserving the order from the corpus, building the **RAKE-instance** dictionary that we use for both quantitative and qualitative evaluation of the selected explanation generator model, as it comprises a more diverse collection of phrases. This dictionary contains 291 candidates for the positive reviews, 66.32% of conveying a positive sentiment, and 290 phrases extracted from the negative reviews, 55.17% of them conveying a negative sentiment. The dictionary contains 323 positive-sentiment phrases and 258 negative phrases. The explanation candidates are listed in Appendix D.

5.2 Word Embeddings

Due to the unstructured nature of the textual data that we are using for our experiments, we have trained word embeddings of size 300, randomly initialized using the uniform distribution. This approach not only leads to a satisfactory baseline, but it also alleviates the out-of-vocabulary issue of the pre-trained embeddings. The informal language used in the movie review dataset abounds with abbreviations, slang words, and proper nouns that are often not part of the GloVe vocabulary of words. This aspect is also reflected by the accuracy and average F1-score obtained on the test set and reported in Table 5.1.

Word Embedding	**Acc. (%)**	**Avg. F1**
Glove	86.06	0.7587
Trained (uniformly initialized)	87.75	0.7750

Table 5.1: Word Embeddings for VLSTM - performance comparison on the classification task

The VLSTM model used for this experiment has 2 stacked LSTM layers with 256 features of the hidden state, trained using Adam Stochastic Optimization (with the default parameters from PyTorch[1]) and a dropout rate of 0.5.

5.3 Base Classifier

The model that we consider in our framework for interpretability as the baseline classifier is a recurrent neural network bidirectional LSTM architecture. This category of models proved to obtain satisfactory results in previous work for text classification, including sentiment analysis. This method is used for sequential tasks in particular since it captures the contextual information. BiLSTM networks have a better performance for NLP tasks since they capture information of the neighboring words from both directions (before and after the current token).

[1]https://pytorch.org/

To further establish the hyperparameters of the model that will be explained, we perform manual hyperparameter tuning and report the accuracy and average F1 score in Table 5.2. We aim to determine the optimal number of stacked LSTM layers, together with the number of hidden units of the hidden states, and also determine the dropout value that helps the model to avoid overfitting.

The reported results have been obtained by using trained 300-dimensional embeddings randomly initialized using the uniform distribution. The models were trained for maximum of 40 epochs, the best model is chosen based on the minimum loss value obtained on the validation set. We have also used an early stopping condition with a patience value of 20, which means that after 20 epochs (not necessarily consecutive) that increased the validation loss value compared to the previous epoch, the training stops. For each experiment, for evaluating we choose the model that obtained the lowest loss value on the validation set and report its results for the test set.

The experiments show that for the 1-layer biLSTM the model reaches a performance plateau within the first 5 epochs, the network capabilities of generalization being exceeded very quickly. The dropout is only applied to the non-final layers of the stack, so it has no effect on single-layer architectures. For the multi-layer models, the better performance is further improved when the number of hidden units is increased, as well as the dropout value, to avoid overfitting.

The models that we have considered for the rest of the thesis as baselines are VLSTM, VLSTM+, and VLSTM-experim. For VLSTM-experim, we have performed some of the early experiments with different vocabulary types and experimented with the jointly trained architecture of the explainer.The VLSTM and VLSTM+ models were chosen for an experiment that uses two models with comparable performance, yet one performing slightly better on the test set, compared to the other. Thus, we used those two models for a comparative analysis of the final explainer in order to assess which model would perform better on the test set when only the explanations are shown to the human evaluators.

Although from the results of the experiments that we have performed in order to choose the baseline model, we can estimate that larger models would further improve the accuracy

Num. Layers	**Hidden units**	**Dropout**	**Acc.(%)**	**Avg. F1**	**Model name**
4	256	0.3	85.33	0.7569	VLSTM-4-256
		0.5	86.05	0.7574	
2	256	0.05	85.77	0.7601	**VLSTM+**
		0.1	86.70	0.7651	
		0.3	86.61	0.7658	
		0.5	**87.75**	**0.7653**	
2	128	0.05	84.51	0.7420	VLSTM-2-128
		0.1	85.54	0.7526	
		0.3	86.66	0.7642	
		0.5	87.11	0.7684	
1	128	-	83.45	0.7304	VLSTM-1-128
2	64	0.05	83.89	0.7366	**VLSTM**
		0.1	84.66	0.7440	
		0.3	**85.01**	**0.7486**	
1	64	-	82.08	0.7194	VLSTM-1-64

Table 5.2: Manual tuning for the baseline biLSTM classifier

rate, we continue our research for explainability with the 2-layer bi-LSTM architecture and 256 hidden units since it obtains satisfactory results while also having a model size that allows us to perform multiple experiments in a timely manner. We emphasize again that the purpose of our experiments is to provide insightful explanations without losing performance and not obtaining state-of-the-art results on the classification task, thus this model-size - performance compromise is suitable for our use-case.

5.4 Explanation Generation Model

As described in Section 4.5.2, the explainer aims to determine the phrase that improves the prediction of a given test instance, by choosing one of the candidate-phrases from the dictionary of explanation extracted from the training set. Selecting a sample of explanations is a non-differentiable operation that does not permit the error backpropagation in the training process of the neural network. To alleviate this, we employ the categorical reparametrization method in order to generate a probability distribution for the explanation candidates. We analyze the performance impact of both a soft selection of the explanations and the one-hot categorical distribution.

5.5 Categorical Reparameterization

To determine the probability distribution of the dictionary of possible explanations, we compared the impact of Softmax and Gumbel-Softmax as activation functions for the final layer of the explanation generator model.

5.5.1 Softmax

Our experiments show that by applying the Softmax activation function, especially for dictionaries with a large number of entries, the explainer outputs a uniform distribution of very low probabilities (values ranging between 0.0064 and 0.0071). The resulting context vector obtained by aggregating the possible explanations will thus contain very little information extracted uniformly from the embeddings of all the dictionary's entries. This only leads to an additional noisy context vector that does not have a positive impact on the classification task. The results also confirm this claim. When using Softmax as the activation function of the explanation generator's last layer, we obtain a decrease in performance of approximately 1-2% for all the considered metrics (compared to the baseline classifier). We can conclude that the bi-LSTM classifier is able to detect that this additional vector is noisy since the performance on the classification task is not impacted drastically.

5.5.2 Gumbel-Softmax

To alleviate the aforementioned shortcoming, we continued our experiments using Categorical Reparametrization with Gumbel-Softmax (Jang et al., 2017). This way, the obtained context vector will comprise information aggregated from one or a few explanations, rather than a mixture of all the dictionary's entries. Our claims are empirically confirmed by the results reported in Appendix, Table A.1.

5.6 MLP Explainer Experiments

We have performed experiments with the MLP explainer both by jointly training it with the classifier and by using the pre-trained classifier only for the forward pass of the text instances. We describe both approaches and findings in the following subsections, concluding that the pre-trained classifier with the weights frozen during the training of the explainer achieves a satisfactory performance faster and it is more feasible for real use cases when the model is already trained when the justifications for predictions are required. To ensure that the classification performance is not impacted by the choice of the dictionary acquisition method or the architecture of the MLP model, we report the accuracy and average F1 when classifying the test instances concatenated to the corresponding generated explanations.

5.6.1 Jointly-Trained MLP

To explain the biLSTM classifier, we have run experiments where the MLP explainer and the classifier are trained in an end-to-end manner on the classification task. This approach requires only the parameter configuration of the classifier and the training data in order to provide individual prediction justifications. For this approach, the classifier is retrained together with the explanation generator, which adds complexity to the training process. The increase in the number of trainable parameters translates into higher training time.

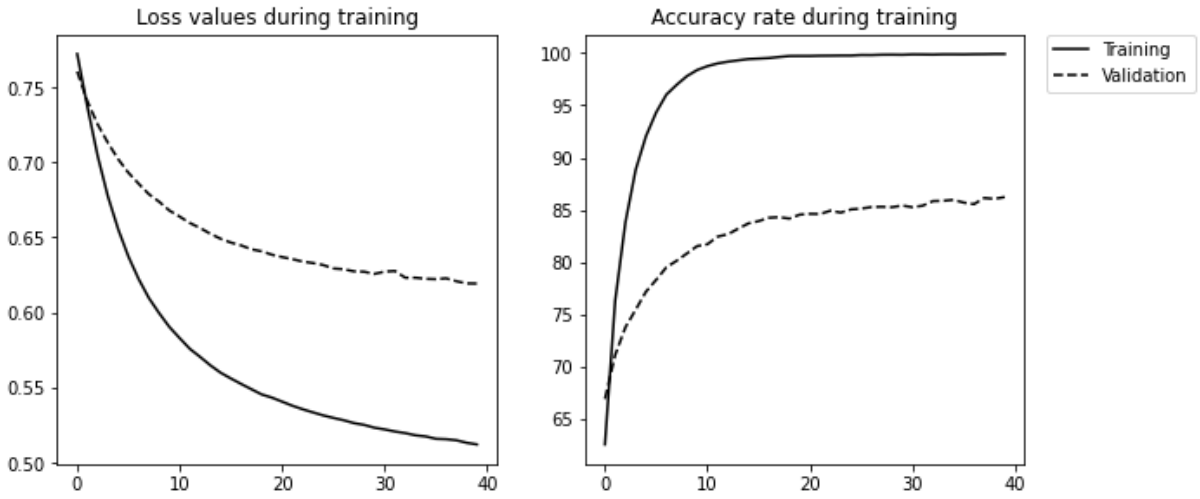


Figure 5.1: Training loss and accuracy for the jointly trained MLP explainer

Dictionary. The dictionary employed for this experiment is the small dictionary described in Section 5.1, acquired with RAKE-corpus.

Baseline model. The baseline model that we have considered for this experiment is VLSTM+ form Table 5.2.

MLP explanation generator. We employed a 40-layer Deep Neural Network (15 layers with 512 neurons and 25 layers with 256 neurons), trained using a dropout rate of 0.05, Adam Optimization with decoupled weight regularization (Loshchilov and Hutter, 2019), and learning rate of 0.02. The alpha hyperparameter of the loss function is set to 0.7 since our experiments reported in the next section have proven it is the optimal value for the trade-off between the two desiderata of the explainer.

Classification results. The model obtains the lowest validation loss value after the last epoch of training (40), the results of this model on the test set being 83.88% accuracy and 0.7355 average F1-score.

Conclusions. As it can also be observed from the loss and accuracy values in Figure 5.1, the jointly trained model has a slow but steady learning progress. Solutions to this

phenomenon of slow convergence of the model such as increasing the learning rate can be further employed in order to obtain a steeper learning curve. The increased number of parameters (38 million trainable parameters) of the joint model leads to a slower training process, which might not be feasible for real-life applications when interpretability is required, especially when hyperparameter tuning is also needed.

5.6.2 MLP Explainer Using the Pretrained biLSTM Classifier

Since jointly training the explainer and the classifier has a slow convergence, also due to the change of the input of the explainer during the training process for the same training data point, we are further analyzing the performance of the MLP explanation generator when the biLSTM model has the parameters fixed.

Model	**Params**	**Acc. (%)**	**Avg. F1**
VLSTM		85.77	0.7601
MLP gen	$\alpha = 0.7$	**86.12**	0.7603
	$\alpha = 0$	84.62	0.7448
MLP-sum	$\alpha = 0.5$	84.99	0.7485
	$\alpha = 0.7$	**86.05**	0.7589
	$\alpha = 0.75$	73.90	0.6331
	$\alpha = 0.8$	74.66	0.6398
JT MLP	$\alpha = 0.7$	79.43	0.6947

Table 5.3: Hyperparameter and loss function comparisons

Hyperparameter for the loss function. The desiderata for the explanation generator model G are obtained by minimizing the loss function that incorporates both goals. The weight of each of the two goals is controlled through the hyperparameter $\alpha \in [0, 1]$ from the loss function in Equation 4.3: $\mathcal{L} = \alpha\mathcal{L}_1 + (1 - \alpha)\mathcal{L}_2$. We have empirically determined that the optimal value for α is 0.7, by performing experiments using the *VLSTM-experim* model.

The results obtained on the classification task of the ensemble of explanation generator and classifier are reported in Table 5.3. We compare here the results obtained using different aggregation methods for the loss function $\mathcal{L}_2$ that focuses on improving the contribution. The model that performs best, **MLP gen**, uses average reduction, while the model MLP-sum uses a discount factor (100) for the summed losses for each batch. The jointly trained model (explanation generator and classifier), referred in the table by *JT MLP*, has a slower learning curve and a more time-consuming training process, underperforming (when using the same number of epochs for training as with the explanation generator trained for the *VLSTM* model).

Model architecture. To choose the number of layers and hidden units of the explainer, we perform experiments aiming to preserve the performance of the considered baselines for the test set. For the experiments reported in Table 5.4, we use the RAKE-instance dictionary to determine the optimal number of layers and dropout rates for the MLP explanation generator. Our experiments focus on determining the best performing model for the VLSTM+ baseline that preserves or improves the accuracy rate and then using it for the smaller model (VLSTM), to test whether it has the same power of generalization and it can preserve the same performance.

The experiments performed for the VLSTM+ model show that a MLP explanation generator with a smaller number of parameters underperforms compared to the baseline, which means that the contribution values of the generated explanations do not have high enough values to change a correct prediction and even have a negative influence on some of the data points. However, larger models manage to preserve the same classification rate when the dropout rate is also increased.

The 60-layer MLP explanation generator that obtains the best classification results is further used to assess its performance when different dictionaries are used. Table 5.5 reports the results when the small (RAKE-corpus) and large (RAKE-instance) dictionaries are used.

With this setting of the explanation generator that improves the accuracy of the base

classifier, we further report the results both as quantitative (in Section 6.1) and qualitative analysis (in Section 6.2) for the interpretability task.

5.6.3 Maximizing the Cosine Similarity

With the VLSTM+ model, together with the RAKE-instance dictionary, we performed another set of experiments that aim to determine whether enforcing the semantic similarity between the generated justification and the explained text has a positive impact on the interpretability of the final model, while still preserving a similar classification rate. To achieve this goal, we use a model variation obtained by modifying the loss function with an additional function (Equation 5.1) that aims to maximize the cosine similarity between the text embedding x and the generated justification's embedding e_x.

$$\mathcal{L}_3 = 1 - Cos(x, e_x) \tag{5.1}$$

Loss function The final loss function contains, thus, the terms of the original equation (4.3): $\mathcal{L}_1$ – the binary cross-entropy loss function, and $\mathcal{L}_2$ – the function comprising the contribution score, and the additional term $\mathcal{L}_3$, weighted by the hyperparameter γ. Equation 5.2 aggregates all three criteria of the alternative model: the binary cross-entropy loss, the contribution maximization function, and the semantic similarity measure. Table 5.6 reports the classification impact of this enforcement of the semantic similarity between the justification and text instance, also summarizing the interpretability metrics. While the purpose of this change is to improve the quality of the generated explanations, enforcing them to be contextual similar to the text of the review, this is a compromise with the fidelity to the model. When the classifier relies on some semantic concept unrelated to the review to make the classification, it might become difficult to correlate with the classified text, but it also conveys the fact that the model is not relying on concepts that should in reality improve the performance. Thus, a justification that is similar to the context of the review may be easier to interpret by humans, but not faithful to the model. This phenomenon can happen for the classifiers that do not generalize well and have numer-

ous incorrect classifications. In such cases, the trade-off between model faithfulness and interpretability tends to incline towards interpretability, the explanation generator being inclined to choose a semantic similar justification that does not maximize the contribution value.

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 \tag{5.2}$$

Hyperparameter tuning To tackle the trade-off problem between the model faithfulness and interpretability, we perform grid search for the values of the hyper-parameters α, β, and γ of the loss function and we report in Table 5.6 metrics for both the faithfulness of the model and the interpretability contributions on the unique reviews of the test set. This subset of the original subset contains 24828 instances.

Model faithfulness. The baseline model (VLSTM+) obtains an accuracy rate of 87.75 for the sentiment analysis classification task. To measure the faithfulness of the explaining model to the original classifier, we report the obtained accuracy rate for the same test set.

Our experiments test the hypothesis that an equal weight for the three criteria would obtain the best results for the G_cos proposed explainer. This is confirmed by the experiments for the triplets $(\alpha, \beta, \gamma) \in \{(0.35, 0.3, 0.35), (1, 1, 1)\}$ that obtain the best accuracy rates. Comparably, the value of 0.7 for the cross-entropy function's weight, which has been chosen in the previous experiments for the eVLSTM+ model, manages to preserve a similar performance rate (87.31) when the similarity measure and the contribution criterion have equal weights (0.15). This set of experiments shows that the increase of the γ parameter is inversely proportional to the variation of the total number of changed predictions. The cosine similarity seems to have a negative impact on the contribution values, which means that high contribution values are associated with phrases that are not optimally similar to the embedding of the review that is being classified.

The baseline for the explanation generator model that only relies on the contribution criterion, without the cross-entropy or the cosine similarity functions, obtains a classifica-

tion rate of 82.70. The generalization power of the model when there is no cross-entropy loss function involved is lower than the classifier's performance by 5%. This baseline is not improved by adding the cosine similarity, although these experiments report the highest number of correctly changed predictions.

5.7 Chapter Summary

This chapter summarizes the experiments that we have performed in order to determine the hyperparameters of the explanation generator model that preserves the overall performance of the text classifier on the test set, while also generating insightful explanations. We experimented with different architectural settings for the explanation generator G in order to select two models for further analysis for the interpretability task. The two selected models (eVLSTM+ and eVLSTM) use different biLSTM classifiers that have comparable performance and that have been chosen for semantic model comparison and interpretability, the results of the evaluation being detailed in the next chapter.

Model name	Number of layers	Hidden units	Dropout rate	Accuracy (%)	Avg. F1
VLSTM	2	64	0.30	85.01	**0.7486**
eVLSTM	30; 30	512; 256	0.80	**85.13**	0.7483
VLSTM+	2	256	0.50	86.08	0.7650
eVLSTM+	10; 10	512; 256	0.05	85.96	0.7571
	15; 25	512; 256	0.30	86.78	0.7654
	15; 25	512; 256	0.50	87.33	0.7721
	15; 25	512; 256	0.60	87.42	0.7716
	20; 20	512; 256	0.65	87.51	0.7742
	20; 25	512; 256	0.60	87.40	0.7730
	20; 25	512; 256	0.65	87.68	0.7736
	20; 25	512; 256	0.70	87.65	0.7747
	25; 25	512; 256	0.60	87.69	0.7749
	25; 25	512; 256	0.65	87.68	0.7760
	25; 25	512; 256	0.70	87.67	0.7747
	25; 30	512; 256	0.50	87.22	0.7710
	25; 30	512; 256	0.60	87.02	0.7672
	30; 30	512; 256	0.30	86.54	0.7647
	30; 30	512; 256	0.40	86.69	0.7650
	30; 30	512; 256	0.50	87.30	0.7727
	30; 30	512; 256	0.60	87.13	0.7732
	30; 30	512; 256	0.65	87.32	0.7703
	30; 30	512; 256	0.70	**87.88**	**0.7756**
	40; 40	512; 256	0.50	87.41	0.7725
	40; 40	512; 256	0.60	87.15	0.7690
	40; 40	512; 256	0.70	87.36	0.7707
	50; 50	512; 256	0.30	86.50	0.7624
	50; 50	512; 256	0.50	87.27	0.7713
	50; 50	512; 256	0.60	87.42	0.7736
	50; 50	512; 256	0.70	87.46	0.7747

Table 5.4: MLP explanation generator classification performance (large dictionary–RAKE-instance-for-hyperparameter-tuning) for the 25000-instance test set

Model name	Dictionary	Accuracy (%)	Avg. F1
VLSTM	-	85.01	0.7486
eVLSTM	RAKE-instance-for-hyperparameter-tuning	85.13	0.7483
	RAKE-instance	**85.71**	**0.7551**
	RAKE-corpus	84.71	0.7458
VLSTM+	-	86.08	0.7650
eVLSTM+	RAKE-instance-for-hyperparameter-tuning	**87.88**	0.7756
	RAKE-instance	87.77	0.7752
	RAKE-corpus	87.75	**0.7766**

Table 5.5: MLP explanation generator classification performance for the 25000-instance test set

α	β	γ	Acc. (%)	Correctly changed	Incorrectly changed	Total changed pred.	Correctly changed pred.(%)	Correct pred. positive contrib. (%)	Incorrect pred. positive contrib. (%)	Total positive contrib. (%)
VLSTM+ baseline			**86.08**	-						
0.7	0.3	0.001	87.41	1469	1139	2608	56.33	65.53	**43.04**	62.70
		0.01	87.59	1641	1267	2908	56.43	65.53	36.09	61.88
		0.1	**87.68**	1560	1165	2725	**57.25**	59.24	35.39	56.30
		0.2	87.47	1601	1256	2857	56.04	60.14	35.56	57.06
		0.3	87.48	1721	1375	3096	55.59	74.58	31.33	69.17
		0.4	87.65	1631	1242	2873	56.77	68.04	33.01	63.71
		1	87.67	1692	1299	2991	56.57	65.44	32.56	61.38
	0.15	0.15	**87.68**	1653	1256	2909	56.82	78.60	30.15	72.64
0	1	0	82.75	1698	2526	4224	40.20	**88.63**	27.83	78.14
		0.3	85.27	1920	2121	4041	47.51	87.90	30.69	**79.47**
		0.4	85.11	1981	2223	4204	47.12	87.64	28.97	78.91
		0.7	83.68	**2093**	**2689**	**4782**	43.77	88.29	25.67	78.03
		1	85.29	1894	2091	3985	47.53	77.22	33.08	70.73
0.3	1	1	86.54	1866	1753	3619	51.56	84.71	31.42	77.54
0.5			86.56	1814	1697	3511	51.67	79.07	31.37	72.66
0.7			86.85	1773	1583	3356	52.83	78.59	31.21	72.36
1			87.54	1745	1383	3128	55.79	68.03	32.23	63.57

Table 5.6: Hyperparameter tuning - explanation generator using the cosine similarity (G_cos) on the unique reviews test set (24828 instances). The VLSTM+ baseline is the accuracy obtained on this test subset. The accuracy obtained for $(\alpha, \beta, \gamma) = (0, 1, 0)$

Chapter 6

Interpretability Evaluation

This chapter focuses on evaluating the interpretability level of the proposed framework, along with the analysis of the generated explanations and their contribution to the classification task.

For the interpretability evaluation, we analyze the performance of the models that perform best on the classification task (eVLSTM and eVLSTM+) together with the RAKE-instance dictionary acquisition method, proposing both a quantitative analysis and a qualitative analysis.

For the **quantitative analysis**, we report and interpret the contribution values of the generated explanations for the test instances and the coherence between the sentiment of the generated explanation and the gold-truth label.

The **qualitative analysis** consists of human evaluation tasks, the analysis of particular examples of the review, and their corresponding explanations, along with examples of the most frequent explanations in order to draw insights about the classifier's behavior.

6.1 Quantitative Analysis

The natural language justifications are chosen to maximize the marginal contribution to the classification task. To assess the impact of adding a keyphrase to a text instance that is

classified by the pre-trained deep learning model, we compute the contribution metric for each test instance, according to the formula described in Section 4.6.2.1. To better assess the contribution value distribution for the test set, we provide summary statistics: the average, minimum, maximum, and standard deviation for the entire test set and for the correctly classified, and incorrectly classified instances, respectively. Furthermore, metrics such as the percentage of positive contributions and the number of the predictions that have been changed by the justification allow us to compare and identify the models that better generalize on the interpretability task while still remaining faithful to the model's predictions. This section will focus on the justification generated for the VLSTM and VLSTM+ models, using the dictionary extracted using RAKE-instance, but we also report our automatic evaluation metrics in Appendix B for RAKE-corpus.

6.1.1 Explanation's Polarity

Using the VADER sentiment analyzer, we report metrics related to the polarity of the generated explanations, relative to the model's predictions and expected labels, in a breakdown depending on the contribution value (positive or negative), and the correctness of the predictions. Tables 6.1 and 6.2 reflect the metrics for eVLSTM, while Tables 6.3 and 6.4 show the metrics for the eVLSTM+ model.

Contribution sign	**ACC**	**AIC**	**All**
Positive	61.75	23.42	56.65
Negative	32.83	63.72	37.55
All	48.91	42.92	48.06

Table 6.1: Polarity consistency with the **prediction** of eVLSMT (%): ACC – All Correctly Classified Instances, AIC – All Incorrectly Classified Instances, and All Instances

Contribution sign	**ACC**	**AIC**	**All**
Positive	61.75	76.58	63.73
Negative	32.83	36.28	33.36
All	48.91	57.08	50.07

Table 6.2: Polarity consistency with the **label** of eVLSMT (%): ACC – All Correctly Classified Instances, AIC – All Incorrectly Classified Instances, and All Instances

Contribution sign	ACC	AIC	All
Positive	53.93	44.03	53.07
Negative	43.50	57.72	46.05
All	50.18	51.75	50.37

Table 6.3: Polarity consistency of the explanations with the **prediction** of eVLSMT+ (%): ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances

Contribution sign	ACC	AIC	All
Positive	53.93	55.97	54.10
Negative	43.50	42.28	43.29
All	50.18	48.25	49.94

Table 6.4: Polarity consistency of the explanations with the **label** of eVLSMT+ (%): ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances

6.1.2 Contribution Values

The contribution values obtained by appending the explanation to the input text in the classification process, described in Section 4.6.2.1 represents an automated evaluation method for assessing the impact of the justification in the classification process of the original model.

Tables 6.5, 6.6, 6.7, and 6.8 report the aggregated metrics for the contribution values obtained for the explanations of the test instances.

Metric	**ACC**	**AIC**	**All**
Mean	0.0024	-0.0080	0.0011
Min	-0.3430	-0.4539	-0.4539
Max	0.3947	0.2868	0.3947
Std. dev.	0.0314	0.0670	0.0377
# of pos. C	14079	1285	15364
Total	21835	3009	24844
% of pos. C	64.48	42.71	61.84
# of changed predictions	1544	1093	2637

Table 6.5: Individual contribution values (C) for: ACC – All Correctly Classified Instances, AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM+ (RAKE-instance-for-hyperparameter-tuning)**

Tables 6.5 and 6.8 present the metrics for the different acquired dictionaries (RAKE-instance-for-hyperparameter-tuning and RAKE-corpus), in order to showcase the impact the dictionaries have on the same model (VLSTM+). Such analyses, along with the manual evaluation of examples of the possible explanations, are suitable and has been used for deciding the type of dictionary that can be further used for the interpretability task.

When the dictionary acquisition method is chosen, preserving the dictionary among experiments for different classifiers is required. We, thus, report in Tables 6.6 and 6.7 the generalization power of the models eVLSTM and eVLSTM+ for the RAKE-instance dictionary. The MLP architecture of the explanation generator proves to be able to capture the semantic features of the pre-trained text classifier, obtaining positive contribution values for approximately 62% of the test instances with the eVLSTM+ model (Table 6.7).

Metric	ACC	AIC	All
Mean	0.0022	-0.0062	0.0010
Min	-0.3138	-0.5004	-0.5004
Max	0.4281	0.3702	0.2810
Std. dev.	0.0292	0.0549	0.0342
# of pos. C	11847	1819	13666
Total	21303	3525	24828
% of pos. C	55.61	51.60	55.04

Table 6.6: Individual contribution values (C) for: ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM - Rake-instance**

Metric	ACC	AIC	All
Mean	0.0021	-0.0091	0.0007
Min	-0.3949	-0.5682	-0.5682
Max	0.5549	0.4249	0.5549
Std. dev.	0.0328	0.0730	0.0400
# of pos. C	13958	1324	15282
Total	21794	3034	24828
% of pos. C	64.05	43.64	61.55

Table 6.7: Individual contribution values (C) for: ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM+ - Rake-instance**

Metric	ACC	AIC	All
Mean	0.0017	-0.0044	0.0010
Min	-0.3087	-0.3779	-0.3779
Max	0.3809	0.3313	0.3909
Std. dev.	0.0249	0.0527	0.0298
# of pos. C	13795	1278	15073
Total	21782	3046	24842
% of pos. C	63.33	41.96	60.71
# of changed predictions	1521	1112	2633

Table 6.8: Individual contribution values (C) for: ACC – All Correctly Classified Instances, AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM+ (RAKE-corpus)**

6.2 Qualitative Analysis

While the quantitative analysis focuses on reporting metrics related to both the classification task (reported in the previous chapter with the goal of preserving the original classifier's prediction rate) and the impact of the generated explanations on the prediction confidence, with the qualitative analysis we are taking a closer look to examples of explanations for the classified reviews. While the interpretability task is intended for humans' guidance for understanding how the classifier makes decisions, and there is no ground truth for the generated explanations, the human evaluation is required. We, thus, select a sample of 100 reviews and employ human evaluation for two tasks, reporting and interpreting the results. Furthermore, we aim to obtain model interpretability through an analysis of the most frequently generated explanations, and instance interpretability by analyzing a few examples of explanations.

6.2.1 Human Evaluation

The generated explanations can be used not only for individual prediction analysis and model interpretability but also for model selection. The following subsections describe the model comparison and model interpretability evaluation methods that we performed using two human annotators (one of the authors of the current research work and another Computer Science Master student with a Machine Learning background).

6.2.1.1 Model Interpretability

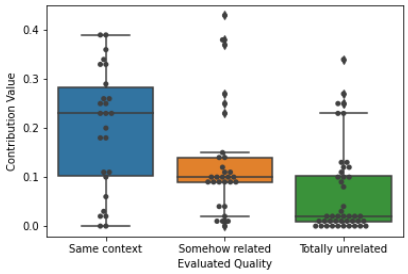
Sample selection. For this human evaluation task, we selected 100 sample reviews: the top 50 highest contributions for which the models have different predictions and explanations and 50 highest contribution explanations for which the models make the same prediction. These reviews are shuffled in order to ensure that the labeling is unbiased.

Experiment. We show to the two human evaluators the text of the review, together with the explanations and corresponding values of the contributions for two different classifiers.

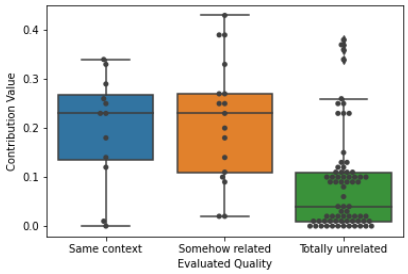
The annotators are asked to predict what each of the models' output, given that the presented explanation is improving the confidence of the classifier towards the correct prediction, and to classify the explanation according to its relatedness to the context of the review as "Same context", "Somehow related", "Totally unrelated".

Results interpretation. Having predictive power over the behavior of the classifier is one of the interpretability indicators of an approach. The ability of the human evaluators to predict which one out of the two models correctly classifies the review in question translates into a high faithfulness of the generated explanations to the models. This aspect is also supported by the high contribution values that ensure that the original classifier is highly influenced by the semantic information of the justification. The proposed contribution metric allows one to automatically evaluate the performance of the explanation generator model in terms of the quality of the generated explanations and their faithfulness to the model, the human evaluation being a confirmation of it. For both the models, the results of the human evaluation is in agreement with the quantitative evaluation discussed in Section 6.1.2. For the VLSTM model, 86% of the human predictions matched the true label, but only 49% matched the model's prediction, whereas the explanations for the VLSTM+ model led to 81% human predictions matching the true label and 74% the model's output. The agreement percentage between the two annotators for the VLSTM model is 82%, with a Cohen's kappa score of 0.3539 (fair agreement), while for the VLSTM+ model it is 89%, with a Cohen's kappa score of 0.6746 (substantial agreement). The values of interest are the percentages of the human predictions that overlap with the model's predictions, which is higher for the VLSTM+ model.

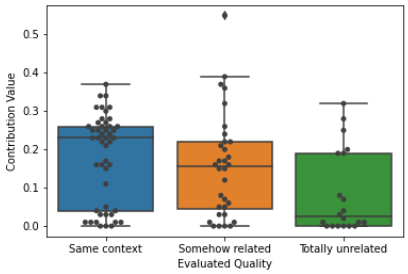
Relatedness of the explanations. The results of the relatedness of the explanations to the explained review are depicted through the boxplots in Figure 6.1, that show the contribution values corresponding to each of the annotated review, according to the quality category chosen by the human annotators, ("Same context", "Somehow related", "Totally unrelated") and the corresponding value counts. Since there are no gold truth labels for the generated explanations, and there is no formalization for the quality of an explanation,



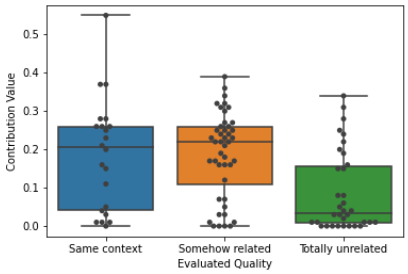
(a) Annotator 1: VLSTM.
Value counts: 26, 30, 44.



(b) Annotator 2: VLSTM.
Value counts: 12, 17, 71.



(c) Annotator 1: VLSTM+.
Value counts: 48, 32, 20.



(d) Annotator 2: VLSTM+.
Value counts: 22, 44, 34.

Figure 6.1: Relatedness of the explanation to the review's context and the corresponding distribution of the contribution values, according to the two human annotators

the contribution values help one correlate the influence an explanation has on the model's prediction and its relatedness to the context of the review. We can observe that for the VLSTM model, there are explanations with high contribution values (highly influential) that are not related to the context of the review (Figure 6.1b), while for the VLSTM+ model, the highly influential explanations are either in the "Same context" category or "Somehow related" according to both annotators (Figure 6.1c, 6.1d). This is an evidence that the VLSTM model is highly influenced by phrases that are not related to the context of the classified instance, which means that it makes predictions without identifying the semantic meaning conveyed by the author of the review, while the VLSTM+ makes more informed decisions. We also measure the agreement between the two annotators using Cohen's kappa coefficient, obtaining a score of 0.1856 for the VLSTM model and of 0.1978

for the VLSTM+ model. These scores translate into a slight agreement.

Further analysis. To further analyze the results of the explainers, we show in Table 6.9 examples of the most frequent explanations chosen by the explanation generators for the predictions of the two evaluated models (eVLSMT and eVLSTM+), using the test set. A complete list of the explanations and their frequencies can be found in Appendix C.

The most frequently chosen explanations reveal insights about both the classifier and the capacity of the MLP explanation generator to choose appropriate justifications (through the frequency distribution of the dictionary's entries). We note that the distribution of the explanations on the test set for the VLSTM+ model is closer to a uniform distribution, which is an expected result due to the high diversity of the topics, while for the VLSTM we observe a more skewed distribution of frequencies that is closer to a Pareto Distribution, the highest frequency explanation being a very general phrase that is less insightful than other phrases of the dictionary.

	VLSTM	VLSTM+
Most **frequent** explanations	wide white ready grin	bargain basement production values
	respectable affluent rich america	offered depression era audiences
	leslie still look great	deliberately shoot people dead
	kung fu u scenes	one real bad movie
	king give impressive performances	movie horribly overlook due
	prestigious hollywood award nomination	one particularly terrifying scene
	pulitzer prize winning play	really great indie films
	average predictable romantic comedy	former award winning actresses

Table 6.9: Examples of the positive-contribution explanations for VLSTM and VLSTM+

Most influential explanations. We also list the breakdown for each model according to the class prediction in Tables 6.10 and 6.11. We call those explanations the most influential ones for each of the reviews' categories since they are the most frequent explanations in

the top 100 highest contributions. This closer look at the explanations helps one grasp a better understanding of the most common topics of the positive and negative instances, depending on the prediction of the models.

		True label	
		Positive	**Negative**
Prediction	**Positive**	respectable affluent rich america wide white ready grin leslie still look great definitely worth seeing despite raging alcoholic matt sorenson great comedy writing team great curtis mayfield wrote chance like john gulager	respectable affluent rich america wide white ready grin leslie still look great best women character actresses us wars recently made fascinating yet unsettling look kung fu u scenes excellent carter burwell score
	Negative	wide white ready grin respectable affluent rich america leslie still look great kung fu u scenes	wide white ready grin respectable affluent rich america leslie still look great beloved younger sister nettie absolutely fabulous comedy series delicately humorous moments mingled

Table 6.10: Most influential explanations according to the prediction and true label for eVLSTM (most frequent explanations from the top 100 phrases with the highest contribution values)

Comparative analysis. The skewness of the frequency distribution of the explanations generated by eVLSTM is visible through the breakdown analysis presented in Table 6.10, too. Since the hyperparameter tuning of the explainer plateaued with the performance of eVLSTM that these reports are presented for, we can claim that the VLSTM model is not able to correctly capture the semantic meaning of the classified reviews compared to the

VLSTM+ model, even though their performance in terms of accuracy score is similar.

True positives. The aforementioned conclusion can be drawn from the true positive explanations, where the most influential ones for VLSTM contain neutral or not very insightful justifications (such as "raging alcoholic matt sorenson"), while for the larger model (VLSTM+), all the explanations convey a positive sentiment with insightful aspects closer related to the common movie review topics (for example "fantastic roving camera angles").

		True label	
		Positive	**Negative**
Prediction	**Positive**	excellent movies ever produced outstanding opening pan shot fun movie without preaching time absolutely savage denunciation fantastic roving camera angles beautiful fraternal love story true john woo masterpiece positively roast every scene fun holiday family movie great gordon harker provides excellent motion picture archives funniest sequences takes place highly motivated probation officer true romantic comedy classics	terry anderson deserve gratitude fine actor robert hardy war sixty years ago greatest movies ever made weird misguided masochistic belief simple slow love story think italian horror cinema amazing career topping performances fascinating yet unsettling look excellent movies ever produced perfect english directly back stagecoaches seem thrillingly alive really good job playing perfectly syched moog background

Prediction	**Negative**	glitzy hollywood plastic fantastic decent indie horror films sending apparent death threats really well done honestly talented visual directors working painfully little character development horrible piano music crescendoing stagecoaches seem thrillingly alive	reasonably brutal rape scene bad sheena easton song poor copy steals lots chance like john gulager bad wig till kidnapped painfully obvious horse costume bad songs every minute often disturbing morality tale worst cinematic con trick painfully little character development glitzy hollywood plastic fantastic worst movie ever made ultimately sympathetic police investigator making particularly bad films murdered sister every time bad things naturally occur

Table 6.11: Most influential explanations according to the prediction and true label for eVLSTM+ (most frequent explanations from the top 100 phrases with the highest contribution values)

True Negatives. The positive-sentiment explanation generated for the correctly classified negative instances of the eVLSTM model ("absolutely fabulous comedy series") with a contribution value of 0.10 translates into the model being influenced by a positive sentiment phrase to make a negative prediction. Thus, VLSTM does not correctly understand the meaning of the review that was explained through this phrase. By further analyzing the movie review explained with this phrase, one can draw conclusions about the out of vocabulary words present in the review, considering the topic of the review, writing style,

or used language a semantic outlier compared to the other training instances. We explore and explain this example in Section 6.2.2. The absence of positive explanations for the VLSTM+ model confirms once more that this model is able to generalize for the negative predictions, and for the predictions that were originally predicted positive, the negative phrases that influence the correct prediction manage to have a high enough contribution value in order to improve the confidence of the model.

False positives. The positive sentiment of the explanations for both evaluated models convey the fact that some of the negative reviews tend to comprise appreciations of the movies, even though the final rating or conveyed sentiment is a negative one. Both models seem to struggle to correctly identify those scenarios, being mislead by the positive context that has a high influence over the final prediction.

False negatives. Analyzing the negative-sentiment explanations for the false negative predictions of the VLSTM model, we can reason that the classifier finds it difficult to correctly classify instances that contain negative details. Explanations such as "worst movie ever made" (0.36) or "terribly gone wrong right" (0.21), with a high contribution value that can alter an initially positive prediction indicate that the classifier is highly influenced by negative contexts. Furthermore, even when the contribution value is low (having lower chances of altering a prediction), the classifier is still making incorrect predictions. The presence of negative-sentiment phrases for the VLSTM+ model (for example: "horrible piano music crescendoing") reveals the fact that the testing set contains positive reviews that also describe negative aspects. The incapacity of both models to correctly classify such instances can translate into this category of reviews being underrepresented in the training set. This claim of having a **semantically imbalanced training set** can be confirmed by analyzing, in a similar manner, the explanations generated for the training instances.

6.2.1.2 Interpretability-based Model Selection

Comparing the performance of two different classifiers and making an informed decision about which one would best perform on new real data is challenging when only classification metrics such as accuracy are reported since there is no additional information about what the model has learned. The ability to predict how the model would perform when only the review and multiple explanations are presented to the user is insightful and helpful for predicting how the model would work on new data points. We propose an experiment for comparing the models eVLSTM and eVLSTM+ that obtain a similar classification rate on the sentiment analysis.

Experimental setting. We use a sample of classified movie reviews in order to compare the performance of the two models (eVLSTM and eVLSTM+). This method represents an alternative to the classical approach of assessing models' generalization power through classification metrics on unseen datasets that rely on the prediction rate without any assumptions on how the model works or what it has learned from the training data. We claim that once humans are shown the explanations for each prediction of the compared models, they are able to determine what knowledge (semantic features) the classifier uses to make new predictions. When the used knowledge (explanation) is totally unrelated to the review under classification or when it conveys the opposite sentiment (compared to the gold truth), then the model is influenced by the wrong semantic concepts, thus it did not manage to learn and "understand" the context of the review. Further assumptions can be made based on a pair of such reviews and explanations, depending on the context.

Sample selection. Out of all the explanations with positive contribution values sorted in descending order, we select the top 50 for which the two models have different predictions and different explanations to be able to compare the models. For each of the two models, we selected 25 highest contribution values. This choice favours one of the models since the contribution value of the explanation for the other model can have a low value that does not allow one to make an informed decision. Since we are working with binary classification,

one of the two models in this sample will incorrectly classify the movie review, thus allowing us to also automatically determine when the annotator chooses the correctly classifying model.

Experiment. For this analysis, we present a movie review, the generated explanations for two classifiers, along with their corresponding contribution values and we ask the evaluators to determine the relatedness of the explanation to the context of the review, to predict what the model has outputted, and to say which model performed better, based on the explanation that the model considers as useful information for the current prediction. The annotators also have the option of not choosing any of the models when the explanations are not insightful enough to make an informed decision.

Results. Out of the 50 reviews with different predictions for the two considered classifiers, 13 reviews (26%) had less informative justifications for both the models, the human annotators being unable to make a decision. For the remaining reviews, the reviewers correctly chose the better performing model 75.68% of the time. The agreement between the two annotators is of 74%.

Error analysis. Out of the 22 instances that the annotators were not able to correctly identify the correctly classifying model, only 9 were incorrectly chosen (the annotator chose model 1 instead of 2 or conversely), the rest being the category that the annotators did not make a decision at all. In all those 22 instances, the contribution value of at least one of the two explanations is very low (< 0.1), making it difficult to make an informed decision. For the incorrect choices, even though one of the explanations had a high enough contribution value, the annotators classified the explanation as being unrelated to the context of the review. The other annotator was not able to make a decision for only 10 instances.

Conclusions. This experiment is proposed as an interpretability aided method for model selection. We employ human judgment to assess the relatedness of the explanations with high contribution values to the review's context, along with their prediction on which model

would better perform on new data. We also compare the agreement between the accuracy-based evaluation and the interpretability-based human evaluation and report our findings. While the majority of times, the human-evaluation is in agreement with the accuracy-based comparison between the models, it is worth noting that one of the challenges faced by this experiment is to acquire a sample of reviews for which both models have high-contribution explanations that are also related to the context of the review.

6.2.2 Semantic Outlier Detection Through Explanation Exploration

Table 6.12 showcases a few examples of explanations for each prediction category (TP, FP, TN, FN) for the eVLSTM+ model. The table contains the original review texts – some of them being reduced due to the high length, the explanations with their contributions, the prediction of the original model (VLSTM and VLSTM+) and of the enriched models (eVLSTM and eVLSTM+), and a short comment that interprets the results, especially for the eVLSTM+ model.

The examples focus on showcasing the explanations for the VLSTM+ model, also revealing the less interpretable justifications generated for the same review for the VLSTM model (Example review 3). While both models manage to generate high contribution explanations, they also seem to be complementing each other: for a review that the justification that has a high quality for one model is less informative and less interpretable for the other.

This comparison allows us to conclude that while VLSTM does not generalize well enough on the test set compared to VLSTM+, the latter manages to distinguish between the two classes and to identify more granular concepts and aspects about the context of the movie review. This assumption is also validated by the prediction rate of the two models, reported in Table 5.5.

<table>
<tr><th>Review ID</th><th>VLSTM explanation and contribution</th><th>VLSTM prediction & Raw prediction</th><th>VLSTM+ explanation and contribution</th><th>VLSTM+ prediction & Raw prediction</th><th>Gold label</th></tr>
<tr><td>1</td><td colspan="5">“ ”I am ... proud of ’Head’,” Mike Nesmith has said. He should be, because this film, which either has been derided by many of us or studied and scrutinized by film professors, works on many levels.
Yes, it’s unconventional. To many, frustrating. It’s almost as if the producers hand you the film and tempt: ”You figure it out.”
[...]
My guess is that ”Head” is the culmination of motivations somewhere between intended and unintended.
Largely, the insiders responsible for ”Head” seem to enjoy themselves in the revelries that take place in the film, but there is anger - anger at the chaos that characterized the late ’60s and anger at the way the media, television especially, had changed culture in negative ways. Drugs and violence were strong negative forces in the late ’60s and still are, but the producers of ”Head” want you to know that poor ”information” is a far greater danger.
[...]
”Head” is either a movie that creates itself ”as we go along”, or is a deliberate statement. Perhaps, perhaps not. [...]
Cheers: A true guilty pleasure. Very funny. Intelligent. Will please the fans. Find the substance, it’s there. Unabashedly weird. Bizarre collection of characters. Good tunage. Length is appropriate. Lots of great one liners, including my all time prophetic favorite: ”The tragedy of your times, my young friends, is that you may get exactly what you want.”
Caveats: Dated. Drugs. No plot. No linear delivery of any thought in particular. At least twenty-five stories that interweave in stop-and- go fashion. So, may easily frustrate. May seem pretentious to some. People who can’t stand The Monkees need not watch, though that in itself is no reason to avoid it. The psychedelic special effects may kill your ailing picture tube or your acid burnt- out eyeballs.
Match, cut. ”</td></tr>
<tr><td></td><td>excellent movies ever produced (0.08)</td><td>positive(0.99)</td><td>excellent movies ever produced (0.10)</td><td>positive (0.47)</td><td>positive</td></tr>
<tr><td></td><td colspan="5">Interpretation:
This is an example of a changed prediction, where the model VLSTM+ was initially making a wrong prediction (negative - 0.47) about the polarity of the review due to the lengthy description filled with negative phrases, but in the presence of the positive explanation, the same model manages to improve its confidence towards the correct prediction.</td></tr>
<tr><td>2</td><td colspan="5">“Ruggero Deodato is often credited for inventing the cannibal subgenre with JUNGLE HOLOCAUST in 1975. But director Umberto Lenzi, usually acknowledged as a Deodato rip-off, directed THE MAN FROM DEEP RIVER 3 years earlier in 1972. Is it a worthy start for the genre? Well....not really.....
A photographer accidentally kills a man in self-defense and while retreating into the jungles of an Asian country, is captured by a native tribe who hold him captive, force him into slave labor, and eventually accept him when he marries the chief’s daughter. Throughout the whole film, I never felt this was a horror film. It was more reminiscent of a drama, like A MAN CALLED HORSE, which I liked better. Ivan Rassimov is pretty good as the photographer, but it is Me Me Lai as the chief’s daughter who is memorable and great. I have always been a Me Me Lai fan ever since her breathtaking performance in JUNGLE HOLOCAUST and she is never given credit for her acting chops because she hardly speaks in her films. She is still very talented and charming. Lots of real animal mutilation is the one thing about DEEP RIVER that could make it a horror film, but even that doesn’t execute well.
THE MAN FROM DEEP RIVER is good to see for those who want to see what started the cannibal subgenre, but as an entry in the genre, is easily eclipsed by Deodato’s entries and even Lenzi’s own later entries. Recommended only for completists and Me Me Lai fans.”</td></tr>
</table>

	absolutely fabulous comedy series (0.10)	negative (0.11)	absolutely fabulous comedy series (-0.001)	positive(0.97)	negative
	Interpretation: While the VLSTM model correctly classifies the negative review, the model is influenced by a positive-sentiment explanation, which means that the model learned that a review with positive phrases can be classified as negative. On the other hand, the VLSTM+ model is confident that the review is positive, but is slightly negatively influenced by the positive explanation, which means that the model learned that reviews containing such positive phrases cannot be classified as negative instances.				
3	"Although I found the acting excellent, and the cinematography beautiful, I was **extremely disappointed** with the adaptation. [...] The character changes in Mattie and Zenna are almost non-existent. While in the novella they almost change places, at the end of this adaptation it appears as if they are both invalids. Lastly that Mattie and Ethan consummate their relationship fully nearly destroys the power and poignancy of the finale. The change of the narrator being a preacher was one effective change. Neeson and Arquette are superb in their portrayals. Joan Allen was also wonderful, however her character was much watered down from Whartons novella. I do not expect films to faithfully portray novels, but this one went to far and in the process nearly **destroyed the story**. Overall, I **would not recommend watching** this film unless you have read the book as you will come away confused and disappointed."				
	inventively gory murder sequences (-0.0007)	negative (0.03)	talented visual directors working (0.10)	negative (0.32)	negative
	Interpretation: Although the VLSTM+ model is influenced by the positive aspects mentioned in the review, it manages to preserve the correct negative prediction due to the sufficiently low value of the contribution.				
4	"This is a great film for pure entertainment, nothing more and nothing less. It's **enjoyable**, and a **vaguely feel-good movie**. A minor, but nonetheless irritating thing about the movie is that we don't know why Justine and Chas broke up. Okay, most first relationships don't work for one reason or another, but they more or less seemed like a **nice couple**. In a nutshell, it's worth a watch to escape reality."				
	simple slow love story (0.04)	positive (0.8)	simple slow love story (0.12)	negative (0.2)	positive
	Interpretation: This is an example where the model is influenced by a phrase that is contextually related to the review and has the same polarity, but the contribution value is not high enough to improve the prediction. We can, thus, deduce that the training set is underrepresented semantically by positive reviews that contain negative contexts.				

Table 6.12: Insightful examples from the test set, along with their explanations. The underlined phrases express the opposite sentiment compared to the gold label, while the bold text expresses the same sentiment as the gold truth.

6.3 Chapter Summary

In this chapter, we presented the experimental setting and the employed evaluation methods for the two selected models (eVLSTM and eVLSTM+) for the interpretability task that

has no gold truth labels. Thus, we rely on the confidence score that assesses the fidelity of the explanation to the original classifier, based on the idea that an explanation with a high value of the explanation should have a high influence on the final prediction of the classification task. We also report the results of the quantitative and qualitative analysis. The quantitative, automatic evaluation is based on the polarity of the phrases and of the contribution scores, while the qualitative analysis relies on human judgment for predicting the output of the classifier given the explanation, for semantic model comparison, and for semantic outlier detection based on the most influential explanations of the testing dataset. The next chapter summarizes our contributions that answer the four research questions addressed in the current work, highlighting the challenges that we tackled, and proposing future research avenues.

Chapter 7

Conclusion and Future Work

This chapter summarizes our findings, highlighting the challenges that we have encountered, the main contributions of our research work, and ideas for improving and extending the proposed framework.

7.1 Summary of Contributions

This research thesis comprises an extensive literature review of the emerging XAI field and proposes an unsupervised generative method for generating justifications for a biLSTM text classifier. Inspired by the human behavior where the decisions are motivated by past experiences, the present methodology is a neural prediction approach that identifies the most influential features of the training set for individual predictions. This solution has the advantages that it does not require human annotations, like the explainability tasks do, and it is informative for both interpreting individual predictions and model interpretability.

The contributions of this research work mentioned below address the research questions mentioned in Section 1.4 of the Introduction chapter and consist of a number of novel approaches applied for improving the interpretability of deep learning text classifiers.

1. Proposed a novel neural prediction approach for identifying the most influential phrases of a training set in an unsupervised manner through two desiderata: of

improving the confidence of individual predictions and of not loosing the explained text classifier's performance. (RQ1)

2. Performed experiments for model architecture selection, hyperparameter tuning, and determining the appropriate dictionary acquisition method for the set of possible explanations such that the original classification performance is preserved. (RQ2)

3. Proposed a novel automatic evaluation method for measuring the faithfulness of the generated justifications to the model – the contribution metric – that is also supported by the results of the human evaluation. (RQ3)

4. Described numerous analysis methods for the interpretability task, suitable for semantic outlier detection, semantic model selection, and model interpretability. (RQ4)

5. Proposed an end-to-end framework that does not require labeled data for the interpretability task and that contains independent components that can be adapted to alternative (NLP) text classification tasks (other than sentiment analysis). (RQ1)

7.2 Challenges

During our research, we have tackled a number of challenges that we highlight in the following subsections in order to summarize our findings and the architectural decisions that we have made, proposing solutions that are backed up by extensive experiments.

7.2.1 Dictionary Acquisiton

In an ideal setting, the implementation of the proposed methodology would use the words and phrases of the entire training set as potential candidates for the explanation generation model or a set of predefined insightful justification. Our experiments show that the choice of the dictionary has an impact not only on individual predictions' contribution values, but also on the overall performance of the model. One of the challenges that we faced was determining the most informative dictionary acquisition method that manages to extract

the explanation candidates in an unsupervised manner, while also preserving the coherence of the phrases for a high interpretability level. While keywords and short phrases are more coherent, they do not convey as much information as multi-word phrases.

When performing experiments and analyzing the impact of different keyphrase extraction methods on the performance of the explanation generator, hyperparameter tuning is also required, thus making the fair comparison between the dictionaries becomes challenging. In our experiments, we have preserved one dictionary and evaluated different models in order to assess the generalization power and effectiveness of the proposed framework, but we also emphasize the fact that replacing the dictionary acquisition method and improving the informativeness of the phrases also leads to an improved level of interpretability.

7.2.2 Hyperparameter Tunning

The high number of variables included in the proposed framework (dictionary acquisition method, classifier's hyperparameters, explanation generator's architecture, and hyperparameters) represented a challenge in structuring, organizing, and defining the experiments that are informative for the end task of interpretability. While it is computationally expensive and probably infeasible to perform an exhaustive search for all those variables, making informed decisions for each of the modules of the framework can lead to insightful results. Furthermore, having independent components that can be easily adaptable to different tasks and domains, not only to sentiment analysis, represents another advantage of the proposed framework.

7.2.3 Model Evaluation

Evaluating the interpretability level, the quality of the generated explanations, and the impact that an explanation generation model can have are challenging topics for the XAI domain due to an incompleteness of the problem formalization. While there are no established evaluation methods universally accepted by the research community, we try to assess the usefulness of our approach through both quantitative analysis – by defining measur-

able desiderata – and qualitative analysis – by exploring the generated explanations for interpretability tasks. Our findings for the qualitative analysis highlight that, aside from the individual prediction's interpretability, the generated explanations can also be informative for model interpretability – allowing the user to answer the question "what would the model predict for a particular text instance?" –, model selection, and semantic outlier detection.

7.3 Future Work

For improving the results that we report in this research work, we propose alternative methods for the aforementioned challenges.

For the dictionary acquisition method, experimenting with alternative text mining approaches suitable for the type of text that is being classified (informal style of social media text, academic writing, articles) could lead to improved interpretability results.

In terms of the model architecture of the generation model, a Natural Language Generation approach could improve the qualitative results, leading to new insights about the model. While this is a compelling approach, it comes with a high increase of complexity which can also be reflected in longer training times. Moreover, one can also experiment with alternative architectures for generating multiple explanations for the same classified instance.

The human evaluation method could also benefit from improvements such as increasing the sample size of the human-annotated test instances, measuring the agreement between multiple annotators, and elaborating multiple experimental settings for different tasks such as assessing the informativeness of the explanations.

Another research direction considered as future work is applying this framework on text classification for different languages.

References

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1823–1832, New York, NY, USA. Association for Computing Machinery.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks.

Diego Antognini, Claudiu Musat, and Boi Faltings. 2019. Multi-dimensional explanation of ratings from reviews.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. 2020. Agent57: Outperforming the atari human benchmark. *arXiv preprint arXiv:2003.13350*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Shane Barratt. 2017. Interpnet: Neural introspection for interpretable deep learning. *arXiv preprint arXiv:1710.09511*.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.

Pedro Carpena, Pedro Bernaola-Galvan, Michael Hackenberg, Ana Victoria Coronado Jiménez, and Jose Oliver. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79:035102.

Hanjie Chen and Yangfeng Ji. 2019a. Improving the explainability of neural sentiment classifiers via data augmentation.

Hanjie Chen and Yangfeng Ji. 2019b. Improving the explainability of neural sentiment classifiers via data augmentation.

Hanjie Chen and Yangfeng Ji. 2019c. Improving the explainability of neural sentiment classifiers via data augmentation.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 883–892, Stockholmsmässan, Stockholm Sweden. PMLR.

R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao. 2018. Explainable text classification in legal document review a case study of explainable predictive coding. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1905–1911.

Mark W. Craven and Jude W. Shavlik. 1995. Extracting tree-structured representations of trained networks. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, page 24–30, Cambridge, MA, USA. MIT Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2454–2463, Long Beach, California, USA. PMLR.

Juan Herrera and Pedro Pury. 2008. Statistical keyword detection in literary corpora. *The European Physical Journal B - Condensed Matter and Complex Systems*, 63:135–146.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Xinghua Hu and Bin Wu. 2006. Automatic keyword extraction using linguistic features. pages 19–23.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kyoung-Rok Jang, Sung-Hyon Myaeng, and Sang-Bum Kim. 2018. Interpretable word embedding contextualization. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 341–343, Brussels, Belgium. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Naveenkumar Jayakumar. 2012. Keyword extraction through applying rules of association

and threshold values. *International Journal of Advanced Research in Computer and Communication Engineering*, 1:295–297.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Jasmeen Kaur and Vishal Gupta. 2010. Effective approaches for extraction of keywords. *International Journal of Computer Science Issues*, 7.

Richard Khoury, Fakhri Karray, and Mohamed S. Kamel. 2008. Keyword extraction rules based on a part-of-speech hierarchy. *IJAMC*, 2:138–153.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2280–2288. Curran Associates, Inc.

Aykut Koç, Ihsan Utlu, Lutfi Kerem Senel, and Haldun M. Özaktas. 2018. Imparting interpretability to word embeddings. *CoRR*, abs/1807.07279.

Deepthi Kuttichira, Sunil Gupta, Cheng Li, Santu Rana, and Svetha Venkatesh. 2019. *Explaining Black-Box Models Using Interpretable Surrogates*, pages 3–15.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *CoRR*, abs/1707.01154.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jun LI, Guimin HUANG, Chunli FAN, Zhenglin SUN, and Hongtao ZHU. 2019. Key word extraction for short text via word2vec, doc2vec, and textrank. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, 27:1794–1805.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

X. Liu, X. Wang, and S. Matwin. 2018. Improving the interpretability of deep neural networks with knowledge distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 905–912.

Xiaoxuan Liu, Livia Faes, Aditya Kale, Siegfried Wagner, Dun Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph Ledsam, MD Schmid, Konstantinos Balaskas, Eric Topol, Lucas Bachmann, Pearse Keane, and Alastair Denniston. 2019b. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural*

Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. 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. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Warren S. McCulloch and Walter Pitts. 1988. *A Logical Calculus of the Ideas Immanent in Nervous Activity*, page 15–27. MIT Press, Cambridge, MA, USA.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica*, 22:276–82.

R. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Thomas M. Mitchell. 1997. *Machine Learning*, 1 edition. McGraw-Hill, Inc., USA.

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

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*. Https://distill.pub/2018/building-blocks.

M. Ortuño, Pedro Carpena, Pedro Bernaola-Galvan, E. Muñoz, and Andrés Gimeno. 2007. Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, 57:759.

Dong Park, Lisa Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. pages 8779–8788.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4782–4793. Association for Computational Linguistics.

Rajib Rana. 2016. Gated recurrent unit (gru) for emotion classification from noisy speech.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20.

Gerard Salton, C. S. Yang, and C. T. Yu. 1975. A theory of term importance in automatic text analysis. Technical report.

Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. *Lecture Notes in Computer Science*, page 5–22.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951. Association for Computational Linguistics.

Mehrnoush Shamsfard and Ahmad Abdollahzadeh Barforoush. 2003. The state of the art in ontology learning: A framework for comparison. *Knowl. Eng. Rev.*, 18(4):293–316.

Yan Shi. 2016, accessed August 3, 2020. *Understanding LSTM and its diagrams*. https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: A literature review.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Koustuv Sinha, Yue Dong, Jackie Chi Kit Cheung, and Derek Ruths. 2018. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823, Brussels, Belgium. Association for Computational Linguistics.

A. M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Gabriel Vásquez Morales, Sergio Mauricio Martínez Monterrubio, Pablo Moreno Ger, and Juan Recio-García. 2019. Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning. *IEEE Access*, PP:1–1.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu. 2019. Sentiment analysis of comment texts based on bilstm. *IEEE Access*, 7:51522–51532.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Hongyang Yu, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Qi Tian, and Nicu Sebe. 2020. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, 128(5):1141–1159.

Xiaolei Zuo, Silan Zhang, and Jingbo Xia. 2017. The enhancement of textrank algorithm by using word2vec and its application on topic extraction. *Journal of Physics: Conference Series*, 887:012028.

Appendix A

Softmax-Gumbel Experiments

Model name	Dict	Accuracy	Average F1
Vanilla	-	**85.77**	0.7601
gen(MLP)_softmax	Rake-1 (inst)	84.05	0.7351
	Rake-1 (corpus)	**86.74**	**0.7706**
	Rake-2 (inst)	84.36	0.7518
	Rake-2 (corpus)	83.77	0.7402
	Rake-3 (inst)	86.17	0.7634
	Rake-3 (corpus)	83.57	0.7364
	Rake-5 (inst)	85.72	0.7578
	Rake-5 (corpus)	86.52	0.7669
	TextRank	85.39	0.7560
	Yake	86.26	0.7634
	TF-IDF	85.48	0.7565
gen(MLP)_gumbel	Rake-1 (inst)	86.81	0.7698
	Rake-1 (corpus)	85.23	0.7563
	Rake-2 (inst)	87.02	**0.7754**
	Rake-2 (corpus)	86.21	0.7629
	Rake-3 (inst)	87.02	0.7754
	Rake-3 (corpus)	87.00	0.7718
	Rake-5 (inst)	86.04	0.7636
	Rake-5 (corpus)	85.45	0.7550
	TextRank	85.41	0.7548
	Yake	86.15	0.7606
	TF-IDF	**87.18**	0.7726

Table A.1: MLP joinlty train: Softmax vs. Gumbel

Appendix B

Quantitative Analysis

Metric	ACC	AIC	All
Mean	0.0023	-0.0005	0.0019
Min	-0.3608	-0.4606	-0.4606
Max	0.4730	0.2871	0.4730
Std. dev.	0.0287	0.0511	0.0331
# of pos. C	7584	2399	9983
Total	21045	3799	24844
% of pos. C	36.04	63.15	40.18

Table B.1: Individual contribution values (C) for: ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM - Rake-corpus**

Metric	ACC	AIC	All
Mean	0.0017	-0.0045	0.0009
Min	-0.3548	-0.426	-0.4267
Max	0.5377	0.3263	0.5377
Std. dev.	0.0241	0.0541	0.0296
# of pos. C	14329	1267	15596
Total	21770	3074	24844
% of pos. C	65.81	41.21	62.77

Table B.2: Individual contribution values (C) for: ACC – All Correctly Classified Instances,
AIC – All Incorrectly Classified Instances, and All Instances for **eVLSTM+ - Rake-corpus**

Appendix C

Qualitative Analysis

C.1 Most Frequent Explanations – eVLSTM

You can find below a list of the most frequent explanations, along with their frequency for the explained test predictions of the model eVLSTM.

1. wide white ready grin 4583
2. respectable affluent rich america 3932
3. leslie still look great 923
4. kung fu u scenes 214
5. beloved younger sister nettie 205
6. king give impressive performances 172
7. prestigious hollywood award nomination 153
8. illegal immigrant feels sick 129
9. sending apparent death threats 76
10. harshly lighted sets rob 76
11. us wars recently made 64
12. rape scenes might seem 59
13. subsequent cuban missile crisis 58
14. terminal stage lung cancer 53

15. subsequent post traumatic stress 48
16. still probably best left 40
17. racing action beautiful pictures 33
18. eventually several people die 33
19. physically impressive man capable 31
20. two nasty little details 31
21. less interesting stuff worthwhile 27
22. depression following almost ruined 24
23. movie horribly overlook due 23
24. inventively gory murder sequences 20
25. reasonably utal rape scene 20
26. special entertainment nine excerpts 20
27. gorgeous maruschka detmers takes 19
28. dick cheney mutters mostly 19
29. post traumatic stress syndrome 19
30. pulitzer prize winning play 19
31. stopped crying almost instantly 18
32. average predictable romantic comedy 18
33. creating commercially attractive films 18
34. newer maelstrom like attraction 17
35. extravagant climactic super ballet 17
36. murderous character called dominic 17
37. serial killer get caught 16
38. disney story gone horribly 16
39. lost valuable precious moments 16
40. intensely talented performer around 16
41. dashing police officer hero 16
42. instant commercial success may 16
43. anyone could easily enjoy 15
44. every successful director gives 15

45. still worked terrifically well 15
46. stagecoaches seem thrillingly alive 15
47. gorgeous techicolor production telling 15
48. special body part helping 14
49. kill forty five minutes 14
50. tiny laugh deep within 14
51. fellow italian horror buffs 14
52. dishonest crook follows suit 14
53. fascinating yet unsettling look 14
54. sexy esther muir plays 14
55. gorgeous looking isabelle adjani 13
56. flawless stock haryanvi accent 13
57. murderous rampage must end 13
58. first bad signs come 13
59. placed countless glorious animal 13
60. whole affair tremendously disappointing 13
61. intellectually honest person would 13
62. best women character actresses 13
63. smiled like three times 12
64. really wo nt enjoy 12
65. mentally retarded boy makes 12
66. also enjoyed two bouts 12
67. supporting cast lent credit 12
68. enraged samurai utally murders 12
69. yet another comic masterpiece 12
70. give special undue advantages 12
71. best movies ever made 12
72. mother superior would ask 12
73. really great indie films 12
74. whilst alive enjoyed staying 12

75. certainly something worth watching 12
76. callous villain without resorting 12
77. arcane imagery proudly displayed 12
78. indeed amazing filmic imagery 11
79. two stupid old men 11
80. popular spoofs like airplane 11
81. long slow takes encourage 11
82. picture looks terribly drab 11
83. mechanized armies stealing generators 10
84. standard ugly blonde chick 10
85. master manages happy endings 10
86. standard comic book villain 10
87. enjoys receiving root canals 10
88. wonderful job tricking people 10
89. beautifully filmed vampire flick 9
90. time absolutely savage denunciation 9
91. weird misguided masochistic belief 9
92. quite successful launching campaign 9
93. pulitzer prize winning journalist 9
94. good fun political satire 9
95. still absolutely horrid makes 9
96. sadists apparently running free 9
97. good gore scenes like 9
98. really good job playing 9
99. violent barbershop vocal duo 8
100. whole show offensively loud 8
101. man whose great grandmama 8
102. classic angry mob style 8
103. means embarrassingly bad like 8
104. weakest vehicle day found 8

105. sympathetic policewoman officer sudow 8
106. delightfully zany museum curator 8
107. especially loved ray psyching 8
108. successful snack foods company 8
109. extremely fascinating character study 8
110. sequence sounds really crude 8
111. violent young criminal visiting 8
112. absolutely pathetic action scenes 8
113. still thoroughly enjoy watching 8
114. successful beverly hills dentist 8
115. really fake italian accent 8
116. horror veteran george kennedy 8
117. thoroughly engaging romantic drama 8
118. gore effects got stupider 8
119. escaped killer nick grainger 8
120. free speech high horse 8
121. popular among prominent actors 8
122. young hero comes back 8
123. terry anderson deserve gratitude 8
124. chance like john gulager 7
125. typically fun hipster pretenses 7
126. best major studio films 7
127. personally liked russell better 7
128. doomed space craft orbiting 7
129. crime fighting android thing 7
130. modern audience certainly accepts 7
131. excellent facial expressions combine 7
132. beautiful attracts excellent idea 7
133. rather conventional romantic comedy 7
134. picture warrants special praise 7

135. incredibly bad process work 7
136. painfully obvious dim view 7
137. drug war etc etc 7
138. woman kills two couples 7
139. please hire convincing actors 7
140. intentionally stupid title song 7
141. watch racism slowly dissipate 7
142. regular love interest holds 7
143. still magnificent béatrice dalle 7
144. really bad buddy movie 7
145. nice sets help keep 7
146. old unspenseful horror films 7
147. hopelessly simplistic conflicts like 7
148. pretty original twist indicating 7
149. really well done honestly 7
150. pleasantly surprised upon rediscovering 7
151. illegal bare knuckle fight 7
152. horribly written havoc preceded 7
153. avuncular superior officer assures 7
154. felt like huge parts 6
155. either beautiful party girls 6
156. pretty painless ninety minutes 6
157. one honest movie goer 6
158. great location cinematography right 6
159. one particularly terrifying scene 6
160. racist guard avoids confrontation 6
161. lovely lilting itish accent 6
162. ugly loud hawaiian shirts 6
163. emmy award winning performance 6
164. ultra evil white slaver 6

165. successful dental hygienist witnesses 6
166. anger management problem due 6
167. know micheal elliott loves 6
168. gets rich married men 6
169. corrupt treasury official keen 6
170. think italian horror cinema 6
171. find super spy lovejoy 6
172. jackie gleason type villain 6
173. one low life loser 6
174. tom cruise made ugly 6
175. sympathetic grumpy shop owner 6
176. ultimately sympathetic police investigator 6
177. far better work like 6
178. labeled box office poison 6
179. soon softened gorgeous lady 6
180. cold war tension portrayed 6
181. great gordon harker provides 6
182. divine violet kemble cooper 6
183. hippy purist propaganda crap 6
184. severely decomposing male corpse 6
185. blind woman defeating hundreds 6
186. racist local cops raid 6
187. funniest opera ever put 6
188. darryl hannah avoid humiliation 6
189. one real bad movie 6
190. sometimes artistic liberties would 5
191. found something worth giving 5
192. masterpieces zu neuen ufern 5
193. happy happy joy joy 5
194. best films ever made 5

195. nazi hall monitors love 5
196. low budget horror movie 5
197. real life war machine 5
198. ugly deformed serial killer 5
199. anger towards mainly neal 5
200. terribly boring new age 5
201. excellent movies ever produced 5
202. always admired susan sarandon 5
203. mostly beautiful young women 5
204. great films get better 5
205. opera novel seem like 5
206. painfully obvious horse costume 5
207. beautiful wealthy salon owner 5
208. absolutely fabulous comedy series 5
209. admit creating great expectations 5
210. lovely susannah york provides 5
211. huge star trek fan 5
212. beautiful ritual seem vapid 5
213. really worth greatest attention 5
214. beautiful menzies art design 5
215. two uncomfortably prolonged assaults 5
216. absolutely amazing production number 5
217. already anxious ide insane 5
218. liberty mutual insurance commercial 5
219. serial killer sounds intriguing 5
220. lovely young alisan porter 5
221. best work angie dickinson 5
222. recent asian horror films 5
223. rapidly decaying nitrate negative 5
224. wonderful exciting story together 5

225. huge box office success 5
226. best cult comedy films 5
227. often subsequent night terrors 5
228. ugly little creature locked 5
229. many superb character actors 5
230. often disturbing morality tale 5
231. good many terrific shots 5
232. old vietnam war vet 5
233. horrifying shock therapy session 5
234. love survive song performed 5
235. abused next door neighbor 5
236. negative reviews mainly complained 5
237. graphical content extremely disturbing 5
238. female characters slutty idiots 5
239. gorgeous female actresses got 5
240. great comedy writing team 5
241. loathing teens often feel 5
242. us three hundred lovers 5
243. good mann flick thanks 5
244. remember one horrifying incident 5
245. raging alcoholic matt sorenson 5
246. ridiculously forced freudian slips 5
247. many horror fans may 4
248. heroine get happily married 4
249. war sixty years ago 4
250. appreciated paolo conte used 4
251. declared missing presumed killed 4
252. true jane austen fan 4
253. hurricane nonchalantly devastating everything 4
254. former award winning actresses 4

255. worst movies ever made 4
256. happy coincidence inadvertently saved 4
257. highly respected russian family 4
258. glorious technicolor cinematography leaps 4
259. horrible opening scene establishes 4
260. highly laughable special effects 4
261. delightfully named nettlebed becomes 4
262. gorgeously fixating patty shepard 4
263. hugely winning comedic team 4
264. pandering treacly love letter 4
265. stupid dialogue never matches 4
266. compelling true romance story 4
267. perfectly smug kyle maclachlan 4
268. huge jane austen fan 4
269. greatest motion picture trilogies 4
270. dick grayson alias nightwing 4
271. merry men must find 4
272. horror icons tom savini 4
273. decent indie horror films 4
274. terribly gone wrong right 4
275. also includes dvd bonuses 4
276. israeli independence war triggers 4
277. another great orchestral score 4
278. sexy charlotte lewis asks 4
279. play two slacker friends 4
280. bad rock music blares 4
281. horribly miscast dean cain 4
282. portray another major pain 4
283. courageous personnage played magnificently 4
284. earth destroying everything around 4

285. excellent carter burwell score 4
286. another pathetic chase sequence 4
287. lightweight dark comedy entertains 4
288. television horror movie hostess 4
289. talented visual directors working 4
290. main characters best friend 4
291. portrays incredibly cruel treatment 4
292. various wonderful period costumes 4
293. making particularly bad films 4
294. surprisingly good special effects 4
295. beautifully cheerful musical score 4
296. rob gets accidentally hypnotized 4
297. simple slow love story 4
298. kids become suicide terrorists 4
299. personal best films ever 4
300. grown ups hate horrors 4
301. true john woo masterpiece 4
302. best live action short 4
303. make us laugh afterwards 4
304. many people hate black 4
305. fun movie without preaching 4
306. fairy tale loving niece 4
307. defeat real live evil 4
308. protagonists get super powers 4
309. oh dear oh dear 4
310. offered depression era audiences 4
311. average ufo conspiracy theory 4
312. sick sad world thinks 4
313. positively roast every scene 4
314. great curtis mayfield wrote 4

315. dumb low budget porno 4
316. completely destroyed within moments 4
317. dead fleet street journalist 4
318. greatest movies ever made 4
319. talented director jack gold 4
320. glitzy hollywood plastic fantastic 4
321. new stupid stuff instead 4
322. basically birth till death 3
323. dead guy would chose 3
324. producer dick powell saw 3
325. towners ultimately fails miserably 3
326. fifty attackers never seem 3
327. bargain basement production values 3
328. fantastic digitally remastered transfer 3
329. highly motivated probation officer 3
330. many bad movies later 3
331. time please please please 3
332. drexler looked pretty good 3
333. sweet little god daughter 3
334. bad things naturally occur 3
335. stupid characters running around 3
336. hopeful suitor allan lane 3
337. true hollywood happy ending 3
338. outstanding opening pan shot 3
339. killers never loose track 3
340. war either required transplanting 3
341. hello dave burning paradise 3
342. american flag stands proudly 3
343. love chinese epic films 3
344. freeman aka morgan freeman 3

345. gorgeous fashion plate poses 3
346. rather sweet value system 3
347. good laugh although unintended 3
348. really stupid ending tacked 3
349. dumb lead actress thinks 3
350. stunningly beautiful art direction 3
351. bad sheena easton song 3
352. post traumatic stress disorder 3
353. nasty old hag neighbor 3
354. yummy blonde marisol santacruz 3
355. bafta anthony asquith award 3
356. really really really upset 3
357. cool glamorized sexual moments 3
358. seemingly petty criminals weigh 3
359. stereotyped bad guys waiting 3
360. prison escapee finds time 3
361. sinister cinema vhs copy 3
362. walks around looking stricken 3
363. simply wonderful little gem 3
364. know everyone makes fun 3
365. happy black natives dying 3
366. died two years later 3
367. always wonderful lili taylor 3
368. individual personalities creating clear 3
369. definitely worth seeing despite 3
370. unseen maniac slashing teens 3
371. almost acquire negative points 3
372. pretty well tell participants 3
373. greatest performance would ever 3
374. best entries like cleo 3

375. surprisingly effective gretchen moll 3
376. beautiful fraternal love story 3
377. enter world war ii 3
378. delightful summer entertainment hits 3
379. always marvelous holland taylor 3
380. beautiful autumnal ontario province 3
381. another easy romantic comedy 3
382. features three excellent performances 3
383. worst scripts ever written 3
384. painfully obvious stagebound sets 3
385. former lover terry anderson 3
386. devastatingly icy david strathairn 3
387. uniformly superb acting qualifies 3
388. huge adam sandler fan 3
389. done better special effects 3
390. snidely whiplash villain roy 3
391. stupid blond girl told 3
392. everyone lives happily ever 3
393. always enjoyed science fiction 3
394. nearly perfect digital copy 3
395. translation ha ha ha 3
396. slightly less poverty stricken 3
397. james bond type villain 3
398. actually quite scary seeing 3
399. love slapstick comedy shows 3
400. background music perfectly chosen 3
401. state taking illegal action 3
402. bizarre murder mystery plot 3
403. illegal dice game going 3
404. top five worst movies 2

405. others may feel betrayed 2
406. year old horribly ashamed 2
407. thank god abc picked 2
408. perfectly well rank among 2
409. amazing martial arts fantasy 2
410. trio get hopelessly lost 2
411. provide free ukulele picks 2
412. usual socialist economic miracle 2
413. nothing good comes easy 2
414. fake death ending coming 2
415. mean horrible beyond belief 2
416. deeply satisfying x file 2
417. policemen anally rape male 2
418. greatly appreciated second look 2
419. many talented persons involved 2
420. nice sizable supporting part 2
421. undergoing statutory rape charges 2
422. big budget blockbuster faire 2
423. delicately humorous moments mingled 2
424. really goofy love triangle 2
425. genuinely creepy horror flick 2
426. delightful little thriller opens 2
427. second favourite horror setting 2
428. picture looks incredibly handsome 2
429. perfectly syched moog background 2
430. entertainment value worth watching 2
431. basic love triangle story 2
432. murdered sister every time 2
433. worst movie ever made 2
434. poor copy steals lots 2

435. worst acting performance ever 2
436. great clive owen fans 2
437. poetic masterpiece told clearly 2
438. quickly paced murder mystery 2
439. sure razzie award winner 2
440. hardly say anything positive 2
441. late sixties definitely worth 2
442. love pad quite well 2
443. best supporting actor award 2
444. nice talented memorable oddity 2
445. great tonino delli colli 2
446. rocky horror picture show 2
447. slap stick comedy like 2
448. police guard gets killed 2
449. lovable pup ever shown 2
450. say enough bad things 2
451. worst editing ever combined 2
452. spent many happy hours 2
453. totally beautiful chick flick 2
454. god help micheal ironsides 2
455. terribly acted kids kill 2
456. painfully obvious twist ending 2
457. angry low budget filmmaker 2
458. events seem incredibly forced 2
459. another prisoner centuries ago 2
460. truly powerful social satire 2
461. fantastic roving camera angles 2
462. excellent motion picture archives 2
463. rich old men bidding 2
464. went back home happy 2

465. utterly terrible acting performances 2
466. wonderfully touching human drama 2
467. adventurous lively farraginous chronicle 2
468. religious civil war going 2
469. inning let alone lose 2
470. dreadful one missed call 2
471. dead race left switched 2
472. much higher entertainment value 2
473. heroic leading couple visits 2
474. perfect english directly back 2
475. worst columbo ever dreamt 2
476. hysterically bad pregnant girl 2
477. another splendid muppet night 2
478. die hard smap fans 2
479. wished dead gets dead 2
480. deliberately shoot people dead 2
481. clichéd love interest subplot 2
482. fine actor robert hardy 2
483. almost goofy romantic comedy 2
484. enough bad character stereotypes 2
485. best known display occurs 1
486. many seemingly impressive scientists 1
487. true romantic comedy classics 1
488. low budget horror films 1
489. suicidally stupid cannon fodder 1
490. ecstatic bastille day throngs 1
491. excellent program notes mark 1
492. totally pointless creepy prison 1
493. college girl gets assaulted 1
494. horror writer stephen king 1

495. dead bodies hanging feet 1
496. amazing career topping performances 1
497. loved long way round 1
498. rubbish horrible cgi makes 1
499. criminals serving long sentences 1
500. untalented cast free reign 1
501. worst cast choice ever 1
502. strongly suggest steering clear 1
503. horror movies usually lack 1
504. almost ruin every scene 1
505. old man getting sick 1
506. facing unknown terrors might 1
507. worst cinematic con trick 1
508. low budget horror movies 1
509. embarrassing vanity project known 1
510. loved ones taken hostage 1
511. bizarre tale torn right 1
512. love lucy reruns instead 1
513. bitchy thai inmate get 1
514. superb special effects work 1
515. showcase magnificent special effects 1
516. three sexy gals venture 1
517. second truly powerful character 1
518. best tv shows ever 1
519. ridiculously bad opening song 1
520. lead character remotely sympathetic 1
521. extremely talented new filmmaker 1
522. hearse lying dead like 1
523. totally predictable crime drama 1
524. fun holiday family movie 1

525. truly funny santa number 1
526. flying government plane failed 1
527. best martial arts movies 1
528. talented young cast makes 1
529. best supporting actress oscar 1
530. love opening sequences like 1
531. aged pretty well despite 1
532. ridiculous bathtub rape scene 1
533. damn fine scary movie 1
534. nasty looking sharp piece 1
535. worst thing ever made 1
536. funniest sequences takes place 1
537. seemingly overwhelming positive views 1
538. nobody till somebody loves 1
539. charlize theron found inspiration 1
540. action movies unfortunately dominating 1
541. pathetic asiaphile fantasies without 1
542. experiencing much pain throughout 1
543. whole story seemed stupid 1
544. prison tearjerker manslaughter finds 1
545. got gary cooper killed 1

C.2 Most Frequent Explanations – eVLSTM+

You can find below a list of the most frequent explanations, along with their frequency for the explained test predictions of the model eVLSTM+.

1. bargain basement production values 81
2. offered depression era audiences 75
3. deliberately shoot people dead 73

4. newer maelstrom like attraction 71
5. one real bad movie 70
6. movie horribly overlook due 69
7. events seem incredibly forced 68
8. two uncomfortably prolonged assaults 64
9. post traumatic stress syndrome 64
10. one honest movie goer 63
11. horrifying shock therapy session 62
12. big budget blockbuster faire 62
13. really bad buddy movie 62
14. rocky horror picture show 61
15. old man getting sick 61
16. post traumatic stress disorder 60
17. glitzy hollywood plastic fantastic 60
18. special entertainment nine excerpts 59
19. former lovers bake baker 58
20. heroine get happily married 57
21. old vietnam war vet 57
22. supporting cast lent credit 56
23. one particularly terrifying scene 56
24. second truly powerful character 56
25. subsequent cuban missile crisis 56
26. illegal dice game going 55
27. defeat real live evil 55
28. really great indie films 54
29. pretty original twist indicating 53
30. really well done honestly 53
31. former award winning actresses 53
32. really good job playing 52
33. two nasty little details 51

34. anyone could easily enjoy 50
35. wished dead gets dead 50
36. crime fighting android thing 50
37. delightfully zany museum curator 49
38. personal best films ever 49
39. delightfully named nettlebed becomes 49
40. pretty painless ninety minutes 49
41. racist guard avoids confrontation 49
42. academy award winning performers 48
43. know micheal elliott loves 48
44. really stupid ending tacked 48
45. terribly boring new age 47
46. negative reviews mainly complained 47
47. special body part helping 47
48. time please please please 46
49. think italian horror cinema 46
50. nazi hall monitors love 45
51. gore effects got stupider 45
52. respectable affluent rich america 45
53. watch racism slowly dissipate 45
54. masterpieces zu neuen ufern 45
55. really wo nt enjoy 45
56. downright strikingly beautiful girls 45
57. cool glamorized sexual moments 44
58. background music perfectly chosen 44
59. know everyone makes fun 44
60. genuinely creepy horror flick 44
61. really really really upset 44
62. former lover terry anderson 44
63. pretty well tell participants 44

64. israeli independence war triggers 44
65. rather conventional romantic comedy 44
66. murderous character called dominic 43
67. severely decomposing male corpse 43
68. reasonably utal rape scene 43
69. means embarrassingly bad like 43
70. stagecoaches seem thrillingly alive 43
71. gorgeous female actresses got 42
72. suicidally stupid cannon fodder 42
73. damn fine scary movie 42
74. almost acquire negative points 41
75. real life war machine 41
76. two stupid old men 41
77. pulitzer prize winning journalist 40
78. woman kills two couples 40
79. corrupt treasury official keen 40
80. raging alcoholic matt sorenson 40
81. sympathetic policewoman officer sudow 40
82. serial killer get caught 40
83. pretty damn good compared 39
84. horror icons tom savini 39
85. gorgeous fashion plate poses 39
86. dishonest crook follows suit 39
87. trigger happy macho pilot 38
88. tiny laugh deep within 38
89. regular love interest holds 38
90. successful snack foods company 38
91. new stupid stuff instead 38
92. fake death ending coming 38
93. good gore scenes like 38

94. gorgeous techicolor production telling 38
95. hippy purist propaganda crap 37
96. successful dental hygienist witnesses 37
97. second favourite horror setting 36
98. sending apparent death threats 36
99. stunningly beautiful art direction 36
100. horror writer stephen king 36
101. dead fleet street journalist 36
102. portray another major pain 36
103. sadists apparently running free 36
104. rapidly decaying nitrate negative 36
105. us wars recently made 35
106. embarrassing vanity project known 35
107. chance like john gulager 35
108. popular spoofs like airplane 35
109. extravagant climactic super ballet 35
110. hysterically bad pregnant girl 35
111. king give impressive performances 35
112. subsequent post traumatic stress 35
113. heroic leading couple visits 35
114. translation ha ha ha 34
115. slightly less poverty stricken 34
116. really worth greatest attention 34
117. dick cheney mutters mostly 34
118. showcase magnificent special effects 34
119. rape scenes might seem 34
120. experiencing much pain throughout 34
121. successful beverly hills dentist 34
122. delicately humorous moments mingled 33
123. recent asian horror films 33

124. illegal bare knuckle fight 33
125. dashing police officer hero 33
126. greatest movies ever made 33
127. stopped crying almost instantly 33
128. killers never loose track 33
129. poetic masterpiece told clearly 33
130. doomed space craft orbiting 33
131. racist local cops raid 33
132. soon softened gorgeous lady 33
133. cold war tension portrayed 33
134. ultimately sympathetic police investigator 33
135. prison tearjerker manslaughter finds 32
136. fifty attackers never seem 32
137. year old horribly ashamed 32
138. nearly perfect digital copy 32
139. really fake italian accent 32
140. freeman aka morgan freeman 32
141. rather sweet value system 32
142. evil sardo numspa ca 32
143. love slapstick comedy shows 32
144. thoroughly engaging romantic drama 32
145. good fun political satire 31
146. done better special effects 31
147. fine actor robert hardy 31
148. sick sad world thinks 31
149. kill forty five minutes 31
150. yet another comic masterpiece 31
151. policemen anally rape male 31
152. unseen maniac slashing teens 31
153. nice sizable supporting part 31

154. indeed amazing filmic imagery 31
155. dead bodies hanging feet 31
156. one low life loser 31
157. horror veteran george kennedy 31
158. various wonderful period costumes 31
159. kids become suicide terrorists 31
160. sexy esther muir plays 30
161. ugly loud hawaiian shirts 30
162. nice sets help keep 30
163. charlize theron found inspiration 30
164. good laugh although unintended 30
165. many horror fans may 30
166. went back home happy 30
167. seemingly overwhelming positive views 30
168. find super spy lovejoy 30
169. utal murder set pieces 30
170. long slow takes encourage 30
171. young hero comes back 30
172. make everyone else laugh 30
173. terry anderson deserve gratitude 30
174. aged pretty well despite 30
175. old unspenseful horror films 30
176. enraged samurai utally murders 29
177. either beautiful party girls 29
178. low budget horror movie 29
179. labeled box office poison 29
180. fantastic roving camera angles 29
181. placed countless glorious animal 29
182. many superb character actors 29
183. average predictable romantic comedy 29

184. gorgeous maruschka detmers takes 29
185. weird misguided masochistic belief 29
186. perfectly syched moog background 29
187. especially loved ray psyching 29
188. yummy blonde marisol santacruz 28
189. pulitzer prize winning play 28
190. greatly appreciated second look 28
191. hurricane nonchalantly devastating everything 28
192. slap stick comedy like 28
193. fun movie without preaching 28
194. violent young criminal visiting 28
195. uniformly superb acting qualifies 28
196. enter world war ii 28
197. excellent movies ever produced 28
198. still worked terrifically well 28
199. kung fu u scenes 28
200. state taking illegal action 28
201. actually quite scary seeing 28
202. weakest vehicle day found 27
203. strongly suggest steering clear 27
204. lovely lilting itish accent 27
205. arcane imagery proudly displayed 27
206. hearse lying dead like 27
207. dick grayson alias nightwing 27
208. criminals serving long sentences 27
209. absolutely fabulous comedy series 27
210. basically birth till death 27
211. simple slow love story 27
212. provide free ukulele picks 27
213. always enjoyed science fiction 27

214. petty thief tough guy 27
215. whilst alive enjoyed staying 27
216. college girl gets assaulted 27
217. ugly deformed serial killer 27
218. music studio get dead 27
219. likely win another oscar 26
220. declared missing presumed killed 26
221. thank god abc picked 26
222. still probably best left 26
223. excellent motion picture archives 26
224. dead race left switched 26
225. female characters slutty idiots 26
226. basic love triangle story 26
227. liberty mutual insurance commercial 26
228. prison escapee finds time 26
229. beautifully cheerful musical score 26
230. bad things naturally occur 26
231. earth destroying everything around 26
232. devastatingly icy david strathairn 26
233. horribly written havoc preceded 26
234. sinister cinema vhs copy 26
235. girl friend either hug 26
236. popular among prominent actors 26
237. foul mouthed unlovable rogue 26
238. mentally retarded boy makes 26
239. absolutely pathetic action scenes 26
240. decent indie horror films 26
241. blind woman defeating hundreds 25
242. ruggedly handsome honorable man 25
243. many bad movies later 25

244. appreciated paolo conte used 25
245. avuncular superior officer assures 25
246. drexler looked pretty good 25
247. war sixty years ago 25
248. low budget horror films 25
249. many talented persons involved 25
250. enjoys receiving root canals 25
251. many people hate black 25
252. terminal stage lung cancer 25
253. standard ugly blonde chick 25
254. callous villain without resorting 25
255. good mann flick thanks 25
256. dead guy would chose 25
257. escaped killer nick grainger 25
258. seemingly petty criminals weigh 25
259. gorgeous looking isabelle adjani 25
260. often subsequent night terrors 24
261. illegal immigrant feels sick 24
262. often disturbing morality tale 24
263. nasty looking sharp piece 24
264. flying government plane failed 24
265. certainly something worth watching 24
266. remember one horrifying incident 24
267. main characters best friend 24
268. individual personalities creating clear 24
269. still thoroughly enjoy watching 24
270. dumb lead actress thinks 24
271. divine violet kemble cooper 24
272. ecstatic bastille day throngs 24
273. sympathetic grumpy shop owner 24

274. totally predictable crime drama 24
275. beautiful autumnal ontario province 24
276. murderous rampage must end 24
277. love pad quite well 24
278. lost valuable precious moments 24
279. god help micheal ironsides 24
280. making particularly bad films 24
281. free speech high horse 24
282. play two slacker friends 24
283. got gary cooper killed 24
284. modern audience certainly accepts 24
285. serial killer sounds intriguing 24
286. found something worth giving 24
287. almost goofy romantic comedy 24
288. snidely whiplash villain roy 23
289. beautiful ritual seem vapid 23
290. late sixties definitely worth 23
291. sure razzie award winner 23
292. far better work like 23
293. painfully obvious horse costume 23
294. terribly gone wrong right 23
295. excellent program notes mark 23
296. mostly beautiful young women 23
297. eventually several people die 23
298. sometimes artistic liberties would 23
299. talented young filmmaker releases 23
300. delightful summer entertainment hits 23
301. glorious technicolor cinematography leaps 23
302. features three excellent performances 23
303. violent barbershop vocal duo 23

304. painfully obvious dim view 23
305. also enjoyed two bouts 23
306. enough bad character stereotypes 23
307. lovely young alisan porter 23
308. master manages happy endings 23
309. excellent facial expressions combine 23
310. police guard gets killed 23
311. sweet little god daughter 23
312. graphical content extremely disturbing 23
313. religious civil war going 23
314. others may feel betrayed 22
315. hardly say anything positive 22
316. definitely worth seeing despite 22
317. beloved younger sister nettie 22
318. poor copy steals lots 22
319. stupid dialogue never matches 22
320. first bad signs come 22
321. ultra evil white slaver 22
322. intentionally stupid title song 22
323. wonderfully touching human drama 22
324. outstanding opening pan shot 22
325. dumb low budget porno 22
326. absolutely amazing production number 22
327. terribly acted kids kill 22
328. truly powerful social satire 22
329. good many terrific shots 22
330. producer dick powell saw 22
331. incredibly bad process work 22
332. beautifully filmed vampire flick 22
333. lead character remotely sympathetic 22

334. sexy charlotte lewis asks 22
335. courageous personnage played magnificently 22
336. us three hundred lovers 22
337. rich old men bidding 22
338. wonderful job tricking people 21
339. hello dave burning paradise 21
340. another prisoner centuries ago 21
341. highly laughable special effects 21
342. idle rich help give 21
343. positively roast every scene 21
344. less interesting stuff worthwhile 21
345. highly respected russian family 21
346. mechanized armies stealing generators 21
347. physically impressive man capable 21
348. almost ruin every scene 21
349. huge jane austen fan 21
350. great location cinematography right 21
351. horrible opening scene establishes 21
352. leslie still look great 21
353. true jane austen fan 21
354. darryl hannah avoid humiliation 21
355. intensely talented performer around 21
356. love opening sequences like 21
357. nobody till somebody loves 21
358. murdered sister every time 21
359. anger management problem due 21
360. bad sheena easton song 21
361. rob gets accidentally hypnotized 21
362. wide white ready grin 20
363. beautiful fraternal love story 20

364. love survive song performed 20
365. great gordon harker provides 20
366. average ufo conspiracy theory 20
367. best live action short 20
368. huge star trek fan 20
369. top five worst movies 20
370. rubbish horrible cgi makes 20
371. classic angry mob style 20
372. whole story seemed stupid 20
373. mean horrible beyond belief 20
374. pleasantly surprised upon rediscovering 20
375. war either required transplanting 20
376. please hire convincing actors 20
377. bizarre murder mystery plot 20
378. already anxious ide insane 20
379. biggest murder mysteries hollywood 20
380. died two years later 20
381. best women character actresses 20
382. clichéd love interest subplot 20
383. walks around looking stricken 20
384. still absolutely horrid makes 20
385. make us laugh afterwards 20
386. many seemingly impressive scientists 20
387. usual socialist economic miracle 20
388. nice talented memorable oddity 20
389. bafta anthony asquith award 20
390. give special undue advantages 19
391. towners ultimately fails miserably 19
392. opera novel seem like 19
393. horror movies usually lack 19

394. many true life romances 19
395. gets rich married men 19
396. happy happy joy joy 19
397. world war ii salute 19
398. compelling true romance story 19
399. three sexy gals venture 19
400. sequence sounds really crude 19
401. intellectually honest person would 19
402. another great orchestral score 19
403. die hard smap fans 19
404. standard comic book villain 19
405. great clive owen fans 19
406. fairy tale loving niece 19
407. abused next door neighbor 19
408. bitchy thai inmate get 19
409. surprisingly effective gretchen moll 19
410. merry men must find 19
411. gorgeously fixating patty shepard 19
412. amazing career topping performances 19
413. amazing martial arts fantasy 19
414. bad rock music blares 19
415. true hollywood happy ending 19
416. whole affair tremendously disappointing 19
417. say enough bad things 19
418. really goofy love triangle 19
419. huge box office success 19
420. typically fun hipster pretenses 19
421. perfect english directly back 19
422. still magnificent béatrice dalle 19
423. loved ones taken hostage 19

424. inventively gory murder sequences 19
425. racing action beautiful pictures 19
426. fascinating yet unsettling look 18
427. completely destroyed within moments 18
428. greatest performance would ever 18
429. avid agatha christie fan 18
430. prestigious hollywood award nomination 18
431. fun holiday family movie 18
432. nasty old hag neighbor 18
433. jackie gleason type villain 18
434. time absolutely savage denunciation 18
435. true romantic comedy classics 18
436. ugly little creature locked 18
437. true john woo masterpiece 18
438. tom cruise made ugly 18
439. action movies unfortunately dominating 18
440. beautiful menzies art design 18
441. every successful director gives 18
442. television horror movie hostess 18
443. perfectly well rank among 18
444. simply wonderful little gem 17
445. superb special effects work 17
446. picture looks terribly drab 17
447. pandering treacly love letter 17
448. another pathetic chase sequence 17
449. spent many happy hours 17
450. james bond type villain 17
451. delightful little thriller opens 17
452. drug war etc etc 17
453. fellow italian horror buffs 17

454. also includes dvd bonuses 17
455. best films ever made 17
456. harshly lighted sets rob 17
457. everyone lives happily ever 17
458. beautiful attracts excellent idea 17
459. admit creating great expectations 17
460. facing unknown terrors might 17
461. truly funny santa number 17
462. great films get better 17
463. talented visual directors working 17
464. mother superior would ask 17
465. stereotyped bad guys waiting 16
466. untalented cast free reign 16
467. worst scripts ever written 16
468. much higher entertainment value 16
469. strong female role model 16
470. man whose great grandmama 16
471. fantastic digitally remastered transfer 16
472. funniest opera ever put 16
473. american flag stands proudly 16
474. ridiculous bathtub rape scene 16
475. quite successful launching campaign 16
476. disney story gone horribly 16
477. surprisingly good special effects 16
478. would die tragically within 16
479. extremely talented new filmmaker 16
480. instant commercial success may 16
481. undergoing statutory rape charges 16
482. great tonino delli colli 16
483. nothing good comes easy 16

484. portrays incredibly cruel treatment 16
485. entertainment value worth watching 15
486. bad guys hires gunmen 15
487. highly motivated probation officer 15
488. best entries like cleo 15
489. best pictures ever made 15
490. worst columbo ever dreamt 15
491. painfully obvious twist ending 15
492. best work angie dickinson 15
493. always marvelous holland taylor 15
494. grown ups hate horrors 15
495. painfully obvious stagebound sets 15
496. hopeful suitor allan lane 15
497. killed without thinking anything 15
498. totally pointless creepy prison 15
499. always admired susan sarandon 15
500. inning let alone lose 15
501. personally liked russell better 15
502. picture warrants special praise 14
503. picture looks incredibly handsome 14
504. lightweight dark comedy entertains 14
505. love chinese epic films 14
506. great comedy writing team 14
507. angry low budget filmmaker 14
508. killer sharks andor crocodiles 14
509. huge adam sandler fan 14
510. trio get hopelessly lost 14
511. protagonists get super powers 14
512. deeply satisfying x file 14
513. worst thing ever made 14

514. stupid blond girl told 14
515. horrible piano music crescendoing 14
516. hugely winning comedic team 13
517. love lucy reruns instead 13
518. loathing teens often feel 13
519. worst cast choice ever 13
520. smiled like three times 13
521. totally beautiful chick flick 13
522. extremely fascinating character study 13
523. anger towards mainly neal 13
524. painfully little character development 13
525. whole show offensively loud 13
526. always wonderful lili taylor 13
527. stupid characters running around 13
528. best supporting actor award 13
529. felt like huge parts 13
530. flawless stock haryanvi accent 13
531. hopelessly simplistic conflicts like 13
532. stupid slasher movies push 13
533. funniest sequences takes place 13
534. utterly terrible acting performances 13
535. pathetic asiaphile fantasies without 13
536. loved long way round 13
537. excellent carter burwell score 13
538. bad songs every minute 12
539. happy black natives dying 12
540. lovely susannah york provides 12
541. worst films ever made 12
542. talented young cast makes 12
543. low budget horror movies 12

544. another easy romantic comedy 12
545. best movies ever made 12
546. adventurous lively farraginous chronicle 12
547. best foreign language movie 12
548. wonderful exciting story together 12
549. another splendid muppet night 12
550. best tv shows ever 12
551. greatest motion picture trilogies 12
552. bizarre tale torn right 12
553. horribly miscast dean cain 12
554. best supporting actress oscar 12
555. creating commercially attractive films 11
556. best major studio films 11
557. worst editing ever combined 11
558. best cult comedy films 11
559. happy coincidence inadvertently saved 11
560. best serious movies ever 11
561. beautiful wealthy salon owner 11
562. emmy award winning performance 11
563. worst movies ever made 10
564. ridiculously forced freudian slips 10
565. bad wig till kidnapped 10
566. oh dear oh dear 10
567. talented director jack gold 10
568. dreadful one missed call 10
569. great curtis mayfield wrote 10
570. best martial arts movies 10
571. worst songs ever included 9
572. worst acting performance ever 9
573. quickly paced murder mystery 9

574. best known display occurs 9
575. perfectly smug kyle maclachlan 8
576. worst movie ever made 8
577. lovable pup ever shown 8
578. depression following almost ruined 8
579. ridiculously bad opening song 8
580. worst cinematic con trick 7

Appendix D

Dictionary of Explanations (RAKE-instance)

The list below contains the phrases extracted using the method RAKE-instance, from the instances of the training set. We emphasize again that the polarity of the extracted phrases does not need to be consistent with the class of the review that the phrase has been extracted from since the reviews often contain bot positive and negative opinion of the movie, one of them dominating the sentiment transmitted for the reviewed movie.

Phrases extracted from the positive classified instances:

1. absolutely amazing production number
2. absolutely fabulous comedy series
3. action movies unfortunately dominating
4. admit creating great expectations
5. adventurous lively farraginous chronicle
6. aged pretty well despite
7. almost ruin every scene
8. also includes dvd bonuses
9. always enjoyed science fiction

10. amazing career topping performances
11. amazing martial arts fantasy
12. american flag stands proudly
13. anger management problem due
14. another easy romantic comedy
15. another great orchestral score
16. another prisoner centuries ago
17. appreciated paolo conte used
18. avid agatha christie fan
19. background music perfectly chosen
20. bad guys hires gunmen
21. bad things naturally occur
22. bafta anthony asquith award
23. basic love triangle story
24. basically birth till death
25. beautiful autumnal ontario province
26. beautiful fraternal love story
27. beautiful menzies art design
28. beautiful wealthy salon owner
29. beloved younger sister nettie
30. best cult comedy films
31. best films ever made
32. best live action short
33. best major studio films
34. best movies ever made
35. best pictures ever made
36. best serious movies ever
37. best supporting actor award
38. best supporting actress oscar
39. best tv shows ever

40. best work angie dickinson
41. big budget blockbuster faire
42. biggest murder mysteries hollywood
43. bitchy thai inmate get
44. bizarre murder mystery plot
45. bizarre tale torn right
46. certainly something worth watching
47. classic angry mob style
48. compelling true romance story
49. corrupt treasury official keen
50. courageous personnage played magnificently
51. creating commercially attractive films
52. criminals serving long sentences
53. damn fine scary movie
54. dashing police officer hero
55. dead fleet street journalist
56. dead race left switched
57. deeply satisfying x file
58. defeat real live evil
59. definitely worth seeing despite
60. delicately humorous moments mingled
61. delightful little thriller opens
62. delightful summer entertainment hits
63. delightfully zany museum curator
64. depression following almost ruined
65. devastatingly icy david strathairn
66. dick grayson alias nightwing
67. die hard smap fans
68. dishonest crook follows suit
69. divine violet kemble cooper

70. doomed space craft orbiting
71. downright strikingly beautiful girls
72. dreadful one missed call
73. drexler looked pretty good
74. ecstatic bastille day throngs
75. either beautiful party girls
76. emmy award winning performance
77. enter world war ii
78. entertainment value worth watching
79. escaped killer nick grainger
80. especially loved ray psyching
81. excellent carter burwell score
82. excellent facial expressions combine
83. excellent motion picture archives
84. excellent movies ever produced
85. excellent program notes mark
86. experiencing much pain throughout
87. extravagant climactic super ballet
88. extremely fascinating character study
89. extremely talented new filmmaker
90. facing unknown terrors might
91. fairy tale loving niece
92. fantastic digitally remastered transfer
93. fantastic roving camera angles
94. fascinating yet unsettling look
95. features three excellent performances
96. fine actor robert hardy
97. flawless stock haryanvi accent
98. flying government plane failed
99. former lover terry anderson

100. former lovers bake baker
101. found something worth giving
102. fun movie without preaching
103. funniest opera ever put
104. funniest sequences takes place
105. genuinely creepy horror flick
106. girl friend either hug
107. give special undue advantages
108. glorious technicolor cinematography leaps
109. good fun political satire
110. good mann flick thanks
111. gorgeous fashion plate poses
112. gorgeous female actresses got
113. gorgeous looking isabelle adjani
114. gorgeous maruschka detmers takes
115. gorgeous techicolor production telling
116. graphical content extremely disturbing
117. great comedy writing team
118. great films get better
119. great gordon harker provides
120. greatest motion picture trilogies
121. greatest movies ever made
122. greatest performance would ever
123. greatly appreciated second look
124. grown ups hate horrors
125. happy coincidence inadvertently saved
126. happy happy joy joy
127. hello dave burning paradise
128. heroine get happily married
129. highly respected russian family

130. horrifying shock therapy session
131. horror movies usually lack
132. huge box office success
133. hugely winning comedic team
134. hurricane nonchalantly devastating everything
135. idle rich help give
136. illegal bare knuckle fight
137. illegal dice game going
138. illegal immigrant feels sick
139. indeed amazing filmic imagery
140. individual personalities creating clear
141. instant commercial success may
142. intensely talented performer around
143. inventively gory murder sequences
144. james bond type villain
145. killed without thinking anything
146. king give impressive performances
147. know everyone makes fun
148. know micheal elliott loves
149. late sixties definitely worth
150. leslie still look great
151. less interesting stuff worthwhile
152. lightweight dark comedy entertains
153. likely win another oscar
154. loathing teens often feel
155. long slow takes encourage
156. lovable pup ever shown
157. love survive song performed
158. loved long way round
159. lovely susannah york provides

160. lovely young alisan porter
161. many bad movies later
162. many horror fans may
163. many people hate black
164. many true life romances
165. master manages happy endings
166. masterpieces zu neuen ufern
167. mean horrible beyond belief
168. mentally retarded boy makes
169. modern audience certainly accepts
170. mostly beautiful young women
171. mother superior would ask
172. movie horribly overlook due
173. murderous character called dominic
174. murderous rampage must end
175. nasty looking sharp piece
176. nazi hall monitors love
177. nearly perfect digital copy
178. negative reviews mainly complained
179. nice sets help keep
180. nice sizable supporting part
181. nice talented memorable oddity
182. nobody till somebody loves
183. nothing good comes easy
184. offered depression era audiences
185. often disturbing morality tale
186. one honest movie goer
187. one low life loser
188. one particularly terrifying scene
189. others may feel betrayed

190. painfully obvious dim view
191. perfectly smug kyle maclachlan
192. perfectly syched moog background
193. perfectly well rank among
194. personal best films ever
195. personally liked russell better
196. petty thief tough guy
197. physically impressive man capable
198. picture warrants special praise
199. placed countless glorious animal
200. play two slacker friends
201. pleasantly surprised upon rediscovering
202. poetic masterpiece told clearly
203. policemen anally rape male
204. popular among prominent actors
205. popular spoofs like airplane
206. portrays incredibly cruel treatment
207. post traumatic stress syndrome
208. prestigious hollywood award nomination
209. pretty damn good compared
210. pretty painless ninety minutes
211. producer dick powell saw
212. pulitzer prize winning play
213. quickly paced murder mystery
214. racing action beautiful pictures
215. racist guard avoids confrontation
216. rapidly decaying nitrate negative
217. rather sweet value system
218. really wo nt enjoy
219. really worth greatest attention

220. recent asian horror films
221. religious civil war going
222. respectable affluent rich america
223. ridiculous bathtub rape scene
224. rubbish horrible cgi makes
225. ruggedly handsome honorable man
226. second favourite horror setting
227. second truly powerful character
228. seemingly petty criminals weigh
229. sending apparent death threats
230. severely decomposing male corpse
231. sexy charlotte lewis asks
232. showcase magnificent special effects
233. simply wonderful little gem
234. slap stick comedy like
235. slightly less poverty stricken
236. special body part helping
237. special entertainment nine excerpts
238. spent many happy hours
239. stagecoaches seem thrillingly alive
240. still thoroughly enjoy watching
241. stopped crying almost instantly
242. strong female role model
243. stunningly beautiful art direction
244. subsequent cuban missile crisis
245. subsequent post traumatic stress
246. successful beverly hills dentist
247. successful dental hygienist witnesses
248. superb special effects work
249. supporting cast lent credit

250. surprisingly effective gretchen moll
251. surprisingly good special effects
252. sympathetic policewoman officer sudow
253. talented director jack gold
254. talented visual directors working
255. talented young cast makes
256. talented young filmmaker releases
257. television horror movie hostess
258. terry anderson deserve gratitude
259. thank god abc picked
260. think italian horror cinema
261. thoroughly engaging romantic drama
262. three sexy gals venture
263. time absolutely savage denunciation
264. totally beautiful chick flick
265. trio get hopelessly lost
266. true john woo masterpiece
267. true romantic comedy classics
268. two nasty little details
269. ugly loud hawaiian shirts
270. ultimately sympathetic police investigator
271. ultra evil white slaver
272. undergoing statutory rape charges
273. uniformly superb acting qualifies
274. us three hundred lovers
275. utal murder set pieces
276. various wonderful period costumes
277. violent barbershop vocal duo
278. violent young criminal visiting
279. walks around looking stricken

280. war either required transplanting
281. war sixty years ago
282. watch racism slowly dissipate
283. went back home happy
284. wide white ready grin
285. wished dead gets dead
286. wonderful exciting story together
287. wonderfully touching human drama
288. world war ii salute
289. would die tragically within
290. year old horribly ashamed
291. yummy blonde marisol santacruz

Phrases extracted from the negative classified instances:

1. absolutely pathetic action scenes
2. abused next door neighbor
3. academy award winning performers
4. actually quite scary seeing
5. almost acquire negative points
6. almost goofy romantic comedy
7. already anxious ide insane
8. also enjoyed two bouts
9. always admired susan sarandon
10. always marvelous holland taylor
11. always wonderful lili taylor
12. anger towards mainly neal
13. angry low budget filmmaker
14. another pathetic chase sequence
15. another splendid muppet night
16. anyone could easily enjoy

17. arcane imagery proudly displayed
18. average predictable romantic comedy
19. average ufo conspiracy theory
20. avuncular superior officer assures
21. bad rock music blares
22. bad sheena easton song
23. bad songs every minute
24. bad wig till kidnapped
25. bargain basement production values
26. beautiful attracts excellent idea
27. beautiful ritual seem vapid
28. beautifully cheerful musical score
29. beautifully filmed vampire flick
30. best entries like cleo
31. best foreign language movie
32. best known display occurs
33. best martial arts movies
34. best movies ever made
35. best women character actresses
36. blind woman defeating hundreds
37. callous villain without resorting
38. chance like john gulager
39. charlize theron found inspiration
40. clichéd love interest subplot
41. cold war tension portrayed
42. college girl gets assaulted
43. completely destroyed within moments
44. cool glamorized sexual moments
45. crime fighting android thing
46. darryl hannah avoid humiliation

47. dead bodies hanging feet
48. dead guy would chose
49. decent indie horror films
50. declared missing presumed killed
51. deliberately shoot people dead
52. delightfully named nettlebed becomes
53. dick cheney mutters mostly
54. died two years later
55. disney story gone horribly
56. done better special effects
57. drug war etc etc
58. dumb lead actress thinks
59. dumb low budget porno
60. earth destroying everything around
61. embarrassing vanity project known
62. enjoys receiving root canals
63. enough bad character stereotypes
64. enraged samurai utally murders
65. events seem incredibly forced
66. eventually several people die
67. every successful director gives
68. everyone lives happily ever
69. evil sardo numspa ca
70. fake death ending coming
71. far better work like
72. fellow italian horror buffs
73. felt like huge parts
74. female characters slutty idiots
75. fifty attackers never seem
76. find super spy lovejoy

77. first bad signs come
78. former award winning actresses
79. foul mouthed unlovable rogue
80. free speech high horse
81. freeman aka morgan freeman
82. fun holiday family movie
83. gets rich married men
84. glitzy hollywood plastic fantastic
85. god help micheal ironsides
86. good gore scenes like
87. good laugh although unintended
88. good many terrific shots
89. gore effects got stupider
90. gorgeously fixating patty shepard
91. got gary cooper killed
92. great clive owen fans
93. great curtis mayfield wrote
94. great location cinematography right
95. great tonino delli colli
96. happy black natives dying
97. hardly say anything positive
98. harshly lighted sets rob
99. hearse lying dead like
100. heroic leading couple visits
101. highly laughable special effects
102. highly motivated probation officer
103. hippy purist propaganda crap
104. hopeful suitor allan lane
105. hopelessly simplistic conflicts like
106. horrible opening scene establishes

107. horrible piano music crescendoing
108. horribly miscast dean cain
109. horribly written havoc preceded
110. horror icons tom savini
111. horror veteran george kennedy
112. horror writer stephen king
113. huge adam sandler fan
114. huge jane austen fan
115. huge star trek fan
116. hysterically bad pregnant girl
117. incredibly bad process work
118. inning let alone lose
119. intellectually honest person would
120. intentionally stupid title song
121. israeli independence war triggers
122. jackie gleason type villain
123. kids become suicide terrorists
124. kill forty five minutes
125. killer sharks andor crocodiles
126. killers never loose track
127. kung fu u scenes
128. labeled box office poison
129. lead character remotely sympathetic
130. liberty mutual insurance commercial
131. lost valuable precious moments
132. love chinese epic films
133. love lucy reruns instead
134. love opening sequences like
135. love pad quite well
136. love slapstick comedy shows

137. loved ones taken hostage
138. lovely lilting itish accent
139. low budget horror films
140. low budget horror movie
141. low budget horror movies
142. main characters best friend
143. make everyone else laugh
144. make us laugh afterwards
145. making particularly bad films
146. man whose great grandmama
147. many seemingly impressive scientists
148. many superb character actors
149. many talented persons involved
150. means embarrassingly bad like
151. mechanized armies stealing generators
152. merry men must find
153. much higher entertainment value
154. murdered sister every time
155. music studio get dead
156. nasty old hag neighbor
157. new stupid stuff instead
158. newer maelstrom like attraction
159. often subsequent night terrors
160. oh dear oh dear
161. old man getting sick
162. old unspenseful horror films
163. old vietnam war vet
164. one real bad movie
165. opera novel seem like
166. outstanding opening pan shot

167. painfully little character development
168. painfully obvious horse costume
169. painfully obvious stagebound sets
170. painfully obvious twist ending
171. pandering treacly love letter
172. pathetic asiaphile fantasies without
173. perfect english directly back
174. picture looks incredibly handsome
175. picture looks terribly drab
176. please hire convincing actors
177. police guard gets killed
178. poor copy steals lots
179. portray another major pain
180. positively roast every scene
181. post traumatic stress disorder
182. pretty original twist indicating
183. pretty well tell participants
184. prison escapee finds time
185. prison tearjerker manslaughter finds
186. protagonists get super powers
187. provide free ukulele picks
188. pulitzer prize winning journalist
189. quite successful launching campaign
190. racist local cops raid
191. raging alcoholic matt sorenson
192. rape scenes might seem
193. rather conventional romantic comedy
194. real life war machine
195. really bad buddy movie
196. really fake italian accent

197. really good job playing
198. really goofy love triangle
199. really great indie films
200. really really really upset
201. really stupid ending tacked
202. really well done honestly
203. reasonably utal rape scene
204. regular love interest holds
205. remember one horrifying incident
206. rich old men bidding
207. ridiculously bad opening song
208. ridiculously forced freudian slips
209. rob gets accidentally hypnotized
210. rocky horror picture show
211. sadists apparently running free
212. say enough bad things
213. seemingly overwhelming positive views
214. sequence sounds really crude
215. serial killer get caught
216. serial killer sounds intriguing
217. sexy esther muir plays
218. sick sad world thinks
219. simple slow love story
220. sinister cinema vhs copy
221. smiled like three times
222. snidely whiplash villain roy
223. sometimes artistic liberties would
224. soon softened gorgeous lady
225. standard comic book villain
226. standard ugly blonde chick

227. state taking illegal action
228. stereotyped bad guys waiting
229. still absolutely horrid makes
230. still magnificent béatrice dalle
231. still probably best left
232. still worked terrifically well
233. strongly suggest steering clear
234. stupid blond girl told
235. stupid characters running around
236. stupid dialogue never matches
237. stupid slasher movies push
238. successful snack foods company
239. suicidally stupid cannon fodder
240. sure razzie award winner
241. sweet little god daughter
242. sympathetic grumpy shop owner
243. terminal stage lung cancer
244. terribly acted kids kill
245. terribly boring new age
246. terribly gone wrong right
247. time please please please
248. tiny laugh deep within
249. tom cruise made ugly
250. top five worst movies
251. totally pointless creepy prison
252. totally predictable crime drama
253. towners ultimately fails miserably
254. translation ha ha ha
255. trigger happy macho pilot
256. true hollywood happy ending

257. true jane austen fan
258. truly funny santa number
259. truly powerful social satire
260. two stupid old men
261. two uncomfortably prolonged assaults
262. typically fun hipster pretenses
263. ugly deformed serial killer
264. ugly little creature locked
265. unseen maniac slashing teens
266. untalented cast free reign
267. us wars recently made
268. usual socialist economic miracle
269. utterly terrible acting performances
270. weakest vehicle day found
271. weird misguided masochistic belief
272. whilst alive enjoyed staying
273. whole affair tremendously disappointing
274. whole show offensively loud
275. whole story seemed stupid
276. woman kills two couples
277. wonderful job tricking people
278. worst acting performance ever
279. worst cast choice ever
280. worst cinematic con trick
281. worst columbo ever dreamt
282. worst editing ever combined
283. worst films ever made
284. worst movie ever made
285. worst movies ever made
286. worst scripts ever written

287. worst songs ever included
288. worst thing ever made
289. yet another comic masterpiece
290. young hero comes back

Appendix E

Small Dictionary of Explanations (RAKE-corpus)

Phrases extracted from the positive corpus:

- highly recommend happy go lovely
- dead hard core terrorist
- murder evil
- fun loving happy young woman
- funny funny funny funny
- amazing true love story
- best supporting actress award
- best supporting actor award essentially
- lovingly innocent movie romance
- one horrific murder
- best supporting actress academy award
- best supporting actor award
- beautiful beautiful beautiful movie
- killer murdered two victims immediately
- suffered miserable abuse

- nyc great comedy great drama
- courageous personnage played magnificently
- good story great humour
- happy happy joy joy
- mad mad mad mad world
- enjoy relaxing easy humor
- bad ass criminal
- ensure great success
- horrific murder sequences

Phrases extracted from the negative corpus:

- great love story looks like
- serial killer killing people
- murder rape
- killer killing
- rapists rape jennifer
- bad bad bad bad
- love interest also help saving
- killed stone cold dead
- dead killed
- kill defeated foe
- evil witch killing
- evil dead
- fake fake fake fake
- international terrorist dead
- torture war criminals
- combines perfectly fabulous special effects
- evil dead ii
- stupid stupid stupid
- infected thief goes mad

- beautiful attracts excellent idea
- perfectly honest looks like
- presumably dead evil professor
- evil dead made
- dear dear dear dear dear
- hated hated hated
- evil vs evil rather
- dull bad horror movie
- kill murderer
- murder suicide
- bad bad bad
- devoted best friend
- bad bad bad bad bad
- killer killed
- excruciatingly bad boring movie
- evil dead set
- horrible horrible horrible