Titre: Title:	Learning to Rank with BERT for Argument Quality
Auteur: Author:	Charles-Olivier Favreau
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Favreau, C.-O. (2022). Learning to Rank with BERT for Argument Quality [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10296/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Learning To Rank with BERT for Argument Quality

CHARLES-OLIVIER FAVREAU
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Avril 2022

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Learning To Rank with BERT for Argument Quality

présenté par **Charles-Olivier FAVREAU**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Michel DESMARAIS, président
Amal ZOUAQ, membre et directrice de recherche
Sarath Chandar ANBIL PARTHIPAN, membre

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor Amal Zouaq for her guidance and insightful advice through all the stages of my project.

I would also like to thank Sameer Bhatnagar for his support and his participation in the annotation process, as well as his advice, as someone with hands-on experience on the topic.

This research is supported by an INSIGHT grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

RÉSUMÉ

À ce jour, la tâche d'ordonnancement de la qualité des arguments demeure un défi. Celle-ci vise à évaluer une mesure de qualité des arguments sous la forme de textes libres. La grande majorité des initiatives faisant partie de l'état de l'art approchent la tâche en utilisant des méthodes d'ordonnancement de type «pointwise», cherchant à prédire un score de qualité absolu. Nous proposons plutôt de chercher à apprendre à ordonner les arguments selon leur mesure relative de qualité. En effet, nous expérimentons avec plusieurs méthodes d'apprentissage d'ordonnancement, tels que des méthodes de type «pointwise», «pairwise» et «list-wise». Nous comparons la performance de chacune de ces méthodes sur la tâche d'ordonnancement de la qualité des arguments.

Pour ce faire, nous utilisons la puissante capacité de l'architecture BERT à construire la représentation d'un argument, combinée avec des méthodes d'apprentissage d'ordonnancement, pour ordonner de manière effective les arguments d'une liste selon leur mesure de qualité. De plus, nous démontrons qu'un ensemble de modèles entraînés avec des fonctions de pertes différentes augmente la performance pour l'identification des arguments les plus convaincants d'une liste. Nous comparons l'architecture BERT, combinée à des méthodes d'apprentissage d'ordonnancement, avec les méthodes de l'état de l'art. Nous effectuons cette comparaison sur tous les ensembles de données majeurs de qualité d'argument et démontrons comment une approche d'apprentissage d'ordonnancement présente une meilleure performance à identifier les arguments les plus convaincants d'une liste.

Finalement, nous explorons la faisabilité d'unifier les ensembles de données de qualité d'argument avec une mesure standardisée de qualité. Plusieurs ensembles de données de qualité d'arguments diffèrent dans la manière dont les scores de qualité sont extraits des annotations collectées, d'où la nécessité d'une mesure commune. Uniformiser ces ensembles de données de qualité d'argument permet de comparer notre approche aux approches de l'état de l'art de manière plus homogène. Nous proposons la métrique WinRate comme mesure standardisée de qualité d'argument et démontrons comment cette métrique permet d'uniformiser les ensembles de données, montrant une performance plus constante sur les ensembles de données.

ABSTRACT

The task of argument quality ranking, which identifies the quality of free text arguments, remains, to this day, a challenge. While most state-of-the-art initiatives use point-wise ranking methods and predict an absolute quality score for each argument, we instead focus on learning how to order them by their relative convincingness. Therefore, we experiment with several learning-to-rank methods for the argument quality ranking task, including pointwise, pairwise and list-wise learning-to-rank approaches. We compare how each of these methods perform on different argument quality datasets.

We leverage BERT's powerful ability in building a representation of an argument, paired with learning-to-rank approaches to rank arguments according to their measure of convincingness. We also demonstrate how an ensemble of models trained with different ranking losses often improves the performance for the identification of the most convincing arguments of a list. We compare BERT coupled with learning-to-rank methods to state-of-the-art approaches on all major argument quality datasets available for the ranking task, demonstrating how a learning-to-rank approach performs better at outlining the topmost convincing arguments.

Finally, we explore the feasibility of unifying argument quality datasets with a standardized convincingness metric, as they differ greatly in the way the quality scores are inferred from collected argument annotations. Standardizing argument quality datasets with a common metric allows for a more consistent evaluation of our solution across datasets and therefore, allows for a better comparison to state-of-the-art solutions. We propose the WinRate as a standardized measure of argument quality, and we demonstrate how it unifies datasets, demonstrating more consistent performance of our solution across datasets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short Term Memory
GloVe	Global Vectors for Word Representation
GPPL	Gaussian Process Preference Learning
GPC	Gaussian Process Classifier
SVM	Support Vector Machine
SVR	Support Vector Regression
MACE	Multi-Annotator Competence Estimation
WA	Weighted-Average
NDCG	Normalized Discounted Cumulative Gain
TF	TensorFlow
TFR-BERT	TensorFlow Ranking - BERT
RoBERTa	Robustly Optimized BERT Pretraining Approach
NLP	Natural Language Processing

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Argumentation is a tool to convince an audience of a stance on a given topic using arguments, which consist of one or many phrases or sentences [4]. Stance is defined as the overall position toward an idea, object or proposition [5]. For example, given the topic "Zoos should be abolished" and the stance "Pro", one could argue that a zoo's whole business model is to take animals from their natural habitats and exploit them for money (Dataset *IBM ArgQ 30k* [6]). This yields the question: how can one identify a convincing argument? The automatic assessment of argument quality, a subfield of *Natural Language Processing*, aims at answering that very question. In fact, convincingness is a primary dimension of argument quality [7] and has been the main focus of argument quality research.

1.1 Motivation

The motivation behind being able to evaluate the quality of an argument using machine learning can be explained with the numerous applications where such capability would prove to be useful. Modeling convincingness is useful to many fields. For example, in Educational Data Mining, [8] explains how evaluating the convincingness of a student rationale is important for *Technology-Mediated Peer Instruction* (TMPI) systems. TMPI systems, which are a form of *Learnersourcing*, ask students to submit explanations to justify their choice in the context of multiple choice questions. Students are then presented with other explanations, submitted by their peers. After considering peer rationales, a student can reconsider his own answer. Therefore, the convincingness aspect of a student's rationale impacts the learning process of his peers.

Many other applications can be listed [9, 10]. For instance, ranking arguments according to their quality is an important step of the process of building an argument search engine for the Web, as presented by [9]. As another example, [10] shows how modeling argument quality is used to annotate arguments, their components and relations in persuasive essays.

Defining the attributes of a strong argument is very subjective [7]. In fact, [11] states that the logical structure of an argument, as well as other factors such as the speaker, the emotions and the context have an impact on the argument's quality. Moreover, [12] demonstrates through experiments that the same argument can be regarded differently depending on the

audience. As [6] outlines, rather than strictly defining argument quality, which is subjective, we can list characteristics typically describing strong arguments, as well as characteristics describing weak arguments. For example, bad grammar and low clarity are clear indicators of a weak argument, whereas a strong argument can generally be described as relevant, with high impact. That being said, [2] demonstrates how a neural approach to modeling argument quality using word embeddings outperforms approaches based on task specific features. Moreover, approaches based on task specific features require a significant greater amount of pre-processing steps. Recent work show the benefits of leveraging deep pre-trained language models for argument quality assessment [6, 7]. Therefore, in this work, we focus on deep neural language models to model argument quality (see section 4.1).

1.2 Research Objectives

Automatic argument quality assessment consists in predicting a measure of *convincingness* for an argument given a topic on which the argument is taking position on. *UKP Lab* ([1]) proposed 2 tasks in the field of computational argumentation: First, the task of predicting the most convincing argument of a pair of arguments and second, the task of ranking a list of arguments, according to their convincingness. While the performance of state-of-the-art models on the first task is impressive, the task of ranking arguments in order of convincingness proves to be more challenging [7]. Another challenge of the argument quality ranking task is how argument quality datasets are different in the way their quality score is inferred. This prevents the comparison of models in a uniform way, across datasets.

In this work, we focus on the second task: ranking a list of arguments for a given topic, in order of convincingness. This is, as one would expect, more complex than simply choosing the most convincing argument out of a pair of arguments ([1]'s first task). Most solutions so far approached the task as predicting an absolute quality score for each argument individually, defined as point-wise ranking [1, 6, 7]. While these methods produce the desired outcome as the list of the predicted scores can be sorted to order the arguments, we hypothesize that there are ranking capabilities potentially lost during the learning process by not comparing the arguments together. We propose to define the problem as a true ranking task, where we do not evaluate each argument's individual measure of quality, but instead focus on evaluating its relative convincingness compared to other arguments. In this work, we try to answer the following research question :

How can learning-to-rank techniques contribute to automatic argument quality

evaluation ?

To answer this research question, we propose to leverage a neural approach to *learning-to-rank*, built on top of BERT [13], a modern neural language model that has shown impressive results on several Natural Language Processing (NLP) tasks. This method combines BERT's strong ability to build an argument's representation, and different ranking loss functions (pointwise, pairwise, list-wise). This solution allows us to evaluate the quality of a group of arguments by ordering them from most convincing to least convincing. We also propose a standardized metric for argument quality to unify major argument quality datasets, as each dataset differs greatly in the way the quality scores are inferred from collected argument annotations (This will be discussed in chapter 3). As our proposed solution for argument quality is evaluated on multiple datasets, we want the quality score from each dataset to be comparable to one another, ensuring a uniform evaluation across datasets. To achieve this, it is necessary to answer more specific research questions:

Q1 : How can learning to rank techniques coupled with pretrained language models contribute to automatic argument quality evaluation?

Q2 : How can argument quality datasets be standardized with a common score, to facilitate the comparison of models' performance on the ranking task?

1.3 Contributions

BERT & Learning-to-rank In this work, we present a different method to argument quality ranking and approach it as a true ranking task. We compare pointwise, pairwise and list-wise learning-to-rank methods for the argument quality ranking task, introducing list-wise learning-to-rank methods to the field of argument quality. Furthermore, we combine learning-to-rank methods with pretrained language models (BERT). We demonstrate how our approach outperforms state-of-the-art solutions on NDCG@K metrics.

Standardized Score Moreover, we explore how argument quality datasets can be standardized with a common score. This allows to unify datasets and represents our solution to the heterogeneity among various methodologies and datasets. We thoroughly compare the presented standardized score to each dataset's original quality score and analyze the benefits gained from using a standardized score.

Ranking task on *IBM EviConv* To the best of our knowledge, we are the only ones who evaluate the argument quality ranking task on dataset *IBM EviConv*, presented by [2]. The authors evaluate their approach on the argument pair classification task only. While they publish individual scores for each argument as part of the dataset, they don't address the ranking task.

1.4 Thesis Outline

The findings of our work will be presented as follows. We first present previous state-of-the-art approaches for the argument quality evaluation task in chapter 2. This literature review allows us to establish the baseline that we use as comparison for any solution we propose. Chapter 3 describes the datasets used to evaluate our solution, describing how they compare in the way they were collected. The 4 datasets presented will be used as an evaluation source, common to both our approach and state-of-the-art methods. In chapter 4, we explain our methodology and the architecture of our proposed solution. We also describe the performance metrics used to evaluate our model. The results obtained are compared to state-of-the-art models. Chapter 5 explores the feasibility of using a normalized score for all argument quality datasets, thus unifying them.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

In this chapter, we first look over the concepts relevant to our work, defining the building blocks used in our approach to tackle argument quality ranking. We define the pre-trained language model BERT, which we later use to build argument representation. We define the concepts behind learning-to-rank methods, as it is fundamental to our approach in solving the argument quality ranking task. Then, we define the argument quality assessment task, and we review the main methods presented in the last 5 years for the automatic argument quality assessment task. This allows us to identify state-of-the-art solutions, and therefore, establish the baseline to which we will later compare our approach.

2.1 BERT

In this section, we discuss Bidirectional Encoder Representations from Transformers (BERT) [13], which we later use to build a representation of an argument in regard to its respective topic. BERT consists of a transformer, an attention-based model, applied to language modelling. Through an attention mechanism, BERT learns contextual relations between words in text. Instead of reading the input from left to right, like multiple models do, BERT reads the total sequence of words at once: this is why it is considered bidirectional. This allows BERT to learn the context of a word by looking at the words surrounding it. BERT's strong ability at building a feature representation of text resides in the fact that it was pre-trained on large corpora: the whole English Wikipedia corpus and the Brown Corpus, producing a model that has a strong initial understanding of the English language and can be fined-tuned on a more specific task. BERT is pre-trained on 2 tasks: masked language modeling and next sentence prediction.

Masked Language Modeling BERT is pretrained on the masked language modeling task, which consists in predicting masked words in sentences. In each sequence of words passed to BERT, 15% of words are masked, and BERT is trained at predicting the masked words using the context of the words that are not masked. Figure 2.1 shows the configuration for training on the masked language modeling task. A classification layer is added on top of the encoder output. The output is multiplied by the embedding matrix to be transformed into words from the vocabulary. Finally, a softmax function is applied to calculate the probability of each word in the vocabulary as candidate for the position of the masked word. During the training phase, the loss function takes only the predictions on masked words into account.

Next Sentence Prediction The second aspect of BERT's pretraining is next sentence prediction. The model is trained on the following task: given a pair of sentences, the model must predict if the second sentence is the sentence after the first sentence in the original corpus. During training, BERT is fed pairs of sentences, of which 50% are subsequent sentences and 50% aren't. To establish the delimitation between the 2 sequences, special tokens are added to the sentences' tokens: [CLS] at the beginning of tokens, [SEP] between the 2 sequences and at the very end. In addition to the tokens embeddings, sentence embeddings and positional embedding are fed as input to the transformer. Sentence embeddings indicate if a token belongs to the first or second sentence. The entire sequence is given as input to BERT, and a classification layer using Softmax is applied to the output of the [CLS] token, predicting the probability that the second sentence is subsequent of the first. Both masked language modeling and next sentence prediction tasks are trained altogether. The training goal is to minimize the combined loss of the 2 tasks.

Being trained on both masked language modeling and next sentence prediction, BERT is a pre-trained model with a strong initial understanding of the English language. It can be fined-tuned on a more specific task. As we will describe in section 4.1, we use BERT as a building block responsible for learning a representation of arguments with respect to their topic, for the argument quality ranking task.

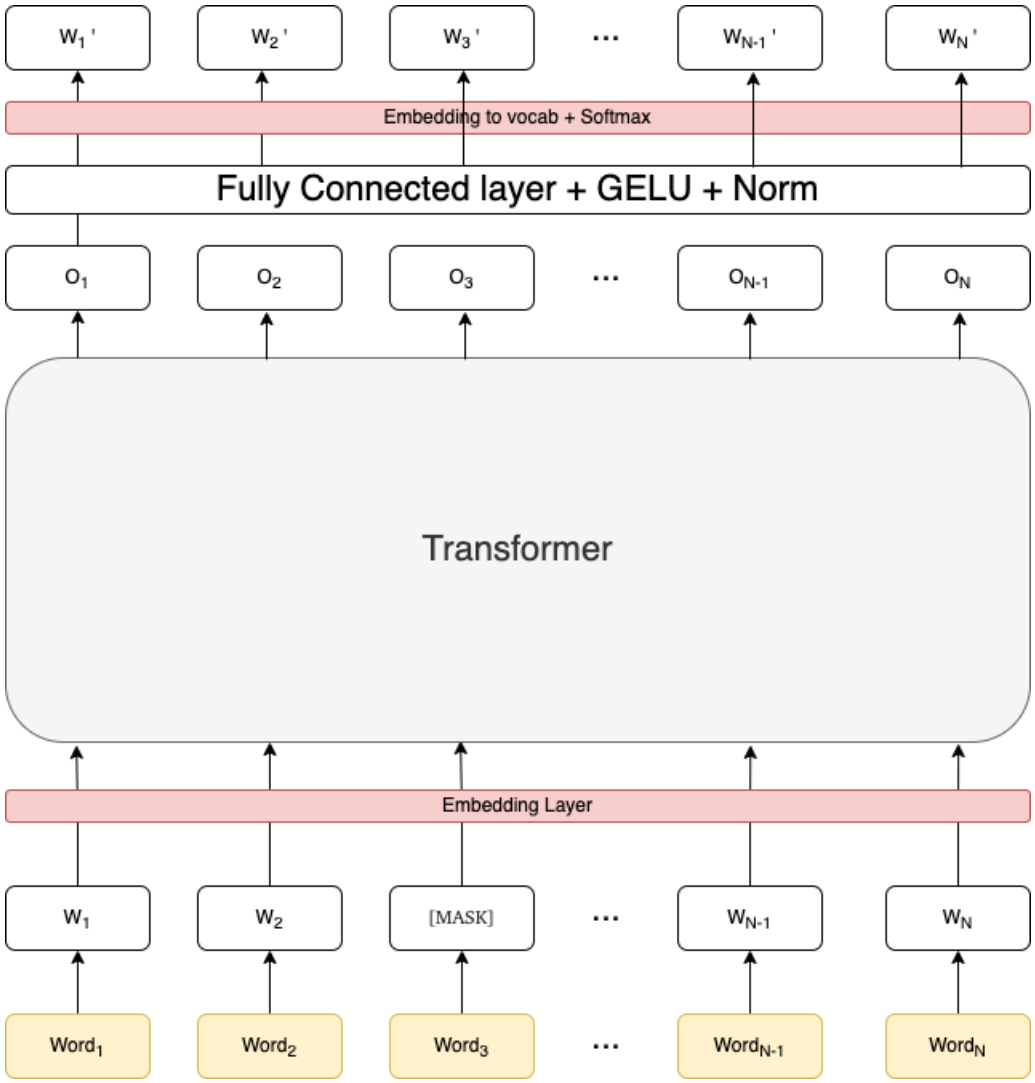

Figure 2.1 BERT architecture for masked language modeling.

2.2 Learning-to-rank

Learning-to-rank methods consist of machine learning applied to the task of ranking a list of items based on the features of those items. These methods focus on the relative order of the items instead of the score predicted for each item. *Learning-to-rank* methods can be divided into 2 groups ([14]): *Learning for Ranking Creation*, which is focused on building a ranking model using machine learning techniques and *Learning for Ranking Aggregation*, which is focused on generating a ranked list of items from multiple ranked lists of items [14]. For the purpose of this work, we focus on *Learning for Ranking Creation* as it applies to our task at hand: ranking a list of arguments according to their relative convincingness.

2.2.1 Learning for Ranking Creation

Data Labeling

For a learning-to-rank task, items to rank are labeled with relevance labels, given topicality. Two types of relevance label can be used to describe items to rank: *binary relevance* and *graded relevance* labels. Assuming the relevance of items is initially represented as a continuous variable, the variable is divided into categories [15]. The choice of the number of categories depends on the application, as it influences how the ranking is modeled. *Binary relevance*, where the variable is divided into 2 categories: relevant or not relevant, is commonly used in Information Retrieval (IR) as it appropriately models the concept of a document being relevant or not relevant to a query. However, it implies that all relevant documents are equally relevant to the query. *Graded relevance* is used when ranking items according to a degree of relevance. In this work, we work with graded relevance labels. For argument quality ranking, we want to rank a list of arguments by their relative convincingness. Therefore, graded relevance labeling is the appropriate labeling technique.

As any supervised task, learning-to-rank methods require labeled data: a gold standard consisting of the ranked list of items for a specific context. Applying this idea to our ranking task, we formulate the problem as follows. We have a set of topics T and a set of arguments A. For each topic T_i, the gold standard assigns a label Y_i from $Y = \{1, 2, ..., r\}$ to each argument a_i from the set of arguments A_i related to the topic T_i. The label Y_i represents a grade. The list of grades consists of a total order between grades: $r \prec r-1 \prec ... \prec 1$, where $\prec$ shows the order relation [14].

Feature extraction The ranking models aim to learn a function $f(x)$ which takes a feature vector x as input. This vector x is a feature vector based on both the topic and the argument. This is important to assure the model is able to generalize to new data and more importantly to new topics. The feature vector x should be a representation of the topic and the argument, appropriately building how they interact with each other, and therefore identifying how the argument is relevant to the topic. We present in section 2.1 how we intend to build a representation of the topic & argument pairs.

Ranking function Through supervised learning, we learn a ranking function $f(x)$ over training examples. This neural function is learned through gradient descent. We show in section 4.2 the different ranking losses we use during training, and then we show in section

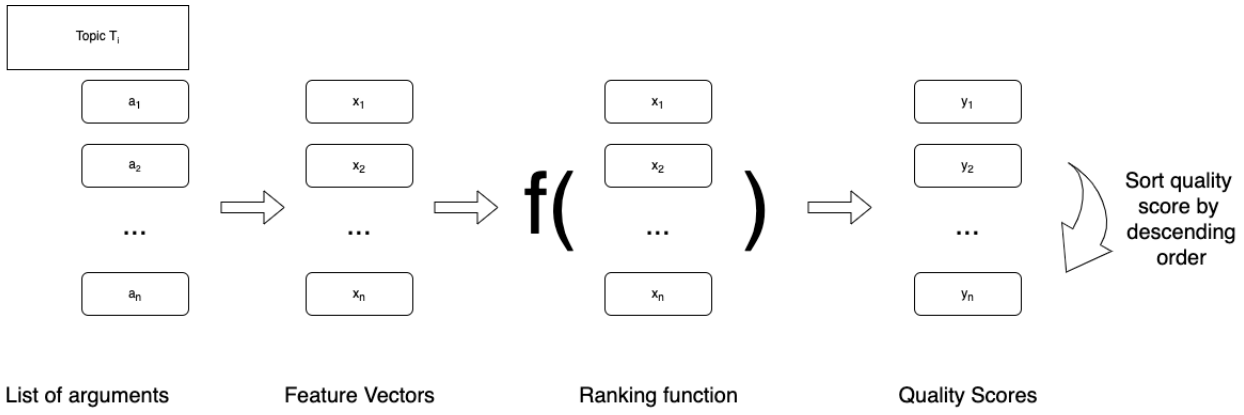

Figure 2.2 Learning-to-rank steps.

4.4 how each ranking loss yields a ranking model performing differently. The ranking model takes an entire list of arguments as input, and learns an ordering that optimizes the relative ordering of the entire list of arguments. The ranking function $f(x)$ outputs a score for each argument, as shown in figure 2.2. Those scores are sorted in descending order, generating a ranked list of arguments, ordered from the highest quality argument, to the lowest quality argument. The major difference between the learning-to-rank approaches lies in the choice of loss function. The loss function can be calculated over individual items of the ranked list, over pairs of items in the ranked list or over the whole ordered list, as we explain in the next sections.

Pointwise Ranking

A pointwise approach to learning to rank considers the ranking problem as a classification, regression or ordinal classification task. Therefore, existing methods for classification, regression or ordinal classification can be applied to learning the ranking function. Pointwise ranking ignores any group structure of the items to rank, and each item is considered individually. In other words, the loss function used to learn the ranking function $f(x)$ is defined on the feature vector of items to rank, considering each feature vector individually [14].

Pairwise Ranking

A pairwise approach to learning-to-rank defines the ranking problem as a binary classification task, where the model learns from preference pairs of features. From the ranked list, a preference pair between $item_i$ and $item_j$ can be defined as positive if $item_i$ is ranked higher than $item_j$ in the list and negative otherwise. Learning from preference pairs classification

provides a model capable of ranking a list of items. Therefore, the ranking model's performance is defined by its performance on the pairwise classification task. The loss function used for pairwise ranking is defined on pairs of features vectors from the ranked list. Similarly to pointwise ranking, during training, pairwise ranking ignores the group structure of the ranked list of items as a whole and focuses on preference pairs. [14].

List-wise Ranking

A list-wise approach to learning-to-rank, compared to pointwise and pairwise approaches, considers the group structure of the ranked list of items as a whole during the learning process. The model learns a ranking function $f(x)$ from the entire list of feature vectors, each labeled with a score, allowing for the ranking function to grasp the group structure of the ranked list of items. As [14] outline, this is a new problem for machine learning and traditional machine learning methods cannot be applied directly. Diverse solutions are proposed to tackle this problem, like the Luce-Plackett model for example, which calculates the permutation probability of items in the list [16]. We further describe Luce-Plackett model in section 4.2.

2.2.2 Ranking Aggregation for pairwise preferences

In this section, we describe the specific case of ranking aggregation for pairwise preferences. The majority of the datasets we present in chapter 3 are collected through crowdsourcing, where the feedback from multiple crowd workers is merged into rankings. Ranking n items can also be considered as the collection of $\binom{n}{2}$ preference pairs [17]. Thus, we first present established solutions to aggregating pairwise preferences into ranking.

Bradley-Terry (BT)

The Bradley-Terry model is a rank aggregation model for pairwise preferences. The probability of a_i being chosen over a_j relies on s_{a_i} and s_{a_j}, which correspond to strength parameters, as shown in equation 2.1 [17].

$$P(a_i > a_j) = \frac{1}{1 + e^{-(s_{a_i} - s_{a_j})}} \tag{2.1}$$

The value of the strength parameter s_{a_i} is evaluated across pairwise preferences using the following maximum a posteriori estimator, shown in equation 2.2:

$$\hat{s} = argmax_s\{Pr(s) \prod_{g \in G} \prod_{a_i \succ_{\rho(g)} a_j} \frac{1}{1 + e^{-(s_{a_i} - s_{a_j})}}\} \tag{2.2}$$

Where, G is a set of graders $G = \{g_1, ..., g_k\}$ and ρ is a set of pairwise preference from grader g.

WinRate

The WinRate metric, applied to pair annotations, consists of the number of times an argument is chosen as the most convincing of the pair over the number of times the argument is shown overall:

$$WinRate(arg) = \frac{\sum_{i=0}^{n} y_i}{n} \tag{2.3}$$

Where y_i is the binary label of the argument for the occurrence i out of n occurrences in total. In other words, this means that for a topic, n argument pairs include this argument. For each argument pair, the label y_i indicates whether the argument is the most convincing argument of the two arguments (1) or is the less convincing argument of the two arguments (0). Therefore, the WinRate is simply the number of times an argument is labeled as more convincing than another argument over the number of times it is compared to another argument [18].

Elo

The Elo rating system can be applied to ranking aggregation of pairwise preferences, as shown by [8]. While it was initially presented to rank chess players [19], [20] demonstrated its application in other fields. The probability of an item a_i being ranked higher than another item a_j is given by equation 2.4:

$$P(a_1 > a_2) = \frac{1}{1 + 10^{(\beta_{a_2} - \beta_{a_1})/\delta}} \tag{2.4}$$

where β_{a_1} and β_{a_2} correspond to the strength of a_i and a_j respectively. δ is a scaling constant. β_{a_i} parameters are updated after each pairwise preference seen as shown in equation 2.5:

$$\beta'_{a_1} = \beta_{a_1} + K(P(a_1 > a_2) - \beta_{a_1}) \tag{2.5}$$

Where K is a constant indicating the maximum adjustment per pairwise preference, called the K-factor, and is fixed to a value [20].

We presented solutions to aggregating pairwise preferences into rankings. While the Bradley-Terry and Elo models are well-established, the WinRate is relatively new, introduced as a pairwise preferences' aggregation method in the context of argument quality ranking by [18]. As we explain in chapter 3, many argument quality datasets are collected as annotated pairs of arguments and, therefore, require aggregating pairwise preferences into rankings to extract ranked lists of arguments.

2.3 Argument Quality Evaluation

Argument quality evaluation or *Argument quality assessment* is the subfield tackling the long-standing challenge of modeling argument quality. The difficulty of the task is mostly explained by its subjectivity [6]. In this section, we look over the state-of-the-art for the argument quality evaluation task, describing the solutions presented in the last 5 years. We first define rigorously the argument and its components. We then present state-of-the-art solutions to the two main tasks of argument quality evaluation: the classification task and the ranking task, both of which have been defined and pioneered by [1]. The classification task consists of choosing the most convincing argument, given 2 arguments on the same topic. The ranking task can be expressed as the ordering of a list of arguments by their relative measure of quality, as previously explained.

2.3.1 Definitions

First, we define the term *argument*. An argument consists of one or more phrases or sentences, composed of the claim and the premise (also called evidence), jointly forming the argument [4]. The claim is either supported or contested by one or multiple premises. The argument is trying to convince the audience of a claim using the premises. In theory, the audience would not believe the claim without evidence of the claim, in the form of premises [21]. In the following example, the argument contains four components: one claim (in bold) and three premises (italic) [22]:

> "(1) **Museums and art galleries provide a better understanding about arts than Internet.** (2) *In most museums and art galleries, detailed descriptions in terms of the background, history and author are provided.* (3) *Seeing an artwork online is not the same as watching it with our own eyes*, as (4) *the picture*

online does not show the texture or three-dimensional structure of the art, which is important to study."

2.3.2 Traditional Machine Learning

Assessing argumentation quality was traditionally based on the evaluation of relevance, sufficiency, acceptability of premises [23] or categorizing fallacies [24, 25]. [26] argues those approaches create "ideal" models, and a gap can be observed between the argument quality modeling of those approaches and real-world arguments.

[27] presents an approach, using linguistic features, to model argument quality. The vast set of 23,345 handcrafted features consists, among others, of semantic density features, discourse and dialogue features, and syntactic property features. For example, the authors use sentence length, word length, specificity of the sentences, the Kullback-Leibler divergence, etc., as features to predict the quality of an argument. Modeling the prediction of argument quality as a regression task, they use 3 different algorithms: Linear Least Squared Error, Ordinary Kriging and Support Vector Machine (SVM) with a radial basis function kernel. Through feature selection, [27] identifies the 10 features most correlated with the annotated quality score where sentence length, the relative frequency of the root node within a sentence (syntactic feature) and lexical n-grams shine as the most correlated linguistic features.

UKP Labs pioneered the task of assessing argument quality by focusing on the relative convincingness of arguments and comparing pairs of arguments having the same stance on a topic [1]. Their initial and main contribution is a dataset of annotated argument pairs, *UKPConvArg1*, which we describe in chapter 3. In assessing argument quality by focusing on the relative convincingness of arguments, they propose 2 tasks which define the argument quality field and remain the basis of evaluation for any new state-of-the-art solution. The first task, the classification task, consists in predicting the most convincing argument of a pair of arguments. The second task, the ranking task, consists in ordering a list of arguments by their relative measure of convincingness. For the classification task, they first present a more traditional method: SVM using a set of rich linguistic features such as unigrams, bigrams, contextuality measures, readability measures, spellchecking, etc.

[28] propose to utilize scalable Gaussian Process Preference Learning (GPPL) to learn from noisy pairwise preferences ([1]'s collected pairwise annotations), producing a classifier which achieves significant improvement over [1]'s models. Compared to [1]'s two approaches

using either linguistic features or word embedding representations, [28] propose to leverage both linguistic features and word embedding representations as input for one single model. Each argument's vector representation consists of 32 010 linguistic features combined with Global Vectors for Word Representation (GloVe) word embeddings. Both feature sets are reused from [1], feeding those features to a scalable Bayesian preference learning model, outperforming [1]'s best performing model. [28], therefore, demonstrate the impact of combining rich linguistic features with embedding representation for the argument quality assessment task.

For the first task, the classification task, [28] leverages Gaussian process preference learning, which they reuse for the second task as this method is directly applicable to the argument quality ranking task. In fact, they argue this approach solves the disadvantages of classifier-based and permutation-based models, by learning a function which outputs a real-valued convincingness score. Therefore, their model, which is trained on pairwise preferences, takes argument features as input and can be used to predict pairwise labels or scores for individual arguments, and consequently, rankings. This makes for a more versatile solution. As a result, [28] reuses the model trained on the classification task, and outperforms both regression models by [1] on the argument quality ranking task. Therefore, they demonstrate the superiority of their approach compared to [1]'s approach, on both the classification and ranking task.

2.3.3 Neural Machine Learning

UKP Labs present a second approach to the classification task: a Bidirectional Long Short Term Memory (BiLSTM) using pre-trained GloVe. Therefore, they compare their first approach using handcrafted features to a word embedding representation of the arguments paired with a BiLSTM, a more modern solution gaining popularity for many natural language task at the time [1]. Evaluated on the task of predicting the most convincing argument of a pair, the SVM using linguistic features slightly outperformed the BiLSTM. However, [1] outlines a noticeable difference: the SVM using linguistic features requires heavier preprocessing prior to training compared to the BiLSTM, which might not justify the slight gain in performance.

As said previously, [1] also introduces a second task to the field of argument quality: the ranking of a given list of arguments on a topic, by their relative measure of convincingness. This is the task we focus on, in this work. [1] initially collects the dataset as pairwise annotations and, to extract rankings from those annotations, applies PageRank algorithm (explained more in details in chapter 3). Many other methods exist to aggregate rankings

from pairwise preferences, as we describe in section 2.2.2. Using PageRank, [1] generates a new dataset, *UKPConvArg1-Ranking*, from their first dataset. The new dataset consists of lists of arguments, ranked by their relative convincingness, each list of arguments being related to a topic. For this ranking task, [1] modifies the SVM using linguistic features and the BiLSTM, both used in the first task, by replacing the output layer with a linear activation function allowing to predict a quality score for every argument of a list, and then order the arguments according to their quality score. Thus, they address the second task as a regression task, which is as pointwise ranking.

[29] performs extensive feature selection over linguistic features and identifies 5 features which stand out: length ratio (length of individual argument over length of average argument) of words, length ratio of sentences, intersection with most common lemmas, stems ratios, percentage of long words and intersection with most common nouns ratio. Using only those 5 features as input to a feed-forward neural network, [29]'s solution delivers performances very close to [1]'s SVM using about 32, 000 features, while being much lighter to train.

[30] propose 2 supervised methods as well as non-supervised methods for the classification task. First, they present a Siamese BiLSTM and a Siamese model using the sum of token embeddings to represent an argument. The Siamese model with sum-of-token-embeddings performs best, and even outperforms [1]'s models. As another approach, [30] proposes to evaluate the similarity of an argument with Wikipedia texts, as a way of measuring its quality. A similarity score of an argument is evaluated by summing the similarity between the argument and each Wikipedia article. The similarity between the argument and a Wikipedia article is calculated using the dot product. Given a pair of arguments, the similarity score is calculated for each argument, and the argument with the highest similarity score is defined as more convincing. This method doesn't outperform [1]'s solutions.

For the ranking task, [18], which extends work from [30], proposes an architecture with an objective similar to RankNet [31]. The model is trained on pairwise annotations, and predict ranks, using a sum of word embeddings as representation of an argument. The model produces scores independently for each argument, normalizing the scores of argument pairs using the Softmax function. Trained on preference pairs of arguments, the model can then be evaluated on the classification task and the ranking task. On the *UKPConvArgRank* dataset, [18]'s solution sets the current benchmark for state-of-the-art performance, using the Spearman ranking metric. As another contribution, [18] proposes an alternative method to aggregate argument pairwise annotations into rankings: the WinRate (described in section 2.2.2) and compares it to [1]'s PageRank algorithm.

[32] proposes a solution to the argument classification task that stands out from other approaches. They further annotate the dataset *UKPConvArg1* by [1] with topic aspects for each argument. For example, for the topic *Ban plastic water bottles* with stance *No*, [32] annotates the argument:

> *The American Water companies are Aquafina (Pepsi), Dasani (Coke), Perrier (Nestle) which provide jobs for the american citizens.*

with the topic aspect *Economy* and, similarly, they annotate the argument:

> *If bottled water did not exist, more people would be drinking sweetened liquids because it would be the only portable drinks! People would become fat!*

with the topic aspect *Convenience and health*. [32] annotates latent topic aspects to leverage the assumption that arguments sharing the same topic aspect are more likely to demonstrate the same level of convincingness. They propose a BiLSTM-GCN for the classification task, consisting of a BiLSTM encoding a representation of each argument, and feeding its output to a Graph Convolutional Network (GCN) which updates the vector representation of each argument utilizing topic aspect information. This architecture yields a stronger performance than other approaches on [1]'s classification task. However, it cannot be directly compared to other approaches as it uses a modified dataset.

[2] presents a Siamese BiLSTM using word2vec embeddings, an architecture designed to build a representation of each argument of the pair and then compare each representation effectively. Each BiLSTM shares the same weights, allowing for each leg of the model to learn a quality representation of an argument in the pair, and both legs' output are compared using a cross entropy classification loss. They argue that their approach, compared to [28]'s solution, requires much lighter preprocessing steps. [28]'s rich linguistic features need heavy preprocessing and might not be suitable for certain languages. [2]'s Siamese BiLSTM, on the other hand, is not dependent on task-specific features, while achieving performance similar to [28]. [2] also contribute a new argument quality dataset to the field, *IBM-EviConv* which we describe in chapter 3.

Similarly to [28], [2] outlines how [1]'s proposed methods, SVM and BiLSTM, are constrained by the fact that if they are trained on pairs of arguments, they can provide pairwise inference only, making those models trained on the classification task not reusable for the ranking task. [2]'s solution, similarly to [28]'s, is trained on pairwise annotations and can provide pairwise as well as pointwise inference. In fact, for the task of ranking a list of arguments, [2]

reuses one leg of their Siamese BiLSTM from the first task. Using that Siamese BiLSTM's leg, they predict a quality score for each argument individually. Evaluated on the ranking dataset *UKPConvArgRank* by [1], [2]'s approach outperformed [28]'s scalable Bayesian preference learning model on the Pearson correlation metric and displays similar performance on the Spearman correlation metric, thus demonstrating a better overall performance. While the authors evaluated the Siamese BiLSTM on dataset *IBM-EviConv* for the classification task, the ranking task was not evaluated on it. This can be explained by the fact that the quality scores assigned to each individual argument, as part of the labeling process of the dataset, come from the predictions of one leg of the Siamese BiLSTM [2] trained on the pairs of arguments, from the first task. Therefore, the quality scores come from the model's predictions and are not collected from crowd annotators, like other datasets.

Pretrained language models

[7] demonstrates an approach outperforming previous solutions on the classification task, on the dataset published by [1], *UKPConvArg1*, by leveraging deep pretrained language models. They apply BERT (described in section 2.1) to the argument classification task. The model takes 2 arguments as input and uses a binary classification head to predict the most convincing argument. BERT's embeddings are fined-tuned on the argument classification task. [7] presents how BERT establishes itself as the state-of-the-art on [1]'s first task: predicting the most convincing argument out of a pair. [7] also presents a new argument quality dataset, *IBM-ArgQ*, which we describe in chapter 3.

To tackle the argument quality ranking task, both [7] and [6], which are initiatives by IBM research, use BERT with a regression head. BERT's strong performance on the first task justifies its use on the second task. To predict a quality score for each argument, they use BERT with 2 sequences as inputs: the topic and the argument. [7] reuses embeddings from the BERT classifier trained on the first task for the BERT Regressor, which they compare to a version of BERT with vanilla embeddings (not fined-tuned on any specific task). BERT with fined-tuned embeddings shows a stronger performance. Performances reported by both [7] and [6] demonstrate the effectiveness of using deep language models for the argument quality ranking task. [6] also contributes to the field by releasing the largest argument quality dataset to date, *IBM-ArgQ-Rank-30k* (see chapter 3).

Looking at state-of-the-art methods in argument quality evaluation, we can observe the effectiveness of using BERT to build embedding representations of the arguments. In fact, methods using BERT deliver the best performance for predicting the most convincing argu-

ment of a pair. However, on the task of ranking a list of arguments, point-wise approaches to learning-to-rank fall short. [18] demonstrates the benefits of using a ranking objective (RankNet). Based on those observations, in this work, we compare a pointwise, pairwise and list-wise approach to learning-to-rank on top of BERT, evaluating our approach on all the major argument quality datasets. Before presenting our solution in details in chapter 4, we thoroughly describe in chapter 3 each dataset included in our study. We compare our approach to state-of-the-art solutions on each of these datasets.

CHAPTER 3 DATASETS AND EVALUATION METRICS

In this section, we describe all major publicly available argument quality datasets released in the last 5 years.

3.1 Datasets

These datasets are used in all our experiments to compare our solution to the state-of-the-art. Descriptive statistics on these datasets are shown in Table 3.1 & 3.2. These datasets present differences in the way the arguments were collected and the way they were annotated.

Table 3.1 Statistics on the most common datasets for the argument quality evaluation task. PC stands for Pair Classification.

Dataset name	Number of arguments	Number of topics	Task	Source
UKPConvArg1Strict	11650 pairs of arguments	32	PC	Extracted from createdebate.com and convinceme.net
UKPConvArg1-Ranking	1052 arguments	32	Ranking	
IBM-ArgQ-Pairs	9100 pairs of arguments	22	PC	Actively collected arguments from crowds
IBM-ArgQ-Args	5300 arguments	22	Ranking	
IBM-EviConv	5697 pairs of arguments	69	PC & Ranking	Automatically retrieved Wikipedia sentences
IBM-ArgQ-Rank-30k	30000 arguments	71	Ranking	Actively collected arguments from crowds

UKPConvArgStrict & UKPConvArgRank

The first datasets used to compare our approach to state-of-the-art solutions is from UKP Lab. The UKP datasets contain arguments extracted from Web debate portals, where the proficiency of writing varies greatly. The collected arguments consist of claims and evidences [2]. A claim poses a statement about a subject, and an evidence is composed of facts presented in support of an assertion (more developed than the claim). Arguments collected take a stance on 16 different topics of various nature, from "Ban Plastic Water Bottles?" to "Christianity or Atheism" [1]. Annotated through crowdsourcing, each argument pair was evaluated by

Table 3.2 Detailed statistics on the arguments of the most common datasets for the argument quality evaluation task.

Dataset name	Arg Length			Topic Length			Mean Topic Arg Count
	Mean	Min	Max	Mean	Min	Max	
UKPConvArg1Strict	263	37	753	56	26	92	32
UKPConvArg1-Ranking							
IBM-ArgQ-Pairs	138	36	275	42	31	63	240
IBM-ArgQ-Args							
IBM-EviConv	189	60	495	34	20	55	26
IBM-ArgQ-Rank-30k	107	35	251	34	21	52	429

5 workers, who each had to choose the most convincing argument out of the pair (with a justification). These are referred to as pair annotations by the authors. Workers could also evaluate 2 arguments as equally convincing. The workers were instructed to be objective, not to judge the truth of the proposition and not express their opinion. The argument pair annotations are part of the dataset named *UKPConvArgStrict*. This dataset is used for the classification task of predicting the most convincing argument out of a pair of arguments. The authors also rank the arguments in order of convincingness for each topic, by computing a score for each argument from the pair annotations. To obtain this score, they build a graph representation where nodes represent arguments and directed edges indicate the most convincing argument: the target of the edge is the most convincing argument of the pair. PageRank is then used to rank all the arguments for each topic. This resulted in the dataset *UKPConvArgRank*. This dataset can be used for the task of ranking a list of arguments given a topic. Table 3.3 gives an example of an argument defending a stance on a topic, from the *UKPConvArgRank* dataset. To summarize, *UKPConvArgStrict* was annotated using pairs of arguments and those annotations needed an extra processing step to extract a pointwise quality score for each argument, creating *UKPConvArgRank*.

Table 3.3 Example of an argument for a topic given by [1].

Topic	Should physical education be mandatory in schools?
Stance	Yes
Argument	PE should be compulsory because it keeps us constantly fit and healthy. If you really dislike sports, then you can quit it when you're an adult. But when you're a kid, the best thing for you to do is study, play and exercise. If you prefer to be lazy and lie on the couch all day then you are most likely to get sick and unfit. Besides, PE helps kids be better at team-work.

IBM-EviConv

[2] introduced the dataset *IBM-EviConv*. The dataset consists of a set of evidence pairs extracted from Wikipedia, a heavily edited corpus, thus assuring a certain level of writing. Table 3.4 gives an example of a collected pair of evidences. The extracted arguments take a stance on 69 different topics of various nature, from "We should legalize prostitution" to "We should introduce universal health care" [2]. Contrary to UKP datasets, which contain claims and evidences, the dataset *IBM-EviConv* consists only of evidences (also referred to as premises, as seen in section 2.3.1). The reason given by the authors for such a decision is to counter an issue known with the UKP dataset: [1] demonstrated how a shallow feature such as the argument length performed very well to predict convincingness. This could be explained by the fact that an evidence is usually longer, providing more details compared to the claim which is more concise. This implies that an evidence could be considered more convincing than the claim by the model for the only reason of its length and not its content [2]. For these reasons, *IBM-EviConv* consists only of evidences of roughly the same length, posing a more challenging task. This forces a model to learn features from the argument's content to evaluate its convincingness instead of relying on shallow features like argument length.

As *IBM-EviConv* is annotated as a set of evidence pairs extracted from Wikipedia, an extra step is needed to rank arguments. To extract rankings from those pairwise annotations, [2] introduces a different ranking aggregation approach than [1]. They first train a Siamese BiLSTM, where parameters are shared by each BiLSTM, both connected through a Softmax layer on top. The Siamese BiLSTM is trained on the classification task using the pairwise annotations extracted from Wikipedia. Then, to infer rankings for each argument, one leg from the Siamese BiLSTM is used to generate a score for each argument. This allows for [2] to extract a pointwise score for each argument from the collected pairwise annotations.

Table 3.4 Example of a pair of collected evidences for topic *We should legalize same sex marriage* given by [2].

Topic	We should legalize same sex marriage.
Evidence #1	The California Supreme Court overturned California's ban on gay marriages on May 15, stating that depriving gays and lesbians of the same rights as other citizens is unconstitutional. (PRO)
Evidence #2	In his 2002 Senate campaign, Coleman pledged support for an amendment to the United States Constitution that would ban any state from legalizing same sex marriage. (CON)

IBM-ArgQ

Dataset *IBM-ArgQ*, proposed by [7], is another argument quality dataset which differs in the way the arguments were collected. Most previous argument quality datasets were collected from online debates. *IBM-ArgQ* was collected actively via a dedicated user interface, where contributors were guided to provide arguments per topic stance. Furthermore, the authors demonstrate how the arguments from the *IBM-ArgQ* dataset are more homogeneous in their length compared to *UKPConvArgRank* by [1]. They argue that this allows for a model to learn argument quality properties not related to argument length. Such properties would be more valuable to properly represent argument quality than a shallow feature like argument length.

A key point about the dataset is the way the arguments were labeled. The previously presented datasets were constructed by annotating argument pairs, and then a transformation step was needed to extract a ranking of arguments from that pairwise annotation [1,2]. In this case, *IBM-ArgQ* was built using two different labelling approaches: each individual argument was annotated with a pointwise quality score, and also, argument pairs were labeled (similar to previous approaches). [7] explains how they explore the two different labeling methods, and analyze how the resulting labels of each method compare. They demonstrate that each method yields consistent results. This resulted in, as per other initiatives, two datasets : one for the task of predicting the most convincing argument of a pair and the other for the task of ranking a list of arguments for a topic. However, in this case, both datasets were built directly from human annotators, avoiding a transformation step such as using PageRank like in [1].

IBM-ArgQ-Rank-30k

Dataset *IBM-ArgQ-Rank-30k* was proposed by [6]. For this dataset, the arguments are annotated directly with an individual score, without the need for argument pair annotations. Using crowd annotation, 30,497 arguments were collected from 280 contributors on 71 controversial topics. The arguments were annotated as a binary decision. For each argument, the annotators were asked if they would recommend a friend to use that argument or not. Each argument was annotated by 10 people. A continuous quality score, between 0 and 1, was then derived from these binary annotations. [6] used two different ways of deriving that score: the Weighted-Average (WA) and the Multi-Annotator Competence Estimation (MACE) probability.

MACE probability [6] uses the MACE probability as a scoring function to infer a quality score from crowd annotations. MACE is an unsupervised item-response generative model [33]. Given annotations, it predicts each label's probability. Moreover, a reliability score is estimated by MACE for each annotator and is used to weight the annotator's annotations. The MACE model maximizes the probability of observed data, maximizing the marginal data likelihood shown in equation 3.1 using Expectation Maximization (EM).

$$P(A;\theta,\xi) = \sum_{T,S}[\prod_{i=1}^{N} P(T_i) \cdot \prod_{j=1}^{M} P(S_{ij};\theta_j) \cdot P(A_{ij}|S_{ij},T_i;\xi_j)] \tag{3.1}$$

Where A is the matrix of annotations (A_{ij} corresponds to observed annotation i from annotator j), S is the matrix of spamming indicators (S_{ij} corresponds to annotator j's trustworthiness on annotation i), and T is the vector of true labels, noting that the true labels and the spamming indicators are unobserved. The annotator reliability score of an annotator j, or in other words, his trustworthiness, is represented as θ_j and ξ_j is a vector representing how an annotator behaves when he is not trustworthy, and he is spamming [33].

Weighted-Average [6] proposes the Weighted-Average as an alternative to MACE probability. It consists of an average of the annotations, weighted by annotator-reliability, similarly to MACE-P. Therefore, weighting each annotation by an annotator-reliability score diminishes the impact of non-reliable annotators on the final argument quality score. The annotator-reliability score is calculated similarly to [7]: using the average of the Cohen's kappa score (see section 5.2 for equation) between the annotator and other annotators (sharing at least 50 common argument judgments). Equation 3.2 shows how the weighted-average score is calculated, where P_a is the set of annotators who labeled argument a as positive and

N_a is the set of annotators who labeled argument a as negative. $Annotator\_Rel_i$ stands for the annotator-reliability score of annotator i.

$$\mathrm{WA}(a) = \frac{\sum_{Annotator_i \in P_a} Annotator\_Rel_i}{\sum_{Annotator_j \in N_a + P_a} Annotator\_Rel_j} \tag{3.2}$$

Both MACE-P and WA scores incorporate an annotator-reliability score to decrease the impact of non-reliable annotators. The process yields a dataset with a continuous quality score for each argument, aiming at the task of ranking the arguments for each topic. In this work, we focus on the score WA instead of MACE-P following [6] who also prioritize WA since they obtain better results using WA as a quality label instead of MACE-P.

3.2 Evaluation Metrics

In this section, we present the metrics used to evaluate our results, on the argument quality ranking task. We first define 4 metrics commonly used by state-of-the-art solutions to evaluate the performance of their approach: Accuracy, Pearson, Spearman and Kendall's Tau. Moreover, we introduce the NDCG metric to argument quality ranking, a metric commonly used in ranking tasks.

3.2.1 Accuracy

The accuracy is a metric of evaluation for the classification task. The accuracy measures the number of correctly predicted data points of all data points. As shown in equation 3.3, the accuracy is calculated using the ratio of the *True Positives (TP)* and *True Negatives (TN)* over the sum of the *True Positives (TP)*, *True Negatives (TN)*, *False Positives (FP)* and *False Negatives (FN)* [34].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

3.2.2 Pearson

The Pearson correlation metric ρ measures the linear relationship between two variables, X and Y. Equation 3.4 and 3.5 show how the correlation is calculated. Pearson's value is between -1 and 1. A value of 1 indicates a perfect positive relationship, a value of 0 indicates no relationship, and a value of -1 indicates a perfect negative relationship [35].

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{3.4}$$

Where $\text{cov}(X, Y)$ represents the covariance between variables X and Y, and σ_x and σ_y represents the standard deviation of variable X and Y, respectively.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{3.5}$$

Where n is the sample size. x_i, y_i are the individual sample points indexed with i. $\bar{x}$ and $\bar{y}$ are the sample mean and can be defined as shown in equation 3.6 and 3.7:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.6}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.7}$$

3.2.3 Spearman

Spearman's rank correlation coefficient is another measure of correlation between the ranking of two variables. It is equal to the Pearson metric between the rank values of two variables. Spearman's rank correlation coefficient identifies if the relationship between two variables is a monotonic function. Whereas Pearson compares the values of two variables, Spearman compares the ordering of the values of the two variables. Spearman allows identifying relationships between two variables that Pearson can't. Equation 3.8 shows how Spearman's rank correlation coefficient r_s is calculated [36]:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.8}$$

where d_i = the distance between the ranks of the variables x_i and y_i and n is the number of samples.

3.2.4 Kendall's Tau

Kendall rank correlation coefficient τ is a measure of ordinal association between two variables. In other words, it is a measure of rank correlation. Compared to Spearman, which is based on deviations, Kendall's Tau is based on concordant and discordant pairs, and is more insensitive to error, generally producing more accurate P-values with smaller sample sizes. Equation 3.9 shows how it is calculated [37]:

$$\tau = \frac{c-d}{c+d} = \frac{S}{\binom{n}{2}} = \frac{2S}{n(n-1)} \tag{3.9}$$

where n corresponds to the sample size, c is the number of concordant pairs and d is the number of discordant pairs in the ranks obtained from ranking the 2 variables, and $S = c-d$. If there are ties between the ranked variables, the equation 3.10 shall be used to calculate Kendall's Tau:

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}} \tag{3.10}$$

$$T = \sum_t t(t-1)/2 \tag{3.11}$$

$$U = \sum_u u(u-1)/2 \tag{3.12}$$

where t is the number of tied observations of variable X and u is the number of tied observations of variable Y.

3.2.5 NDCG

Most initiatives in the argument quality evaluation field used Pearson & Spearman to evaluate the ranking task. As we focus more on the ranking perspective of the task and less on predicting an absolute score for each argument, we employ the Normalized Discounted Cumulative Gain (NDCG) to evaluate our model's performance, as it is a metric commonly used for learning-to-rank. The NDCG can be defined as follows [38]:

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \tag{3.13}$$

Where IDCG is the Ideal Discounted Cumulative Gain. This corresponds to the Discounted Cumulative Gain (DCG) value of the best ranking of the elements. The Discounted Cumulative Gain (DCG) is calculated as follows :

$$\text{DCG} = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log(1+i)} \tag{3.14}$$

Where rel_i is the relevance value (as seen in section 2.2) of the argument at index i and n corresponds to the sample size.

In section 4.4, we report the NDCG@K metrics of our models on all datasets, for various values of K (5, 10 and 15). This allows to evaluate a ranking model's performance at identifying the top K most convincing arguments of a list. We also report our results using Pearson, Spearman and Kendall's Tau metrics, allowing to compare our results to other initiatives of the state-of-the-art on argument quality ranking.

3.3 Performance of State-of-the-art Models

We previously described all the major available argument quality datasets. Those datasets are used to evaluate our learning-to-rank model. To provide a basis for comparison, we present in tables 3.5 and 3.6 a summary of the performance of state-of-the-art approaches described in chapter 2, on the datasets presented in this chapter.

Table 3.5 shows the performance of solutions on the classification task. The dataset *IBM-Rank-30K* is not included in this table as the arguments collected are annotated directly with an individual score, without the need for argument pair annotations, therefore making this dataset not suitable for the classification task. For dataset *UKPConvArgStrict*, we can see that [28]'s Gaussian Process Classifier (GCP) outperforms [1]'s initial SVM and BiLSTM solutions. [2]'s Siamese BiLSTM matches the performance of [28]'s GCP, while not needing any heavy preprocessing. However, all these approaches are outperformed by BERT, evaluated on the dataset *UKPConvArgStrict* by [7]. On the dataset *IBM-EviConv*, [2] compares their Siamese BiLSTM model to [28]'s GCP. The Siamese BiLSTM model outperforms the GCP model by a considerable margin. On dataset *IBM-ArgQ*, [7] first evaluates [28]'s Gaussian Process Preference Learning (GPPL) model and then compares its performance to the model they present for the classification task: BERT. BERT outperforms the GPPL model by a significant margin. We can see in table 3.5 that BERT stands out from other solutions, being the top performing model on [1]'s first task on 2 out of 3 datasets. BERT has not been evaluated on the third dataset, *IBM-EviConv*.

Table 3.6 shows the performance of state-of-the-art models on the argument quality ranking task. This time, the dataset *IBM-EviConv* is not part of the table because this dataset wasn't used to evaluate any state-of-the-art approach on the ranking task. As described earlier, the quality score attributed to each individual argument was generated by one leg

Table 3.5 Performance of notable models of state-of-the-art solutions for the argument pair classification task, as described in chapter 2

Dataset	Model	Features	Accuracy
UKPConvArgStrict	SVM (RBF kernel) [1]	Linguistic features	0.780
	BiLSTM [1]	GloVe word embeddings	0.760
	Forward-Feeding Neural Network	5 Linguistic features	0.770
	GPC (Gaussian Process Classifier) [28]	Linguistic features + GloVe embeddings	0.810
	Siamese Sum-of-tokens [30]	GloVe embeddings	0.825
	Siamese BiLSTM [30]	GloVe embeddings	0.742
	Siamese BiLSTM [2]	word2Vec embeddings	0.810
	BERT Base Uncased for Binary Classification [7]	Fine-tuned BERT embeddings	**0.830**
IBM-EviConv	GPC (Model by [28], evaluated by [2])	Linguistic features + GloVe embeddings	0.670
	Siamese BiLSTM [2]	word2Vec embeddings	**0.730**
IBM-ArgQ	GPPL (Gaussian Process Preference Learning, Model by [28], evaluated by [7])	Linguistic features + GloVe embeddings	0.710
	BERT Base Uncased for Binary Classification [7]	Fine-tuned BERT embeddings	**0.800**

of a Siamese BiLSTM trained on the pair annotations, making this gold standard different from other datasets. This might explain why no attempt to evaluate the ranking task on *IBM-EviConv* was published. On the dataset *UKPConvArgRank*, [1]'s SVM with linguistic features and BiLSTM with GloVe word embeddings are initially outperformed by [28]'s GPPL model using a combination of linguistic features and GloVe word embeddings. These 3 approaches are outperformed, on the ranking task, on *UKPConvArgRank* dataset, by [2]'s Siamese BiLSTM from which only one leg is used to predict a score for each argument. [18] outperforms [2]'s Siamese BiLSTM using a Sum-of-Word Embeddings Feed Forward Neural Net (SWE+FFNN) with GloVe word embeddings. BERT is evaluated on *UKPConvArgRank* by [7] and [6]. Surprisingly, BERT in [7], which takes only the argument as input, outperforms BERT in [6] which takes the argument and the topic as input. BERT in [7] achieves the highest performance using the Pearson metric on the dataset *UKPConvArgRank*, but doesn't outperform [18]'s model using the Spearman metric. On the dataset *IBM-ArgQ*, only BERT is evaluated for the ranking task, making it the state-of-the-art on this dataset. Finally, on the *IBM-Rank-30K* dataset, [6] compares different configurations of BERT to a Support Vector Regression (SVR) and BiLSTM models, which are both outperformed by BERT models. The

best performing configuration of BERT is BERT with fined-tuned embeddings, taking the concatenated topic and argument as input. As for [1]'s first task, BERT stands out as one of the best performing models on the second task, the ranking task, across datasets. The only exception is found when ranking arguments on dataset *UKPConvArgRank*, where [18]'s solution outperforms BERT for the Spearman metric.

Table 3.6 Performance of notable state-of-the-art for the argument quality ranking task, as described in chapter 2

Dataset	Model	Features	Pearson	Spearman
UKPConvArgRank	SVM (RBF kernel) [1]	linguistic features	0.351	0.402
	BiLSTM [1]	GloVe word embeddings	0.270	0.354
	GPPL [28]	Linguistic features + GloVe embeddings	0.440	0.670
	Siamese BiLSTM [2]	word2Vec embeddings	0.470	0.670
	SWE+FFNN [18]	GloVe embeddings	0.480	**0.690**
	BERT [7]	Argument	**0.490**	0.590
	BERT [6]	Argument	0.450	0.630
	BERT [6]	Argument + Topic	0.460	0.620
IBM-ArgQ	BERT [7]	Argument	**0.420**	**0.410**
IBM-Rank-30K (Predictions on WA score)	SVR with RBF Kernel [6]	BOW	0.320	0.310
	BiLSTM [6]	GloVe word embeddings	0.440	0.410
	BERT Vanilla [6]	Argument	0.480	0.430
	BERT Fined-Tuned [6]		0.510	0.470
	BERT Fined-Tuned [6]	Topic + Argument	**0.520**	**0.480**

In this chapter, we described the most notable datasets in the field of argument quality, and their particular features. Since we intend to use them as a common basis for the evaluation of our models and state-of-the-art solutions, it was important to first identify how the datasets differ from each other. It is also noteworthy that the performance of state-of-the-art methods varies from one dataset to another, suggesting that the differences in how the annotations were collected, and how the scores were inferred, might impact the argument quality measure computed for each dataset. We analyze in chapter 5 how the quality measure could be unified in a single quality score for all datasets.

CHAPTER 4 LEARNING-TO-RANK FOR ARGUMENT QUALITY RANKING

In this chapter, we present our approach to argument quality ranking. This approach is based on learning-to-rank methods and BERT as building blocks. We then evaluate our model on the 4 different argument quality datasets presented in chapter 3, and we show how a learning-to-rank approach based on BERT performs compared to state-of-the-art solutions, for each dataset.

4.1 BERT Learning-to-rank Model

Focusing on the ranking task, the learning objective is, given a Topic T, a stance S and a list of arguments A_1, A_2, ... , A_n, to assign a rank to each argument A_i from the most convincing argument to the least convincing argument. We propose to use TFR-BERT (TensorFlow Ranking BERT) [3], a learning-to-rank approach paired with BERT. In this architecture, BERT [13], which has proven to be very efficient in learning text representations, is used as a building block responsible for learning a representation of each argument. A ranking head is used on top of BERT, allowing to apply a ranking loss function (see the section 4.2) over multiple arguments at once. This neural approach to learning-to-rank is implemented using the TF Ranking library [39].

4.1.1 Input Representation

The ranking model needs to be able to grasp the quality of an argument with respect to a topic. The BERT module is responsible for building a representation demonstrating the association between the argument and the topic. Each argument's text is concatenated to its respective topic's text, in a typical BERT pair representation: [CLS] *Topic* [SEP] *Argument* [SEP]. The special token [CLS] indicates the start of a sequence and [SEP] is the separator between the topic and the argument (and also marks the end of the sequence).

4.1.2 Architecture

As shown in figure 4.1, for each argument, the BERT module takes a topic & argument concatenated sequence and outputs the hidden units of the [CLS] token of the last layer. The pooled outputs of each topic & argument sequence, for every argument in the list to rank, are fed into a dense layer which acts as a scoring function. The scoring function learns to

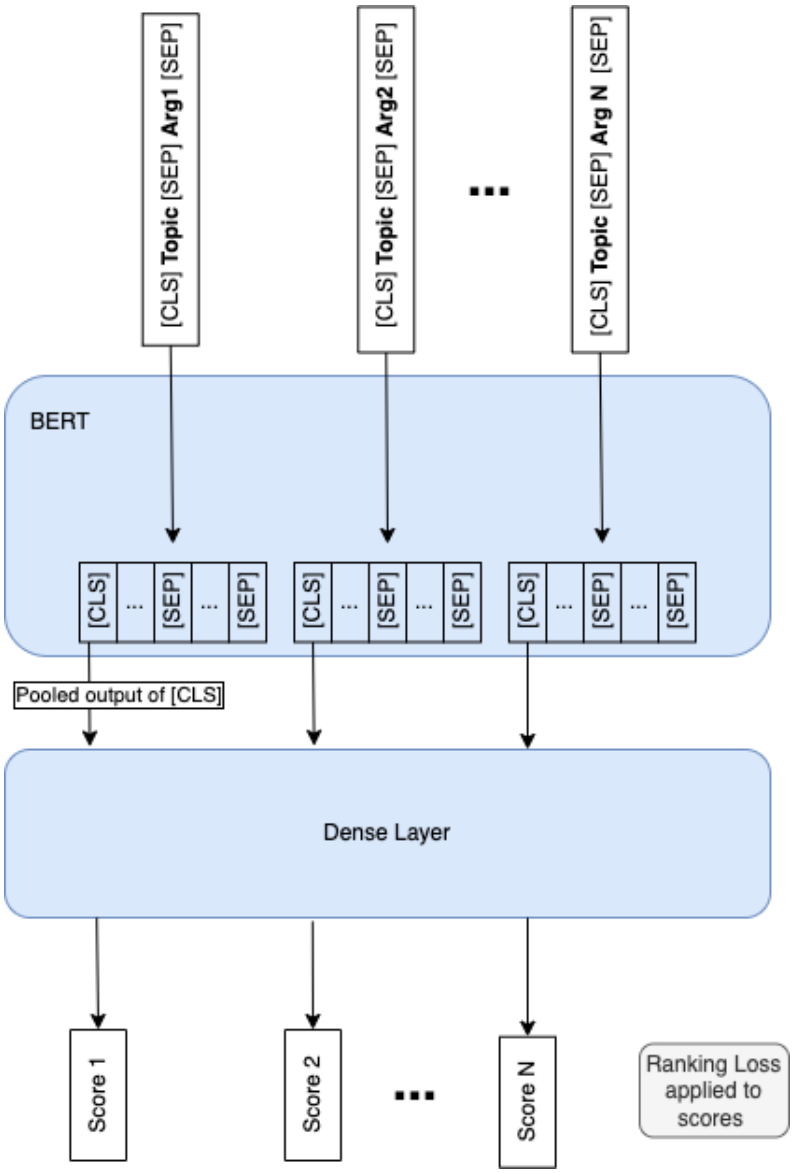

Figure 4.1 Architecture of the BERT Ranking Model based on [3].

associate a score to each argument. A ranking loss function is applied to the scores generated by the neural ranking function (dense layer in figure 4.1) and is used to update the model's weights. The loss function used determines how many arguments are considered at once when calculating the loss for back-propagation over the model's weights.

4.2 Ranking Loss Functions

In this work, we compare the performance of 3 types of ranking losses, and introduce list-wise ranking loss functions to the argument quality evaluation task. When training a ranking model, the loss function can either be applied to the arguments individually (pointwise loss), by pairs (pairwise loss) or altogether (list-wise loss). For pointwise losses, the arguments are considered independently. Therefore, pointwise losses are expected to optimize predicting a score close to label without regard to the ordering. Pairwise losses use argument pairs to calculate the loss. Since pairwise losses treat every pair with the same weight independently of their position in the list, they usually display a lower performance at the top of the rankings and improve at tail level. List-wise losses consider the order of the whole list of arguments, and therefore directly optimize the ranking of the arguments. Table 4.1 shows the specific loss functions we explore for each type of loss. We choose ranking losses which apply for graded relevance labels. Indeed, the argument quality ranking task consists of predicting graded relevance labels for each argument. Therefore, we explore every loss function compatible with graded relevance labels, except for the *Gumbel Approx NDCG Loss*, the *Unique Softmax Loss* and the *Pairwise Soft Zero One Loss*. The time needed for the training of TFR-BERT being considerable, multiplied by the 4 datasets we work with, we discard these losses as they are special cases of the loss functions we already explore.

Table 4.1 Ranking loss functions presented in this section.

Loss Type	Loss Function
Pointwise	Mean Squared loss
Pairwise	Pairwise Hinge Loss
	Logistic Loss
Listwise	List MLE Loss
	Softmax Loss
	Approx NDCG Loss

4.2.1 Mean Squared Loss

The *Mean Squared Loss*, also known as Mean Squared Error (MSE), measures the average of the squares of the difference between the predicted scores (s) and the scores from the ground truth (y) [39]. The *Mean Squared Loss* is a pointwise loss function as each score of a list of ranked items is compared individually to its ground truth. Equation 4.1 defines how the *Mean Squared Loss* is calculated over the predicted list of scores s, using the list of scores

y as ground truth for the ranking task. n corresponds to the sample size.

$$\mathcal{L}(\{y\}, \{s\}) = \frac{1}{n} \sum_{i}^{n} (y_i - s_i)^2 \tag{4.1}$$

4.2.2 Pairwise Hinge Loss

The *Pairwise Hinge Loss* is, as the name would suggest, a pairwise loss, and is based on the difference in relevance between the arguments of each pair of the list to rank. Equation 4.2 defines how this loss is calculated. Given a pair of arguments, where $argument_i$'s rank (y_i) is higher than $argument_j$'s rank (y_j) according to the ground truth, the *Pairwise Hinge Loss* evaluates a correctly ordered pair of arguments as a loss of 0 if the difference between the predicted rank of $argument_i$ (s_i) and predicted rank of argument $argument_j$ (s_j) is at least one. Otherwise, the loss is linearly increased with s_i - s_j [39].

$$\mathcal{L}(\{y\}, \{s\}) = \sum_{i} \sum_{j} I[y_i > y_j] \max(0, 1 - (s_i - s_j)) \tag{4.2}$$

Where I is the indicator function, which takes value 1 if the condition inside the brackets is met, 0 otherwise.

4.2.3 Pairwise Logistic Loss

The *Pairwise Logistic Loss* is also calculated using the order of pairs of arguments in the ranked list. Equation 4.3 shows how the loss is calculated over predicted list of scores s, knowing the ranked list of scores y as ground truth [39].

$$\mathcal{L}(\{y\}, \{s\}) = \sum_{i} \sum_{j} I[y_i > y_j] \log(1 + \exp(-(s_i - s_j))) \tag{4.3}$$

Where I is the indicator function, which takes value 1 if the condition inside the brackets is met, 0 otherwise.

4.2.4 List MLE Loss

Part of the list-wise ranking losses, the *List MLE Loss* function utilizes the Maximum Likelihood Estimation of the Plackett-Luce model, which defines a probability distribution on permutations of objects, also known as permutation probability [14].

$$\mathcal{L}(\{y\}, \{s\}) = -\log(P_s(\pi_y)) \tag{4.4}$$

Where $P_s(\pi_y)$ is the Plackett-Luce probability of a permutation π_y conditioned on the list of scores s, which can be defined as follows:

$$P_s(\pi_y) = \prod_{i=1}^{n} \frac{s_{\pi^{-1}(i)}}{\sum_{j=i}^{n} s_{\pi^{-1}(j)}} \tag{4.5}$$

Where $\pi^{-1}(i)$ defines the object at rank i in permutation π (ranked list). $P_s(\pi_y)$ represents the likelihood of permutation π_y knowing the list of scores s. This allows for a very intuitive loss as the highest probability is assigned to the permutation in descending order of scores (ranked list according to s) and, similarly, the lowest probability is assigned to the permutation in ascending order of scores [14].

4.2.5 Softmax Loss

The *Softmax Loss* function, which is a list-wise loss, computes the Softmax cross-entropy over predicted list of scores s and ranked list of scores y, used as ground truth, as shown in equation 4.6.

$$\mathcal{L}(\{y\}, \{s\}) = -\sum_{i} y_i \cdot \log\left(\frac{exp(s_i)}{\sum_j exp(s_j)}\right) \tag{4.6}$$

4.2.6 Approx NDCG Loss

Part of the list-wise ranking losses, the *Approx NDCG Loss* function is an approximation of the NDCG ranking metric, which is presented in section 3.2.

$$\mathcal{L}(\{y\}, \{s\}) = -\frac{1}{\mathrm{DCG}(y, y)} \sum_{i} \frac{2^{y_i} - 1}{\log_2(1 + \mathrm{rank}_i)} \tag{4.7}$$

Where $\mathrm{DCG}(y, y)$ is the Discounted Cumulative Gain, and can be calculated using equation 3.14, as explained in section 3.2.5. rank_i is a differentiable approximation of the non-differentiable ranking function used to calculate the NDCG metric, an approximation based on the logistic function, as shown in equation 4.8 [40].

$$\mathrm{rank}_i = 1 + \sum_{j \neq i} \frac{1}{1 + \exp\left(\frac{-(s_j - s_i)}{\mathrm{temperature}}\right)} \tag{4.8}$$

4.3 Methodology

4.3.1 Transforming Scores into Ranks

For all the datasets mentioned in chapter 3, the quality score is an absolute value and cannot be directly used with a learning to rank model. We must first sort all the arguments for a given topic by quality score, from lowest to highest. From that sorted list of arguments, we attribute a relevancy rank to each of the argument, with the highest rank assigned to the highest score. To transform scores into ranks, we use the function rankdata from the Scipy library [1]. To deal with arguments with tied scores, we choose the strategy 'dense' to transform the scores into ranks in a way that limits the range of the rank values. This implies only a single rank value is assigned to arguments with tied scores, and ensures the ranking model is able to learn to rank two arguments as equal if they have the same quality score.

4.3.2 Training Parameters

Maximum Sequence Length

The maximum length of the sequence passed as input to the BERT module is calculated for every dataset. Looking at the distribution of the argument length for a dataset, we use the 95th percentile as the maximum sequence length, ensuring 95% of the set of arguments isn't truncated (sequence length including the topic & the argument combined as seen in the section 2.2). For the remaining 5%, corresponding to the longest arguments, the total sequence is truncated to the fixed maximum length. The reason for this decision is to facilitate the training. Given that the remaining arguments are the longest ones, keeping them whole usually increased drastically the sequence length and caused the model to be heavier to train. We decided the performance trade-off of losing the truncated tokens of 5% of the data wasn't worth it. For dataset *UKP Rank*, the maximum sequence length is set to BERT's maximum sequence length (512) as the 95th percentile exceeds this value for this dataset. We show in table 4.2 the different maximum sequence length chosen for each argument quality ranking dataset.

Argument Batches

During the training phase, memory limits did not allow fitting the whole list of arguments for a topic. Thus, for each topic in our datasets, we divide the list of arguments into smaller lists of 12 arguments as shown in table 4.3, the same list size used by [3]. At inference time,

[1] https://docs.scipy.org/doc/scipy/reference/generated/ scipy.stats.rankdata.html

Table 4.2 Maximum sequence length values for each argument quality ranking dataset.

Dataset	Maximum sequence length
UKPConvArgStrict & UKPConvArgRank	512
IBM Evi	357
IBM ArqQ Rank	261
IBM Arg 30K	227

however, full lists of arguments are fed to the model for predictions, ensuring the model is evaluated on unmodified data (test set).

Tied Ranks and Scores

The methodology used to divide a list of arguments into smaller lists of 12 arguments must take into account the number of arguments with the same score, otherwise it affects the training of the model. In fact, many arguments have the same score, especially the most convincing ones and the least convincing ones, resulting in equal ranks. Feeding the model with batches of arguments with equal ranks would result in poor training. Consequently, we divide the list of arguments in such a way that each batch has arguments of rank values well spread across the rank range. To do so, we divide the list of arguments, sorted by convincingness, into 12 slices. Each batch_i takes argument i of every slice, generating uniform batches of size 12, while ensuring no argument overlap between batches. In other words, every list of arguments fed to the model for training contains strong and poor quality arguments, as well as arguments considered relatively convincing. This allows for effective learning of the ranking function.

Table 4.3 Training parameters of TFR-BERT for the argument quality ranking task.

Loss Function	learning rate	EPOCH	optimizer	train batch size	dropout rate	list size
MSE Loss	1e-5	2	adam	6	0.1	12
Hinge Loss	1e-6	3				
Logistic Loss						
List MLE						
Softmax Loss						
Approx NDCG Loss						

Training & Validation Loss

During the training of model TFR-BERT on the argument quality ranking task, both training and validation loss are monitored. Monitoring the validation loss ensures the model does not overfit on the training data and is able to generalize to new data examples. Therefore, the final trained model is chosen, through model selection, using the checkpoint corresponding to the lowest point of the validation loss. As an example, figure 4.2 shows the training and validation loss of TFR-BERT trained using *pairwise logistic loss* on dataset *UKP ConvArgRanking*. The training of the model took up to 7 hours.

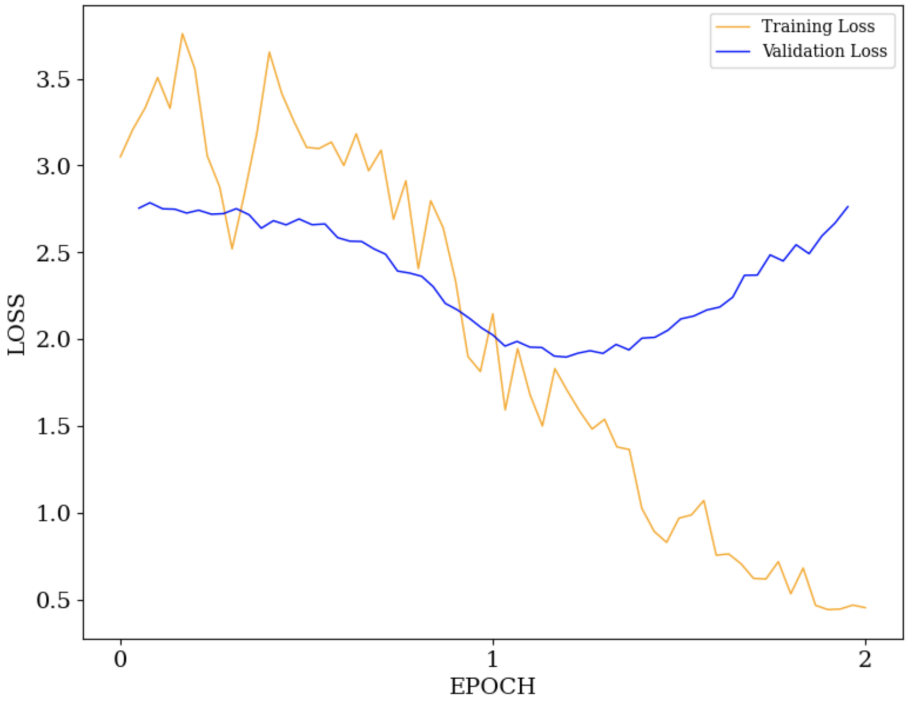

Figure 4.2 Training loss and Validation loss during the training of TFR-BERT using *pairwise logistic loss* on dataset *UKP ConvArgRanking*.

Ensemble TFR-BERT

[3] demonstrates how an ensemble approach to TFR-BERT, combining multiple ranking losses, can improve predictions. We use the same approach for the task of argument quality, combining the predictions of multiple versions of TFR-BERT, each trained using a different ranking loss. For each prediction, we average the list of scores predicted over the different versions of TFR-BERT. This increased the model's performance considerably, as shown in section 4.4.

4.4 Results

In this section, we evaluate the model TFR-BERT, a learning-to-rank approach paired with BERT, presented in section 4.1, on all major argument quality evaluation datasets available, which we described in chapter 3. For each dataset, we compare the performance of TFR-BERT to the state-of-the-art solution on that very dataset. The performance metrics used for this evaluation are Pearson, Spearman, Kendall's Tau and the NDCG@K for values of 5, 10 and 15. We further explore how each metric evaluates the model's performance from a different aspect. Also, to better visualize the variation in performance from one metric to another, we analyze the predictions of the models, looking at the top N most convincing arguments predicted by the model and comparing them to the gold standard.

When a division into train, validation and test sets was not provided in the dataset, we divided it as 20% of the topics assigned to the test set, 20% assigned to the validation set and the remaining to the train set. Table 4.4 shows descriptive statistics on the train, validation and test sets. It is important to note that the test set never contains a topic already seen by the model during training. For reproducibility purposes, we provide all the datasets as lists of ordered arguments following the methodology described in section 4.3.1.

Table 4.4 Division of datasets into train, valid and test sets.

Dataset	Train		Valid		Test	
	Topic	Args	Topic	Args	Topic	Args
UKP Rank	18	602	7	222	7	228
IBM Evi	36	6632	12	2006	21	2756
IBM ArqQ Rank	12	2625	5	1586	5	1087
IBM Arg 30K	49	20974	7	3208	15	6315

4.4.1 UKP Rank

We first evaluate the performance of TFR-BERT on the *UKP Rank* dataset for the argument quality ranking task, comparing different loss functions. Table 4.5 shows Pearson, Spearman, Kendall's Tau as well as the NDCG@K metrics for every model. The first 3 metrics give a measure of how well the model ranks all the arguments of the list. The NDCG@K values show how the model performs at outlining the top K most convincing arguments. We can see in table 4.5 how TFR-BERT compares to BERT on *UKP Rank*: the majority of TFR-BERT variants (TFR-BERT trained with specific loss function) outperform BERT across many metrics, including models trained with pointwise, pairwise and list-wise losses.

The Ensemble TFR-BERT, which combines models trained with *MSE*, *Hinge*, *Pairwise Logistic* and *Approx NDCG* losses respectively, is the best performing variant of TFR-BERT. Comparing it to the state-of-the-art [18]'s Sum-of-Words-Embeddings with Feed Forward Neural Network, we find that Ensemble TFR-BERT performs similarly to their solution for Pearson, Spearman and Kendall's Tau metrics. Ensemble losses and the *Pairwise Hinge* loss are the best performing TFR-BERT variants according to NDCG@K metrics, outperforming BERT by a significant margin. Unfortunately, the NDCG@K metrics were not provided by [18] when evaluating their solution.

Table 4.5 Evaluation of TFR BERT using different ranking losses on *UKP Rank* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
Pointwise	BERT	0.44	0.56	0.40	0.53	0.62	0.68
	TFR-BERT MSE Loss	0.45	0.68	0.51	0.59	0.67	0.72
Pairwise	TFR-BERT Hinge Loss	0.44	0.60	0.46	**0.63**	**0.72**	0.75
	TFR-BERT Logistic Loss	0.38	0.59	0.45	0.43	0.57	0.61
	State-of-the-art: Sum-of-Words-Embeddings + FFNN	**0.48**	**0.69**	**0.52**	-	-	-
List-wise	TFR-BERT Softmax Loss	0.40	0.67	0.51	0.49	0.61	0.66
	TFR-BERT List MLE	0.36	0.61	0.45	0.36	0.54	0.60
	TFR-BERT Approx NDCG Loss	0.47	0.59	0.44	0.54	0.66	0.69
Mix	**TFR-BERT Ensemble Losses**	**0.48**	0.68	0.51	0.60	**0.72**	**0.77**

Prediction of the Top 5 arguments

Looking at the performance metrics in table 4.5, we can see that TFR-BERT outperforms BERT across NDCG@K metrics, for the majority of the loss functions presented, reinforcing the interest of learning-to-rank methods for argument quality ranking. As a better performance according to NDCG@5 metric implies better capability at outlining the 5 topmost convincing arguments of a list, we decided to visualize how that translates into predictions on one topic of the test set. We chose a random topic from the test set: *Is the school uniform a good or bad idea*, with the stance *good*. Table 4.6 shows the top 5 most convincing arguments on that topic, according to *UKP Rank*'s gold standard. We can see that the scores of the top 5 are very close, making the task of comparing top arguments a very difficult task. From there, we can compare the predictions of every model presented in table 4.5 to this gold standard, comparing the top 5 ranked arguments. To help visualize, when an

argument predicted by a model is part of the top 5 according to the gold standard, it is outlined with a bold font in the table. We start by analyzing BERT's predictions, using this model as baseline. Table 4.7 shows that BERT's predicted top 5 arguments contains none of the arguments of the top 5 according to the gold standard. However, we can also notice that the predicted scores are very close to the scores in the gold standard, implying strong performance according to Pearson metric, which allows evaluating a model's performance at predicting the right score for each argument, without consideration of the ordering of the arguments. This demonstrates very well how BERT was trained with a pointwise ranking objective, or in other words, a regression task. This generates a model good at predicting an individual quality score for each argument but showing weaknesses when ordering arguments by their relative measure of convincingness, therefore ranking them.

Table 4.6 Ground Truth of top 5 arguments for the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset.

	Top N Arguments	Score
0	According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.	1.0000
1	I think it's good within certain limits. I went to a school with a uniform, and it was far less stressful than non-uniform college. I'd argue that it's a leveler- prevents people from showing off material wealth/ making others feel bad for not having 'cool' stuff. But it can be taken too far. By the end, we weren't allowed coloured socks, which was idiotic.	0.9990
2	That's really good idea. As i remember every morning i though what was better to wear? It was really problem, i spent quiet a lot of time. I asked my parents to buy new clothes for me, it was happened not rare. I know that not everyone thought as me, but it much better if the school has own uniform and everybody has to follow it. First it looks very good, smart. Secondly there is no envy that somebody have really nice skirt or jeans. Every pupil is the same and it would be easy to study, to not think about another things!!	0.9989
3	1. It makes everyone equal - if children can wear what they want some children will teased and feel less equal to their peers around them vs. uniforms 2. Okay Look school is for learning not how you look and dress but maybe in some levels it matters and most people that go to schools that don't have uniforms take like about 1 hr just to find their pants or shirt I mean really?? When you have a uniform it takes less than 10 min just to take it out and put it on and	0.9984
4	yas,of course . School uniform is important 1.school uniform is a logos for our school 2.to remind us that we are part of the school 3.and if we use the uniform basically student used to think what are they gonna do to, is it positive or negative 4.in the morning we should use our uniform and if were not use our uniform the teacher give us a punishment and from that we can learn to be a discipline student 5.if we go out from the school than the teacher will see we used the school uniform so people will know that we from that school thankyou	0.9979

Table 4.7 Ranking of top 5 arguments by BERT model for the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	I think school uniform is a good idea. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	0.9673	0.9972
1	Uniforms allow an equal and fair social status only based on personality and not looks. I do half to admit wearing what you want is fun and creative but its only fun if everyone can do it and for some children thats not the case and they cant afford to live up to their peers standards so uniforms would make social life much easier and it would give a more mature look to the school.	0.9636	0.9976
2	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.9633	0.9978
3	I believe that the school uniform is a good idea because school uniform improve student attendance and student doesn't spend a lot of time to choosing and buying clothes for school, espeacilly girls. That is why they can use more time to study.	0.9614	0.9899
4	I think that such a policy is a good idea. Uniforms make students equal on an economic level This can be prevent envy and jealousy.	0.9610	0.9896

We previously established in table 4.7 that BERT wasn't able to predict any of the top 5 arguments from the gold standard. We now visualize predictions from TFR-BERT to see how this model performed on the task of outlining the top 5 most convincing arguments, on the same topic, *Is the school uniform a good or bad idea* with the stance *good*. We are particularly interested in identifying how the choice of a ranking loss function, either pointwise, pairwise or list-wise, used during training, impacts the model's performance at outlining the topmost convincing arguments of a list.

Pointwise Loss Looking at table A.1, we can see that TFR-BERT trained with Mean Squared Loss did predict the first top argument in the gold standard. This exemplifies the better performance than the one of BERT at outlining the top 5 most convincing arguments. However, looking at the predicted scores from TFR-BERT trained with Mean Squared Loss, we can see a greater gap between the predicted scores and the gold standard scores, compared to BERT.

Pairwise Losses Tables A.2 and A.3 show the predictions of both variants of TFR-BERT trained with pairwise loss functions: the *Pairwise Hinge Loss* and the *Pairwise Logistic Loss*. The lists of the 5 most convincing arguments predicted by those 2 models are very

similar. They are identical for the first 4 arguments, and their fifth predicted argument is different. While those 2 variants of TFR-BERT present similar top 5 arguments ranking, the quality score predicted for each argument is quite different. TFR-BERT trained with *Pairwise Logistic Loss* predicts scores much closer to the scores in the gold standard. On the other hand, looking at the predicted scores from TFR-BERT trained with *Pairwise Hinge Loss*, we can see a greater gap between the predicted quality scores and the gold standard scores.

List-wise Losses We now analyze how list-wise ranking losses performed on the same topic, looking at the top 5 most convincing predicted arguments. Table A.4 shows the predicted top 5 arguments by TFR-BERT trained with *Softmax Loss*. Similarly to variants of TFR-BERT trained with *Pairwise Logistic Loss* and *Pairwise Hinge Loss*, TFR-BERT trained with *Softmax Loss* predicted one argument of the top 5 according to gold standard as part of its own top 5. We can also see in table A.5 that the predicted top 5 arguments by TFR-BERT trained with *Approx NDCG Loss* contains one argument of the top 5 in the gold standard. TFR-BERT trained with *List MLE Loss* stands out on this topic. As we can see in table 4.8, TFR-BERT trained with *List MLE Loss* predicted 2 arguments of the top 5 according to gold standard as part of its own top 5, showing slightly better capabilities at outlining the topmost convincing arguments of a list, for that particular topic.

Table 4.8 Ranking of top 5 arguments by TFR-BERT model using List MLE Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.9692	1.0000
1	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.9668	0.9978
2	**1. It makes everyone equal - if children can wear what they want some children will teased and feel less equal to their peers around them vs. uniforms 2. Okay Look school is for learning not how you look and dress but maybe in some levels it matters and most people that go to schools that don't have uniforms take like about 1 hr just to find their pants or shirt I mean really?? When you have a uniform it takes less than 10 min just to take it out and put it on and**	0.9043	0.9984
3	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.9016	0.9974
4	I believe that the wearing of the school uniform should be encouraged because it reminds each child that they are equal (at least in school). It also shows unity and children feel included and it helps them to work as a team.	0.8995	0.9960

Ensemble Losses Ensemble TFR-BERT is an average of multiple TFR-BERT trained with different loss functions, as explained in 4.3.2. Its top 5 predicted arguments on the topic *Is the school uniform a good or bad idea* with the stance *good* are shown in table A.6. Ensemble TFR-BERT predicted one argument from the top 5 according to gold standard.

While there were some differences in the predictions of the top 5 from one loss function to another when using TFR-BERT, we note that every TFR-BERT except for TFR-BERT trained with *Approx NDCG Loss*, did predict the most convincing argument from the gold standard in its top 5 arguments. We can't say the same for BERT, which failed to outline this argument in his top 5, nor any argument of the top 5 of the gold standard. Indeed, the argument

> *According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute*

to behavior problems and safety issues in the classrooms and in the hallways of schools today.

which takes a pro stance on the topic *Is the school uniform a good or bad idea*, is the strongest argument of the list according to the gold standard. This high quality score explains itself when looking at the argument more closely: the argument is clearly developed, well formulated and doesn't contain any spelling mistakes. A model not able to identify this argument shows weaknesses in its ranking capabilities.

Prediction of the Bottom 5 arguments

In the previous section, we analyzed the prediction of the top 5 most convincing arguments on a specific topic of dataset *UKP Rank* for different models. To be thorough, we analyze the prediction of the 5 less convincing arguments on the same topic: *Is the school uniform a good or bad idea*, with the stance *good*. In other words, we analyze the bottom 5 arguments of the ranked list predicted by each model and compare it to the gold standard. Table 4.9 shows the 5 less convincing arguments according to the gold standard

Table 4.9 Ground Truth of the ranking of bottom 5 arguments for the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset.

	Top N Arguments	Score
29	Good idea for separate student from other people for at least increase garment job. Why we use soldier uniform? For separate from civil. Why terrorist not use soldier uniform? For harmonious with civil and easy to attack enemy.	0.8999
30	school uniform does no harm to students life emotions now as for the point of expressing oneself imagine one bully wearing shirt,tie,pant all neatly ironed and well polished shoes wont he look good.remember dressing sense also is a part of ur interveiw	0.8960
31	Means you don't have to worry about what you hve to wear! Less awkward when people say wear school uniform and you wear mufti...	0.8933
32	Who in their right mind wants to get rid of Catholic school girl outfits?	0.0005
33	This is very. Bad as the uniforms are also cost effective	0.0000

Table 4.10 and table 4.11 show the prediction of the 5 less convincing arguments on the topic *Is the school uniform a good or bad idea* (the stance being *good*) by BERT and TFR-BERT trained with MSE loss respectively. Both outline 4 arguments as part of the bottom 5 according to gold standard, as part of their own bottom 5, demonstrating equally strong performance at outlining the less convincing arguments of a list. A noticeable difference

between BERT's and TFR-BERT's predictions is the predicted scores for the less convincing arguments of the gold standard. Arguments *Who in their right mind wants to get rid of Catholic school girl outfits?* and *This is very. Bad as the uniforms are also cost effective* are labeled with a score of 0.0005 and 0.0000 respectively. BERT's predicted scores for those 2 arguments are 0.5338 and 0.8055 respectively, which is very far from the gold standard. TFR-BERT's predicted scores for those 2 arguments, on the other hand, are much lower: 0.1911 and 0.3763, and therefore are closer to the gold standard. The analysis on this topic shows that TFR-BERT's prediction of the 5 less convincing arguments is more accurate than BERT's, when looking at the predicted scores of each argument.

Table 4.10 Ranking of bottom 5 arguments by BERT model for the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the bottom 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
29	it is good to follow to proper school code and the right to express emotions is right but not necessary as we have come school for learning	0.8342	0.9343
30	**school uniform does no harm to students life emotions now as for the point of expressing oneself imagine one bully wearing shirt,tie,pant all neatly ironed and well polished shoes wont he look good.remember dressing sense also is a part of ur interveiw**	0.8140	0.8960
31	**This is very. Bad as the uniforms are also cost effective**	0.8055	0.0000
32	**Means you don't have to worry about what you hve to wear! Less awkward when people say wear school uniform and you wear mufti...**	0.6293	0.8933
33	**Who in their right mind wants to get rid of Catholic school girl outfits?**	0.5338	0.0005

Table 4.11 Ranking of bottom 5 arguments by TFR-BERT model using MSE Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the bottom 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
29	**school uniform does no harm to students life emotions now as for the point of expressing oneself imagine one bully wearing shirt,tie,pant all neatly ironed and well polished shoes wont he look good.remember dressing sense also is a part of ur interveiw**	0.4389	0.8960
30	**This is very. Bad as the uniforms are also cost effective**	0.3763	0.0000
31	Wearing school uniform U can be sure that you go to school to study, not showing how fashionable you are	0.2733	0.9628
32	**Means you don't have to worry about what you hve to wear! Less awkward when people say wear school uniform and you wear mufti...**	0.2124	0.8933
33	**Who in their right mind wants to get rid of Catholic school girl outfits?**	0.1911	0.0005

4.4.2 IBM Evi Dataset

Table 4.12 Evaluation of TFR BERT using different ranking losses on *IBM Evi* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
Pointwise	BERT	0.57	0.51	0.37	0.88	**0.90**	0.89
	TFR-BERT MSE	0.56	0.50	0.36	0.90	**0.90**	0.91
Pairwise	TFR-BERT Hinge Loss	0.53	0.46	0.34	0.88	0.88	0.88
	TFR-BERT Logistic Loss	0.37	0.36	0.26	0.86	0.84	0.86
List-wise	TFR-BERT Softmax Loss	0.60	0.54	0.39	**0.91**	**0.90**	**0.92**
	TFR-BERT list MLE	0.38	0.29	0.21	0.77	0.79	0.81
	TFR-BERT Approx NDCG Loss	0.55	0.52	0.36	0.90	0.89	0.80
Mix	**TFR-BERT Ensemble Losses**	**0.61**	**0.56**	**0.41**	**0.91**	0.89	0.89

For model evaluation on *IBM Evi*, we used the exact same test set as in the published dataset [2]. We then reserved 25% of the training set for the validation set, as shown in table 4.4. Table 4.12 shows the performance of TFR-BERT trained with different loss functions on *IBM Evi*, comparing its performance with BERT. We can see that 2 variants of TFR-BERT outperform BERT for Pearson, Spearman & Kendall's Tau metrics: TFR-BERT trained with *Softmax* loss function and Ensemble TFR-BERT, which combines models trained with *MSE*, *Softmax* and *Approx NDCG* losses respectively. Those two variants of TFR-BERT also have the edge on BERT for the NDCG@5 metric. While BERT is not outperformed on the NDCG@10, its performance is matched by both TFR-BERT trained with *MSE* loss

and TFR-BERT trained with *Softmax* loss. TFR-BERT trained with *Softmax* loss function is the best performing model across NDCG@K metrics, outperforming all models on the NDCG@15 metric. Ensemble TFR-BERT is the best performing model across Pearson, Spearman & Tau metrics. Overall, Ensemble TFR-BERT is the best performing model on the *IBM Evi* dataset, demonstrating the effectiveness of an ensemble approach of multiple ranking loss functions, as described in section 4.3.2.

4.4.3 IBM ArqQ Rank

Table 4.13 Evaluation of TFR BERT using different ranking losses on *IBM ArqQ Rank* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
Pointwise	**State-of-the-art: BERT**	**0.42**	**0.41**	0.22	0.55	0.60	0.63
	TFR-BERT MSE	0.30	0.29	0.20	0.63	0.64	0.66
Pairwise	TFR-BERT Hinge Loss	0.31	0.31	0.21	0.61	0.63	0.64
	TFR-BERT Logistic Loss	0.33	0.34	**0.24**	0.60	0.63	0.66
List-wise	TFR-BERT Softmax Loss	0.34	0.33	0.23	0.57	0.61	0.62
	TFR-BERT List MLE	0.32	0.31	0.21	0.58	0.61	0.64
	TFR-BERT Approx NDCG Loss	0.29	0.32	0.22	0.62	0.64	**0.67**
Mix	**TFR-BERT Ensemble Losses**	0.35	0.34	0.23	**0.64**	**0.67**	0.66

Table 4.13 shows the ranking performance of TFR-BERT variants on *IBM ArqQ Rank*, showing how challenging the dataset is for TFR-BERT. BERT, which is the state-of-the-art on dataset *IBM ArqQ Rank*, is the best performing model on Pearson and Spearman metrics. On Kendall's Tau, however, many configurations of TFR-BERT outperform [7]'s state-of-the-art BERT, TFR-BERT trained with *logistic loss* being the best performing model on that metric. Almost all TFR-BERT configurations outperform BERT according to NDCG@K metrics. Ensemble TFR-BERT, which combines models trained with *MSE*, *Hinge*, *Logistic*, *Softmax* and *list MLE* losses respectively, remains the best performing model across most metrics on *IBM ArqQ Rank*.

Table 4.14 Evaluation of TFR BERT using different ranking losses on *IBM Arg 30K* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
Pointwise	**State-of-the-art: BERT**	**0.52**	**0.48**	**0.32**	0.85	0.87	0.86
	TFR-BERT MSE	0.50	0.45	0.32	0.87	0.87	0.87
Pairwise	TFR-BERT Hinge Loss	0.49	0.45	0.31	**0.90**	0.89	0.88
	TFR-BERT Logistic Loss	0.50	0.45	0.31	0.88	0.88	0.88
List-wise	TFR-BERT Softmax Loss	0.49	0.43	0.30	0.86	0.86	0.86
	TFR-BERT List MLE	0.51	0.45	0.32	0.89	**0.90**	**0.89**
	TFR-BERT Approx NDCG Loss	0.43	0.42	0.30	0.88	0.87	0.87
Mix	**TFR-BERT Ensemble Losses**	**0.52**	0.47	**0.32**	0.89	0.89	0.88

4.4.4 IBM Arg 30K

To ensure a proper comparison to the state-of-the-art on dataset *IBM Arg 30K*, we use the exact same division into train, validation and test sets as described in the published dataset [6], as shown in table 4.4. Table 4.14 shows the performance of BERT and different configurations of TFR-BERT on dataset *IBM Arg 30K*. Comparing the TFR-BERT architecture to BERT, we can see that Ensemble TFR-BERT, which combines models trained with *MSE*, *Hinge*, *Logistic*, *Softmax* and *list MLE* losses respectively, matches BERT's performance on Pearson and Kendall's Tau, and performs similarly on Spearman metric. However, almost every configuration of TFR-BERT, including ensemble losses, outperforms BERT on all NDCG@K metrics. TFR-BERT trained with *List MLE* loss is the best performing model over most NDCG@K metrics, while TFR-BERT trained with *Hinge* loss is the best performing model on the NDCG@5 metric.

4.5 Discussion

In earlier chapters, we start by describing learning-to-rank methods and BERT, which are the building blocks for the solutions we present in this chapter. Combining BERT with learning-to-rank methods, we present TFR-BERT and then show how we leverage this architecture, applied to the task of argument quality ranking. To thoroughly evaluate the performance of our presented solution, we rely on 4 different argument quality datasets, which have been described in chapter 3.

The evaluation process, repeated over 4 datasets, demonstrate that TFR-BERT, evaluated on every major argument quality dataset, generally outperforms state-of-the-art solutions on NDCG@K metrics and performs similarly to the state-of-the-art on Pearson, Spearman metrics & Kendall's Tau. We show a summary of the performance of TFR-BERT compared to the state-of-the-art on all datasets in table 4.15. TFR-BERT's performance for the NDCG@K metric shows the model is successful (to a degree) at returning the top K most *convincing* arguments. To properly visualize this aspect of the model's performance, for every variant of TFR-BERT presented, we analyze closely how the top 5 arguments predicted compared to gold standard and also, how it compared to BERT's top 5 arguments. As BERT is considered as the state-of-the-art solution for the argument quality ranking task on 3 out of the 4 datasets, comparing its top 5 arguments predicted for a topic to TFR-BERT's top 5 allowed to properly show an example of how TFR-BERT might have a stronger capability at outlining the topmost convincing arguments of a list. This highlights the value of using learning-to-rank methods for the argument quality ranking task. Considering applications of argument quality ranking, one could say that returning the top K best arguments of a list has more value than the whole ranked list in itself. This reinforces our call for the usage of the NDCG@K metric for the task of argument quality ranking.

Comparing the different types of ranking losses, we can observe the pairwise and list-wise ranking losses usually performed better for the NDCG@K metrics, and thus at identifying top K most convincing arguments of a list. While one loss function did not stand out as generally the best across datasets, an ensemble model of multiple TFR-BERT trained with different loss functions always yielded better results. On almost every dataset it was evaluated on, ensemble TFR-BERT outperforms every TFR-BERT trained using only one ranking loss function, generally performing more uniformly across all metrics, demonstrating a more robust approach to argument quality ranking.

It is important to note that each dataset had its own score of quality, where each score differs in the way it is calculated, transformed from pairwise annotations or the way it was collected. In the next chapter, we explore the feasibility of using a normalized score for all argument quality datasets, thus unifying them.

Table 4.15 Summary table of the evaluation of TFR BERT using different ranking losses on all major argument quality datasets.

	Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
UKP Rank	Pointwise	BERT	0.44	0.56	0.40	0.53	0.62	0.68
		TFR-BERT MSE Loss	0.45	0.68	0.51	0.59	0.67	0.72
	Pairwise	TFR-BERT Hinge Loss	0.44	0.60	0.46	**0.63**	**0.72**	0.75
		TFR-BERT Logistic Loss	0.38	0.59	0.45	0.43	0.57	0.61
		State-of-the-art: Sum-of-Words-Embeddings + FFNN	**0.48**	**0.69**	**0.52**	-	-	-
	List-wise	TFR-BERT Softmax Loss	0.40	0.67	0.51	0.49	0.61	0.66
		TFR-BERT List MLE	0.36	0.61	0.45	0.36	0.54	0.60
		TFR-BERT Approx NDCG Loss	0.47	0.59	0.44	0.54	0.66	0.69
	Mix	**TFR-BERT Ensemble Losses**	**0.48**	0.68	0.51	0.60	**0.72**	**0.77**
IBM Evi	Pointwise	BERT	0.57	0.51	0.37	0.88	**0.90**	0.89
		TFR-BERT MSE	0.56	0.50	0.36	0.90	**0.90**	0.91
	Pairwise	TFR-BERT Hinge Loss	0.53	0.46	0.34	0.88	0.88	0.88
		TFR-BERT Logistic Loss	0.37	0.36	0.26	0.86	0.84	0.86
	List-wise	TFR-BERT Softmax Loss	0.60	0.54	0.39	**0.91**	**0.90**	**0.92**
		TFR-BERT list MLE	0.38	0.29	0.21	0.77	0.79	0.81
		TFR-BERT Approx NDCG Loss	0.55	0.52	0.36	0.90	0.89	0.80
	Mix	**TFR-BERT Ensemble Losses**	**0.61**	**0.56**	**0.41**	**0.91**	0.89	0.89
IBM ArgQ Rank	Pointwise	**State-of-the-art:** BERT	**0.42**	**0.41**	0.22	0.55	0.60	0.63
		TFR-BERT MSE	0.30	0.29	0.20	0.63	0.64	0.66
	Pairwise	TFR-BERT Hinge Loss	0.31	0.31	0.21	0.61	0.63	0.64
		TFR-BERT Logistic Loss	0.33	0.34	**0.24**	0.60	0.63	0.66
	List-wise	TFR-BERT Softmax Loss	0.34	0.33	0.23	0.57	0.61	0.62
		TFR-BERT List MLE	0.32	0.31	0.21	0.58	0.61	0.64
		TFR-BERT Approx NDCG Loss	0.29	0.32	0.22	0.62	0.64	**0.67**
	Mix	**TFR-BERT Ensemble Losses**	0.35	0.34	0.23	**0.64**	**0.67**	0.66
IBM Arg 30K	Pointwise	**State-of-the-art:** BERT	**0.52**	**0.48**	**0.32**	0.85	0.87	0.86
		TFR-BERT MSE	0.50	0.45	0.32	0.87	0.87	0.87
	Pairwise	TFR-BERT Hinge Loss	0.49	0.45	0.31	**0.90**	0.89	0.88
		TFR-BERT Logistic Loss	0.50	0.45	0.31	0.88	0.88	0.88
	List-wise	TFR-BERT Softmax Loss	0.49	0.43	0.30	0.86	0.86	0.86
		TFR-BERT List MLE	0.51	0.45	0.32	0.89	**0.90**	**0.89**
		TFR-BERT Approx NDCG Loss	0.43	0.42	0.30	0.88	0.87	0.87
	Mix	**TFR-BERT Ensemble Losses**	**0.52**	0.47	**0.32**	0.89	0.89	0.88

CHAPTER 5 STANDARDIZED ARGUMENT QUALITY METRIC

5.1 Motivation

In chapter 3, we described in details each dataset we use in this work and showed they differ in many ways. They differ in the way the data was collected. While most of the datasets were collected as pairwise annotations, there are major differences in the way the argument quality scores were induced from pair annotations. The transformation step to extract a point-wise score for each individual argument from argument pair annotations creates heterogeneity among datasets. For example, [1] used PageRank for this transformation step on *UKP ConvArgStrict* dataset, while [2] used a Siamese BiLSTM for this transformation step on dataset *IBM Evi*: training the Siamese BiLSTM on the pair annotations and using one leg of the BiLSTM to predict a quality score for each argument. In this section, we explore the feasibility of using a common score for the transformation step from pair annotations to individual scores, that would allow to make the argument quality datasets more homogeneous.

Moreover, in chapter 4, we show in table 4.15 that the average performance of ranking models varies from one dataset to another. We try to identify the source of those variations. Two reasons could potentially explain the variation of performance. First, as explained earlier, the transformation step from pair annotations to individual scores is different for each dataset, which could explain the difference in performance. Second, we inquire on the quality of the collected argument annotations from human annotators. In other words, if we look at an argument from one of the datasets, would we be in agreement with the quality score it is labeled with? Having those two possible explanations in mind, in this section, we start by performing a qualitative analysis of the datasets in section 5.2, and then we explore the feasibility of using a common metric for the transformation step from pair annotations to individual scores, thus unifying argument quality datasets.

5.2 Qualitative Analysis of datasets

We performed a random qualitative evaluation of the four datasets, analyzing the validity of the quality score of an argument. To do so, for each dataset, we randomly picked 5 topics from the train set and 5 topics from the test set. For each topic, we picked the 3 most convincing arguments and the 3 less convincing arguments. In total, this process yields a sample of 60 arguments per dataset. Having that sample of very strong arguments and

very weak arguments from random topics, we then asked 3 annotators to annotate those arguments by hand. The directives were the following: *would you use this argument if you had to argue on the topic at hand?* The decision had to be binary: yes or no. The annotators carried out this exercise and their annotations were averaged, in a manner similar to [6]'s WA (Weighted Average) score.

Table 5.1 shows the level of agreement of the average of the annotations and the original quality scores on the sample of each dataset. This means that the argument quality scores in each of the dataset are compared to the average score of 3 annotators on high quality arguments and low quality arguments from random topics. From the results in table 5.1, we can see that the annotations collected using our qualitative analysis demonstrated strong correlation with the quality score for dataset *UKP Rank* and dataset *IBM Arg30K*. In fact, in both cases, the correlation is over 0.8. This shows confidence in the correctness of the quality scores for those two datasets. However, the correlation is lower for dataset *IBM ArgQ* and much lower for dataset *IBM Evi*. This puts the correctness of the quality scores of dataset *IBM Evi* into question. Let's not forget that the dataset *IBM Evi* differs from other datasets the most in the way the individual argument quality score was inferred from pairwise annotations. A BiLSTM is initially trained on the argument pair classification task using pairwise annotations and then, one leg of the BiLSTM is used to predict a score for each argument. Therefore, this motivates looking into a common metric allowing to infer a pointwise score from pair annotations, applicable to all datasets.

Table 5.1 Correlation between the average of annotator scores and the original score from the sample of each dataset.

Dataset	Pearson	p-value
UKP Rank	0.8347	< 0.0001
IBM Evi	0.2376	0.0676
IBM ArgQ	0.6779	< 0.0001
IBM Arg30K	0.8860	< 0.0001

Having presented the correlation between the average of the annotations and the original quality scores, we calculate the inter-annotator agreement using the Cohen Kappa Score.

Cohen Kappa Score

The Cohen Kappa score is a measure of the agreement between two annotators classifying N items into C mutually exclusive categories, taking into account the probability of the two annotators agreeing by chance. The Cohen Kappa score is calculated as shown in equation 5.1:

$$\kappa = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} \tag{5.1}$$

Where P_o corresponds to the relative observed agreement among annotators and P_e corresponds to the expected proportion of agreement among annotators. f_{ij} corresponds to the number of times the first annotator assigned an item to category i and the second annotator assigned the same item to category j, generating a k by k confusion matrix. r_i and c_j correspond to the row and column totals of the confusion matrix for category i and j [41].

$$P_o = \frac{1}{N} \sum_{j=1}^{k} f_{jj} \tag{5.2}$$

$$r_i = \sum_{j=1}^{k} f_{ij}, \forall i \tag{5.3}$$

$$c_j = \sum_{i=1}^{k} f_{ij}, \forall j \tag{5.4}$$

$$P_e = \frac{1}{N^2} \sum_{i=1}^{k} r_i c_i \tag{5.5}$$

The Cohen Kappa score being a measure of the agreement, the interpretation of its value is detailed in table 5.2 [42]. For example, a value of 0 indicates an agreement equivalent to chance and a score of 1 indicates perfect agreement. We will use table 5.2 as reference to describe the agreement level between annotators for the quality analysis of the argument quality datasets.

Table 5.2 Cohen Kappa Score interpretation.

Cohen Kappa Score	Interpretation
0	Agreement equivalent to chance
0.1 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Near perfect agreement
1	Perfect agreement

Table 5.3 shows the average Cohen Kappa score between annotators for the quality analysis of the sample for each dataset. From the average of Cohen Kappa score values, we can conclude a moderate agreement between the annotators on datasets *UKP Rank* and *IBM Arg30K* and a fair agreement on datasets *IBM Evi* and *IBM ArgQ*. Interestingly, the datasets where the annotators show weaker agreement prove to be the datasets where the original quality score shows weaker correlation with the average annotator quality score.

Table 5.3 Average Cohen Kappa Score for the annotation process of each dataset's sample.

Dataset	Cohen Kappa Score
UKP Rank	0.5455
IBM Evi	0.2643
IBM ArgQ	0.3229
IBM Arg30K	0.5719

5.3 WinRate Metric

We previously established differences in performance on the argument quality datasets. We hypothesize that those differences are caused by the various ways of computing a point-wise quality score from argument pair annotations. Therefore, we propose to unify argument quality datasets using the WinRate metric, similarly to [18] on *UKP Rank*. The WinRate metric, applied to pair annotations, consists of the number of time an argument is chosen as the most convincing of the pair over the number of times the argument is shown overall, as explained in section 2.2.2.

5.3.1 Correlation with Original Quality Score

For every dataset except *IBM Arg30K*, we extract a pointwise quality score from the pair annotations using the WinRate metric. This exercise can not be done with dataset *IBM Arg30K* because the arguments were not collected through pair annotations but directly with a score for each individual argument as explained in details in chapter 3. The pointwise quality score extracted from the pair annotations using WinRate can be used for the argument quality ranking task similarly to each dataset's original pointwise quality score. Before evaluating any ranking model on the ranking task using the WinRate score, we first analyze the correlation between the WinRate score and each dataset's original pointwise quality score.

UKP ConvArgStrict Dataset

Dataset *UKP Rank* contains a pointwise quality score and is extracted from *UKP ConvArgStrict*, a dataset of argument pair annotations. Recalling from chapter 3, [1] used PageRank algorithm for that transformation step. They build a graph representation where nodes represent arguments and directed edges represent pair annotations. Edge direction indicates the most convincing argument: the target of the edge is the most convincing argument of the pair. PageRank allows to rank the arguments for each topic, using pair annotations. Similarly to [1], we extract rankings from *UKP ConvArgStrict*, generating a new version of *UKP Rank*, this time using the WinRate metric to extract rankings. We then compare the WinRate to the score generated using PageRank. Table 5.4 shows the correlation between the two scores.

Table 5.4 Correlation between WinRate score and PageRank score on *UKP ConvArgStrict* Dataset.

Topic	Stance	Pearson	p-value
if your spouse committed murder and he or she confided in you would you turn them in	yes	0.5713	0.0003
	no	0.5037	0.0020
gay marriage right or wrong	allowing gay marriage is wrong	0.6766	0.0000
	allowing gay marriage is right	0.4910	0.0037
william farquhar ought to be honoured as the rightful founder of singapore	no it is raffles	0.6065	0.0002
	yes of course	0.6620	0.0000
personal pursuit or advancing the common good	advancing the common good	0.6887	0.0000
	personal pursuit	0.6411	0.0000
firefox vs internet explorer	internet explorer	0.7290	0.0000
	firefox	0.6478	0.0003
evolution vs creation	creation	0.8380	0.0000
	evolution	0.7162	0.0000
india has the potential to lead the world	no against	0.6700	0.0000
	yes for	0.7054	0.0000
ban plastic water bottles	yes emergencies only	0.8687	0.0000
	no bad for the economy	0.6339	0.0003
is it better to have a lousy father or to be fatherless	lousy father	0.6445	0.0000
	fatherless	0.5641	0.0008
christianity or atheism	christianity	0.6873	0.0000
	atheism	0.4267	0.0149
should physical education be mandatory in schools	yes	0.7527	0.0000
	no	0.7800	0.0000
pro choice vs pro life	pro life	0.7446	0.0000
	pro choice	0.6789	0.0000
human growth and development should parents use spanking as an option to discipline	no	0.5963	0.0002
	yes	0.4499	0.0067
tv is better than books	books	0.4819	0.0109
	tv	0.8099	0.0000
is porn wrong	yes porn is wrong	0.6984	0.0001
	no is is not	0.6176	0.0002
is the school uniform a good or bad idea	bad	0.5711	0.0003
	good	0.4847	0.0037
Average		0.6450	0.0014

IBM ArgQ Pairs

Dataset *IBM ArgQ* differs from other datasets because it is collected in two different ways, as we explain in chapter 3. First, it is collected as argument pair annotations, as per *UKP ConvArgStrict* and *IBM Evi*. However, it is also directly collected as pointwise argument quality scores. This yields a dataset for the ranking task without the need for a transformation step like [1]. Since there is no need for a transformation step on *IBM ArgQ*, it gives us the perfect opportunity to evaluate the WinRate metric. We apply the WinRate to the argument pair annotations and compare the result to the pointwise argument quality score directly collected by [7]. Table 5.5 shows the correlation between the WinRate and the

original score.

Table 5.5 Correlation between WinRate score and original score on *IBM ArgQ Pairs* Dataset.

Topic	Stance	Pearson	p-value
Flu-vaccination-should-be-mandatory	PRO	0.6655	0.0000
We-should-adopt-cryptocurrency	PRO	0.6782	0.0000
Social-media-brings-more-good-than-harm	CON	0.5822	0.0000
We-should-adopt-vegetarianism	PRO	0.4905	0.0000
We-should-abandon-vegetarianism	CON	0.5546	0.0000
We-should-ban-doping-in-sport	CON	0.3477	0.0010
Gambling-should-be-banned	PRO	0.6516	0.0000
Gambling-should-not-be-banned	CON	0.6290	0.0000
We-should-discourage-information-privacy-laws	CON	0.7205	0.0000
Social-media-brings-more-harm-than-good	PRO	0.5395	0.0000
Online-shopping-brings-more-good-than-harm	CON	0.6890	0.0000
We-should-limit-autonomous-cars	PRO	0.4617	0.0000
Flu-vaccination-should-not-be-mandatory	CON	0.5481	0.0000
We-should-not-ban-fossil-fuels	CON	0.7616	0.0000
We-should-promote-autonomous-cars	CON	0.5310	0.0000
We-should-legalize-doping-in-sport	PRO	0.3464	0.0009
We-should-abandon-cryptocurrency	CON	0.6487	0.0000
We-should-support-information-privacy-laws	PRO	0.6343	0.0000
Online-shopping-brings-more-harm-than-good	PRO	0.5043	0.0000
We-should-allow-the-sale-of-violent-video-games-to-minors	CON	0.3861	0.0019
We-should-ban-the-sale-of-violent-video-games-to-minors	PRO	0.4263	0.0004
We-should-ban-fossil-fuels	PRO	0.4690	0.0001
Average		0.5575	0.0002

IBM EviConv

Dataset *IBM EviConv*'s pointwise quality score is, as described earlier, generated using one single leg from a Siamese BiLSTM trained on the argument pair annotations. Therefore, *IBM EviConv*'s pointwise quality score is very different from other datasets. The score is predicted by a model instead of being directly collected by human annotations or inferred from collected human annotations through a transformation step like PageRank. We compare the pointwise quality score created using WinRate to the score generated by [2]'s leg of the Siamese BiLSTM. Table 5.6 shows the correlation between the two different scores for each topic and the average across the *IBM EviConv* dataset.

Table 5.6 Correlation between WinRate score and original score on *IBM EviConv* Dataset.

Topic	Pearson	p-value
We should end affirmative action	0.0252	0.8775
We should subsidize condoms	0.2206	0.1501
We should legalize prostitution	0.4456	0.0031
We should adopt socialism	0.2054	0.4136
We should prohibit corporal punishment	0.5128	0.0001
We should further exploit wind turbines	0.3348	0.0816
We should ban trans fats usage in food	0.5499	0.0416
We should further exploit hydroelectric dams	0.4036	0.0146
We should ban partial birth abortions	0.3566	0.0385
We should fight illegal immigration	0.2365	0.2659
We should legalize polygamy	0.2005	0.2211
We should adopt open source software	0.7749	0.0003
We should abolish the monarchy	0.2787	0.0817
We should legalize cannabis	0.3339	0.0085
We should adopt a zero tolerance policy in schools	0.2767	0.4102
We should subsidize biofuels	0.4325	0.0021
We should further exploit geothermal energy	0.4782	0.0065
We should lower the drinking age	0.5513	0.0986
We should ban male infant circumcision	0.4371	0.0003
We should introduce universal health care	0.3042	0.0564
We should further exploit wind power	0.4685	0.0002
We should adopt vegetarianism	0.5771	0.0001
Sex education should be mandatory	-0.2634	0.1522
We should prohibit hydraulic fracturing	0.2760	0.1140
We should introduce school vouchers	0.3447	0.0724
We should fight gender inequality	-0.1013	0.8114
We should introduce a flat tax	-0.0234	0.9220
We should abandon coal mining	0.1772	0.4548
We should increase government regulation	-0.0122	0.9558
We should ban corporal punishment in the home	0.2702	0.4820
Bullfighting should be banned	-0.3819	0.0966
We should end international aid	0.0552	0.8277
We should ban fishing	0.6026	0.0293
Abstinence-only sex education should be mandatory	0.1375	0.4769
We should abolish the right to keep and bear arms	-0.1058	0.4740
⋮	⋮	⋮

Topic	Pearson	p-value
⋮	⋮	⋮
We should further exploit solar energy	0.1139	0.5281
We should ban full-body scanners	0.4326	0.0643
We should ban breast implants	0.1810	0.4324
We should ban boxing	0.4193	0.3490
Holocaust denial should be a criminal offence	0.2602	0.6725
We should increase gun control	-0.3642	0.0342
We should abolish zoos	0.4840	0.1314
We should abandon online dating services	0.7494	0.2506
We should increase wealth redistribution	-0.1173	0.7313
Physical education should be mandatory	-0.2005	0.6665
We should abolish intellectual property rights	0.5881	0.0958
Homeschooling should be banned	-0.0210	0.9363
We should cancel the speed limit	-0.2683	0.4536
We should ban human cloning	0.5018	0.0055
The free market should be protected	0.3119	0.0275
We should prohibit flag burning	-0.0469	0.8091
We should ban gambling	0.0090	0.9676
We should limit the freedom of speech	-0.0178	0.9166
We should further exploit nuclear power	0.1064	0.3988
We should increase ecotourism	0.2795	0.1094
We should end mining	-0.0097	0.9631
We should legalize same sex marriage	-0.1868	0.1085
We should subsidize recycling	0.5203	0.0077
We should adopt multiculturalism	0.0389	0.8192
We should legalize the growing of coca leaf	0.2335	0.5162
We should ban the sale of violent video games to minors	0.2058	0.2357
We should fight for Palestinian independence	0.2430	0.1878
We should end censorship	-0.1532	0.5710
We should abolish electronic voting	-0.1224	0.7937
We should protect endangered species	0.0842	0.7655
Big governments should be abandoned	-0.4706	0.2392
We should support water privatization	-0.1025	0.8698
We should limit genetic testing	-0.4694	0.5306
We should adopt blasphemy laws	-0.0577	0.9022
Average	0.1817	0.3584

From the correlation between the WinRate and the original quality score of each dataset, we can see a relationship definitely exists between the 2 different scores for dataset *UKP ConvArgStrict* and dataset *IBM ArgQ Pairs*. Dataset *IBM Evi* shows weaker correlation with WinRate metric. This doesn't come as a surprise, since it is the only dataset where the score comes from model predictions.

5.3.2 Comparison of Manual Scores to WinRate Scores

In the previous section, we evaluated the correlation between the WinRate and the original quality score of each dataset. In this section, we investigate how the WinRate compares to the annotations of the quality analysis of datasets done in section 5.2. The quality analysis consisted in annotating the top 3 most convincing arguments and bottom 3 less convincing arguments for 10 topics for every dataset, each argument annotated by 3 annotators. Earlier, we analyzed the correlation between the average score of the 3 annotators and the dataset quality score. Now we analyze the correlation between the average score of the 3 annotators and the WinRate score. Table 5.7 shows the correlation of interest for all datasets, also showing the correlation with the original quality score for comparison purposes. The WinRate score demonstrates a higher correlation with the average score of the 3 annotators than the original quality score, reinforcing the use of the WinRate score as an argument quality metric to generate a pointwise score from argument pair annotations.

Table 5.7 Comparison of the correlation between the average of annotator scores and the original score versus the correlation between the average of annotator scores and the WinRate score, on the sample of each dataset.

Dataset	Original Score		WinRate	
	Pearson	p-value	Pearson	p-value
UKP Rank	0.8347	0.0000	**0.8981**	< 0.0001
IBM Evi	0.2376	0.0676	**0.4532**	0.0003
IBM ArgQ	0.6779	0.0000	**0.6874**	< 0.0001

5.3.3 Top 5 Arguments According to WinRate

We showed in the last section that the WinRate score is more aligned with the manual score obtained through our qualitative analysis. To further explore how scores extracted using WinRate differ from the original quality scores, we compare the top 5 arguments on a topic (chosen randomly) according to WinRate to the top 5 arguments according to PageRank

on *UKP ConvArg*. While this exercise doesn't reflect the whole dataset, it helps visualize a sample of the difference between the two quality scores, a difference which we demonstrated in the previous section.

Table 5.8 Top 5 arguments according to PageRank on topic *is the school uniform a good or bad idea* with stance *good* of *UKP Rank* dataset. Arguments in bold are common to WinRate's top 5.

	Top N Arguments	Score
0	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	1.0000
1	**I think it's good within certain limits. I went to a school with a uniform, and it was far less stressful than non-uniform college. I'd argue that it's a leveler- prevents people from showing off material wealth/ making others feel bad for not having 'cool' stuff. But it can be taken too far. By the end, we weren't allowed coloured socks, which was idiotic.**	0.9990
2	That's really good idea. As i remember every morning i though what was better to wear? It was really problem, i spent quiet a lot of time. I asked my parents to buy new clothes for me, it was happened not rare. I know that not everyone thought as me, but it much better if the school has own uniform and everybody has to follow it. First it looks very good, smart. Secondly there is no envy that somebody have really nice skirt or jeans. Every pupil is the same and it would be easy to study, to not think about another things!!	0.9989
3	1. It makes everyone equal - if children can wear what they want some children will teased and feel less equal to their peers around them vs. uniforms 2. Okay Look school is for learning not how you look and dress but maybe in some levels it matters and most people that go to schools that don't have uniforms take like about 1 hr just to find their pants or shirt I mean really?? When you have a uniform it takes less than 10 min just to take it out and put it on and	0.9984
4	yas,of course . School uniform is important 1.school uniform is a logos for our school 2.to remind us that we are part of the school 3.and if we use the uniform basically student used to think what are they gonna do to, is it positive or negative 4.in the morning we should use our uniform and if were not use our uniform the teacher give us a punishment and from that we can learn to be a discipline student 5.if we go out from the school than the teacher will see we used the school uniform so people will know that we from that school thankyou	0.9979

Comparing WinRate's top 5 with PageRank's top 5 most convincing arguments, we can see some differences. Table 5.8 and table 5.9 show that they have 2 arguments in common and 3

different arguments. Analyzing the dissimilar arguments, we can notice that arguments based on PageRank do not display as high quality as would be expected. For example, argument #4 of table 5.8 contains familiar language and is missing clarity to a point where the argument's message is in jeopardy. Moreover, argument #2 is missing words and argument #3 uses familiar language, making those arguments of lower quality. Arguments #2, #3 and #4, in our opinion, shouldn't be part of the top 5 most convincing arguments or at least, shouldn't be labeled with a quality score so high (higher than 0.99). On the other hand, arguments part of WinRate's top 5 are of higher quality. Another aspect to consider is the score assigned to each argument of the top 5. PageRank assigned scores ranging from 0.9979 to 1.0000 to 5 arguments which, in our opinion, are of a very different level of quality. For example, argument #0 is clearly of higher quality than argument #4 in table 5.8. However, their assigned score is very close (1.0000 vs 0.9979). WinRate assigned a wider range of scores: from 0.9091 to 1.0000. We can also clearly visualize the difference in quality proportional to the difference in score. Table 5.9 shows how argument #1 being a high quality argument is assigned a score of 1.0000 and argument #4, which is more informal and contains a repetition of a group of words, is assigned a score of 0.9091. WinRate metric seems to demonstrate a better grasp of the quality difference between 2 arguments. This analysis has its limitations: we realize the analysis should be done on the whole dataset instead of just one topic and might not reflect the rest of the dataset.

Table 5.9 Top 5 arguments according to WinRate on topic *is the school uniform a good or bad idea* with stance *good* of *UKP Rank* dataset. Arguments in bold are common to PageRank's top 5.

	Top N Arguments	Score
0	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	1.0000
1	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	1.0000
2	I think school uniform is a good idea. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	0.9615
3	**I think it's good within certain limits. I went to a school with a uniform, and it was far less stressful than non-uniform college. I'd argue that it's a leveler- prevents people from showing off material wealth/ making others feel bad for not having 'cool' stuff. But it can be taken too far. By the end, we weren't allowed coloured socks, which was idiotic.**	0.9524
4	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.9091

5.3.4 Predicting WinRate

In previous sections, we evaluate how the WinRate metric correlates with the argument quality score published with each dataset and also how it correlates with the quality analysis annotations. Now we evaluate the performance when training and evaluating a predictive model on the WinRate score. We evaluated the models presented earlier: BERT and TFR-BERT trained with 6 different ranking loss functions. The training parameters are exactly

the same as seen in 4.3 to ensure consistency in the comparison to the dataset's original quality score.

UKP ConvArgStrict Dataset

Table 5.10 shows the performance of the different models predicting WinRate applied to dataset *UKP ConvArg*. We can see that ensemble TFR-BERT outperforms BERT across all metrics. While ensemble TFR-BERT is the best performing TFR-BERT variant across Pearson, Spearman and Kendall's Tau metrics, TFR-BERT trained with *Approx NDCG* Loss is the best performing TFR-BERT variant across NDCG@K metrics.

Table 5.10 Evaluation of TFR BERT using different ranking losses on the WinRate metric applied to *UKP ConvArg* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
point-wise	BERT	0.73	0.74	0.58	0.82	0.82	0.85
	TFR-BERT MSE Loss	0.68	0.68	0.53	0.83	0.83	0.86
Pairwise	TFR-BERT Hinge Loss	0.67	0.68	0.52	0.81	0.82	0.84
	TFR-BERT Logistic Loss	0.71	0.72	0.56	0.83	0.83	0.86
List-wise	TFR-BERT Softmax Loss	0.70	0.73	0.57	0.85	0.87	0.88
	TFR-BERT List MLE	0.63	0.64	0.49	0.79	0.79	0.82
	TFR-BERT Approx NDCG Loss	0.67	0.72	0.55	**0.90**	**0.89**	**0.90**
Mix	**TFR-BERT Ensemble Losses**	**0.75**	**0.76**	**0.60**	0.89	0.87	0.89

IBM ArgQ Pairs

Table 5.11 shows the performance of models predicting WinRate applied to dataset *IBM ArgQ Pairs*. We can observe that ensemble TFR-BERT outperforms BERT across all metrics. Moreover, ensemble TFR-BERT is the best performing TFR-BERT variant across all metrics, outperforming every TFR-BERT variant trained with one loss function. While ensemble TFR-BERT's performance according to Pearson, Spearman and Kendall's Tau metrics is not much higher than BERT, it outperforms BERT on NDCG@K metrics by a considerable margin.

Table 5.11 Evaluation of TFR BERT using different ranking losses on the WinRate metric applied to *IBM ArgQ Pairs* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
point-wise	BERT	0.39	0.40	0.29	0.79	0.71	0.70
	TFR-BERT MSE	0.30	0.33	0.24	0.70	0.67	0.67
Pairwise	TFR-BERT Hinge Loss	0.40	0.41	0.30	0.80	0.77	0.72
	TFR-BERT Logistic Loss	0.41	0.40	0.30	0.74	0.73	0.72
List-wise	TFR-BERT Softmax Loss	0.39	0.39	0.28	0.72	0.71	0.70
	TFR-BERT List MLE	0.32	0.32	0.23	0.78	0.74	0.70
	TFR-BERT Approx NDCG Loss	0.34	0.37	0.27	0.84	0.77	0.74
Mix	**TFR-BERT Ensemble Losses**	**0.42**	**0.43**	**0.31**	**0.90**	**0.80**	**0.76**

IBM EviConv

Table 5.12 shows the performance of the different models predicting WinRate applied to dataset *IBM EviConv*. We can see that ensemble TFR-BERT and most TFR-BERT variants outperform BERT across all metrics. Ensemble TFR-BERT is the best performing TFR-BERT model across NDCG@K metrics.

Table 5.12 Evaluation of TFR BERT using different ranking losses on the WinRate metric applied to *IBM EviConv* dataset.

Loss	Model	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
point-wise	BERT	0.34	0.32	0.24	0.69	0.70	0.71
	TFR-BERT MSE	0.45	**0.46**	**0.36**	0.73	0.73	**0.76**
Pairwise	TFR-BERT Hinge Loss	0.37	0.35	0.28	0.67	0.67	0.70
	TFR-BERT Logistic Loss	0.39	0.40	0.30	0.74	**0.74**	**0.76**
List-wise	TFR-BERT Softmax Loss	0.40	0.40	0.30	0.72	0.72	0.75
	TFR-BERT list MLE	0.42	0.44	0.34	0.72	0.73	0.75
	TFR-BERT Approx NDCG Loss	0.43	0.44	0.34	0.69	0.71	0.71
Mix	**TFR-BERT Ensemble Losses**	**0.46**	0.45	0.34	**0.75**	**0.74**	**0.76**

5.4 Discussion

In this chapter, we explore the feasibility of standardizing argument quality datasets with a common metric, motivated by the differences in how the quality score is calculated for each dataset and by the variation of performance across datasets. This led to questions

about the interest of the available datasets' scores. Therefore, we first start by performing a qualitative analysis of the datasets: annotating a sample of arguments from each dataset. Those annotations proved to be very correlated with the dataset's quality score for *UKP Rank* and *IBM Arg30K* datasets, and moderately correlated for *IBM ArgQ*. However, the annotations showed weak correlation with *IBM Evi*'s quality score. We then propose the WinRate as a metric to extract pointwise quality scores from argument pair annotations. We thoroughly compare the correlations between the WinRate and each dataset's quality score (except for *IBM Arg30K*, as we explain in section 5.3.1). The WinRate demonstrated a correlation with the dataset's original quality score for datasets *UKP Rank* and *IBM ArgQ*, but very weak correlation for dataset *IBM Evi*. Dataset *IBM Evi*'s quality score stands out as the least correlated to WinRate and qualitative analysis annotations, thus reducing the level of confidence in its validity.

Moreover, we compare the WinRate score to the average score of annotations collected through the qualitative analysis of the datasets. Some noteworthy results are the following: the WinRate score is more correlated with annotations from the qualitative analysis than the original quality score is, for all three datasets: *UKP Rank*, *IBM Evi* and *IBM ArgQ*. Additionally, we show that the WinRate labels topmost convincing arguments more accurately than PageRank, giving the top 5 arguments of topic *is the school uniform a good or bad idea* with stance *good* of *UKP Rank* dataset as an example to properly visualize it. The 5 most convincing arguments according to WinRate are globally of higher quality than the 5 most convincing arguments according to PageRank, in our opinion.

Having demonstrated WinRate as a viable metric to replace each dataset's own quality score, we evaluated BERT and all TFR-BERT variants on predicting the WinRate as the argument quality score. We report the performance for predicting the WinRate in section 5.3.4. We inquire how well a ranking model is able to learn to rank arguments through WinRate compared to using the original quality score. We compare the models' performance to predict the WinRate score to their performance predicting each dataset's original quality score in table 5.13. For good measure, we limit the comparison to BERT and the overall best performing variant of TFR-BERT: ensemble losses. We can see that for dataset *UKP Rank*, both BERT and ensemble TFR-BERT perform better at predicting the WinRate score than the PageRank score, across all metrics, by a significant margin. This demonstrates how a ranking model can learn more from WinRate than PageRank to rank arguments by their measure of quality. For dataset *IBM Evi*, however, the performance of both BERT and TFR-BERT ensemble is higher on the dataset's original quality score. For dataset *IBM ArgQ*, the

performance of ensemble TFR-BERT is higher on the WinRate score than it is on the original quality score. The performance of BERT is higher on the WinRate score for most metrics. Therefore, we can say that for the majority of cases, the WinRate is a metric from which a ranking model is more able to learn to rank arguments by their measure of quality, compared to other scores like PageRank, for example.

Table 5.13 Comparison of the ranking task on WinRate score vs the original score of each dataset.

	Model	Score	PEARSON	SPEARMAN	TAU	NDCG@5	NDCG@10	NDCG@15
UKP Rank	BERT	WinRate	**0.73**	**0.74**	**0.58**	**0.82**	**0.82**	**0.85**
		PageRank Score	0.44	0.56	0.40	0.53	0.62	0.68
	TFR-BERT Ensemble Losses	WinRate	**0.75**	**0.76**	**0.60**	**0.89**	**0.87**	**0.89**
		PageRank Score	0.48	0.68	0.51	0.60	0.72	0.77
IBM Evi	BERT	WinRate	0.34	0.32	0.24	0.69	0.70	0.71
		Original Score	**0.57**	**0.51**	**0.37**	**0.88**	**0.90**	**0.89**
	TFR-BERT Ensemble Losses	WinRate	0.46	0.45	0.34	0.75	0.74	0.76
		Original Score	**0.61**	**0.56**	**0.41**	**0.91**	**0.89**	**0.89**
IBM ArgQ	BERT	WinRate	0.39	0.40	**0.29**	**0.79**	**0.71**	**0.70**
		Original Score	**0.42**	**0.41**	0.22	0.55	0.60	0.63
	TFR-BERT Ensemble Losses	WinRate	**0.42**	**0.43**	**0.31**	**0.90**	**0.80**	**0.76**
		Original Score	0.35	0.34	0.23	0.64	0.67	0.66

CHAPTER 6 CONCLUSION

6.1 Summary of Contributions

In this work, we propose a different view on the task of ranking arguments by quality. Steering away from trying to predict an absolute quality score for each argument, we instead focus on learning how to order them by their relative convincingness. At the beginning, we ask ourselves: *How can learning to rank techniques coupled with pretrained language models contribute to automatic argument quality evaluation?* Therefore, we propose to use an architecture based on learning-to-rank built on top of BERT. We demonstrate that pairing a learning-to-rank approach with BERT's powerful ability in building a representation of an argument yields stronger ranking capabilities. This shows in the results obtained using TFR-BERT, which demonstrate better performance for the NDCG@K metrics, meaning superior capability at outlining the top K most convincing arguments. We argue that this might have more significant applications than focusing on ranking all the arguments with equal importance. We also demonstrate how combining multiple ranking loss functions (pointwise, pairwise and list-wise) as an Ensemble model of TFR-BERT shows better performance across many metrics.

Secondly, we answer the following research question: *How can argument quality datasets be standardized with a common metric, to facilitate the comparison of the ranking task?* We explore the feasibility of standardizing argument quality datasets with the WinRate metric. We demonstrate how the WinRate metric correlates with most dataset's original metric. Moreover, the WinRate metric shows greater correlation with annotations from the qualitative analysis of the sample of the datasets, than the datasets' original quality score does. We also show how, for the majority of datasets, the performance of ranking models is higher predicting the WinRate metric than the original quality score. This shows how the ranking models are able to learn more from the WinRate metric to model argument quality. This positions the WinRate metric as a viable candidate for a standardized metric to unify argument quality datasets collected from pairwise annotations.

Another contribution of this work is the publication of a research paper to the 2022 FLAIRS conference, where we present learning-to-rank methods paired with BERT for the argument quality ranking task. As part of the publication, we demonstrate the effectiveness of TFR-BERT and compare it to state-of-the-art solutions on the 4 argument quality datasets, as

shown in this work.

6.2 Limitations

One limitation of our work lies in the WinRate. Since it is a ratio of the number of times chosen over the number of times shown, it gives the same score to an argument shown once and chosen once than to an argument shown 10 times and chosen 10 times. In this case, both arguments would be given a score of 1. However, we can easily say that in the case of the second argument, we have much more certainty in the score. Therefore, the WinRate metric fails to capture the notion that the number of voters implies a bigger certainty in the assessment of the quality of the argument. In this scenario, a metric used to transform pairwise annotations into a pointwise score should be able to identify the second argument as an argument of higher quality than the first argument.

In this work, we explore pointwise, pairwise and list-wise ranking losses for the argument quality ranking task. However, for all the models we presented, the resulting trained model remained a pointwise scorer. In other words, a model trained with a list-wise loss function takes into account a whole list of arguments during training, but at inference time, however, a score is predicted for each argument individually. While some groupwise scoring learning-to-rank methods exist, scoring multiple arguments jointly ([43]), some problems would arise applying them to the argument quality ranking task. Combining groupwise multivariate scoring learning-to-rank methods with a deep pre-trained language model like BERT would require to decrease the maximum sequence length and decrease the number of arguments to rank at once because of high memory usage for such an architecture. Therefore, it would not be feasible to rank all arguments for a topic, all at once. Dividing the list of arguments into batches to rank doesn't solve the issue, as aggregating the ranked batches into one ranked list induces bias, similarly to [1]'s aggregation of pairwise annotations into ranks. Therefore, memory limitations didn't allow us to explore how groupwise multivariate scoring learning-to-rank methods could contribute to the argument quality ranking task.

6.3 Future Research

As we described in chapter 2, some state-of-the-art solutions approached the argument quality task as a ranking aggregation task. In fact, they trained their model on pairwise annotations and then evaluated their model's performance predicting ranks. This is an exercise to perform in future work, training TFR-BERT on the argument pairwise annotations

and then evaluating its performance at ranking lists of arguments.

In future work, we also would like to explore other metrics to unify argument quality datasets. As described in section 6.2, the WinRate metric entails some limitations. The WinRate metric was introduced to argument quality by [18] and demonstrated strong utility in the field. However, seeing its limitations, a useful future work would be to explore two well established methods to map pairwise annotations to individual scores: the Bradley-Terry-Plackett-Luce model [16, 44, 45] and the Elo model [19], on the datasets used in chapter 5 and compare them to WinRate. This would determine if those alternatives to WinRate overcome the limitations described in section 6.2.

Finally, we would like to explore other language models paired with learning-to-rank methods for the argument quality ranking task. In fact, new language models such as Robustly Optimized BERT Pretraining Approach (RoBERTa) [46] and Electra [47] have demonstrated a performance improvement on other tasks. It would be valuable to evaluate them as alternatives to BERT for all the learning-to-rank models presented for the argument quality ranking task.

REFERENCES

[1] I. Habernal and I. Gurevych, "Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1589–1599. [Online]. Available: http://www.aclweb.org/anthology/P16-1150

[2] M. Gleize *et al.*, "Are you convinced? choosing the more convincing evidence with a Siamese network," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 967–976. [Online]. Available: https://aclanthology.org/P19-1093

[3] S. Han *et al.*, "Learning-to-rank with BERT in tf-ranking," *CoRR*, vol. abs/2004.08476, 2020. [Online]. Available: https://arxiv.org/abs/2004.08476

[4] M.-F. Moens, "Argumentation mining: How can a machine acquire common sense and world knowledge?" *Argument Computation*, vol. 9, pp. 1–14, 07 2017.

[5] S. Somasundaran and J. Wiebe, "Recognizing stances in ideological on-line debates," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, CA: Association for Computational Linguistics, Jun. 2010, pp. 116–124. [Online]. Available: https://aclanthology.org/W10-0214

[6] S. Gretz *et al.*, "A large-scale dataset for argument quality ranking: Construction and analysis," *CoRR*, vol. abs/1911.11408, 2019. [Online]. Available: http://arxiv.org/abs/1911.11408

[7] A. Toledo *et al.*, "Automatic argument quality assessment – new datasets and methods," 2019.

[8] S. R. Bhatnagar, "Technology mediated peer instruction," Ph.D. dissertation, Polytechnique Montréal, 2021.

[9] H. Wachsmuth *et al.*, "Building an argument search engine for the web," in *ArgMining@EMNLP*, 2017.

[10] C. Stab and I. Gurevych, "Annotating argument components and relations in persuasive essays," 08 2014.

[11] J. P. N. Edward Schiappa, *Argumentation: Keeping Faith with Reason*, 2013.

[12] D. S. Hugo Mercier, "Why do humans reason? arguments for an argumentative theory." Behavioral and Brain Sciences, Cambridge University Press (CUP), 2011, pp. pp.57–74.

[13] J. Devlin *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[14] H. Li, "Learning to rank for information retrieval and natural language processing," in *Synthesis Lectures on Human Language Technologies*, 2011.

[15] J. Kekäläinen, "Binary and graded relevance in ir evaluations: Comparison of the effects on ranking of ir systems," *Inf. Process. Manage.*, vol. 41, no. 5, p. 1019–1033, sep 2005.

[16] R. L. Plackett, "The Analysis of Permutations," *Journal of the Royal Statistical Society Series C*, vol. 24, no. 2, pp. 193–202, June 1975. [Online]. Available: https://ideas.repec.org/a/bla/jorssc/v24y1975i2p193-202.html

[17] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1037–1046. [Online]. Available: https://doi.org/10.1145/2623330.2623654

[18] P. Potash, A. Ferguson, and T. J. Hazen, "Ranking passages for argument convincingness," in *Proceedings of the 6th Workshop on Argument Mining*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 146–155. [Online]. Available: https://aclanthology.org/W19-4517

[19] A. E. Elo, *The rating of chessplayers, past and present*. BT Batsford Limited, 1978.

[20] R. Pelánek, "Applications of the elo rating system in adaptive educational systems," *Computers & Education*, vol. 98, pp. 169–179, 2016.

[21] H. Nguyen and D. Litman, "Extracting argument and domain words for identifying argument components in texts," in *Proceedings of the 2nd Workshop on Argumentation Mining*, 2015, pp. 22–28.

[22] C. Stab and I. Gurevych, "Identifying argumentative discourse structures in persuasive essays," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 46–56.

[23] R. Johnson and J. Blair, *Logical Self-defense*, ser. Key titles in rhetoric, argumentation, and debate series. International Debate Education Association, 2006. [Online]. Available: https://books.google.com/books?id=ojNbr4vYooQC

[24] C. L. Hamblin, *Fallacies, by C. L. Hamblin*. Methuen [London], 1970.

[25] C. W. Tindale, *Fallacies and Argument Appraisal*. Cambridge University Press, 2007.

[26] I. Habernal and I. Gurevych, "What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1214–1223.

[27] R. Swanson, B. Ecker, and M. Walker, "Argument mining: Extracting arguments from online dialogue," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 217–226. [Online]. Available: https://aclanthology.org/W15-4631

[28] E. Simpson and I. Gurevych, "Finding convincing arguments using scalable Bayesian preference learning," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 357–371, 2018. [Online]. Available: https://aclanthology.org/Q18-1026

[29] L. A. Chalaguine and C. Schulz, "Assessing convincingness of arguments in online debates with limited number of features," in *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 75–83. [Online]. Available: https://aclanthology.org/E17-4008

[30] P. Potash, R. Bhattacharya, and A. Rumshisky, "Length, interchangeability, and external knowledge: Observations from predicting argument convincingness," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 342–351. [Online]. Available: https://aclanthology.org/I17-1035

[31] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 89–96. [Online]. Available: https://doi.org/10.1145/1102351.1102363

[32] Y. Gu *et al.*, "Incorporating topic aspects for online comment convincingness evaluation," in *Proceedings of the 5th Workshop on Argument Mining*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 97–104. [Online]. Available: https://aclanthology.org/W18-5212

[33] D. Hovy *et al.*, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 1120–1130. [Online]. Available: https://aclanthology.org/N13-1132

[34] N. A. E. A. G. Baratloo A, Hosseini M, "Part 1: Simple definition and calculation of accuracy, sensitivity and specificity." *Emerg (Tehran)*, vol. 3(2), pp. 48–49, 2015.

[35] Wikipedia contributors, "Pearson correlation coefficient — Wikipedia, the free encyclopedia," 2022, [Online; accessed 27-March-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1078965237

[36] ——, "Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia," 2022, [Online; accessed 27-March-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Spearman%27s_rank_correlation_coefficient&oldid=1068343841

[37] ——, "Kendall rank correlation coefficient — Wikipedia, the free encyclopedia," 2022, [Online; accessed 27-March-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Kendall_rank_correlation_coefficient&oldid=1072432336

[38] Y. Wang *et al.*, "A theoretical analysis of NDCG type ranking measures," *CoRR*, vol. abs/1304.6480, 2013. [Online]. Available: http://arxiv.org/abs/1304.6480

[39] R. K. Pasumarthi *et al.*, "Tf-ranking: Scalable tensorflow library for learning-to-rank," *CoRR*, vol. abs/1812.00073, 2018. [Online]. Available: http://arxiv.org/abs/1812.00073

[40] T. Qin, T.-Y. Liu, and H. Li, "A general approximation framework for direct optimization of information retrieval measures," Tech. Rep. MSR-TR-2008-164, November 2008. [Online]. Available: https://www.microsoft.com/en-us/research/publication/a-general-approximation-framework-for-direct-optimization-of-information-retrieval-measures/

[41] D. G. Altman, *Practical Statistics for Medical Research*. London: Chapman & Hall / CRC, 1991.

[42] A. Agresti, *Categorical data analysis*, ser. A Wiley-Interscience publication. New York [u.a.]: Wiley, 1990.

[43] Q. Ai *et al.*, "Learning groupwise multivariate scoring functions using deep neural networks," *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1145/3341981.3344218

[44] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952. [Online]. Available: http://www.jstor.org/stable/2334029

[45] R. D. Luce, "On the possible psychophysical laws." *Psychological Review*, p. 66(2):81, 1959.

[46] Y. Liu *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[47] K. Clark *et al.*, "ELECTRA: pre-training text encoders as discriminators rather than generators," *CoRR*, vol. abs/2003.10555, 2020. [Online]. Available: https://arxiv.org/abs/2003.10555

APPENDIX A PREDICTION OF THE TOP 5 ARGUMENTS ON UKP RANK

Table A.1 Ranking of top 5 arguments by TFR-BERT model using Mean Squared Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.8717	0.9978
1	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.8659	1.0000
2	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.8570	0.9974
3	year,i support this view. when i was studying in the school,for me wasn't a problem what to wear . In the university....i spend too much time in choosing clothes. firstly, it is wasting a time. secondly, when all students wear one uniform, there wouldn't be any discrimination, dividing into social status groups. A uniform shows students equality.	0.8270	0.9889
4	yes, i believe it's nice to have a school uniform. Each school 's uniform signifies its goal for instant i wore white shirt and blue skirt in my school days, white color is an indication of peace and blue of fidelity in relationships moreover identical uniform also removes the wall of status. it also depicts that whether a child comes from high or low class they all are treated equally under one roof .	0.7873	0.9930

Table A.2 Ranking of top 5 arguments by TFR-BERT model using Pairwise Hinge Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.8809	1.0000
1	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.8401	0.9978
2	I think school uniform is a good idea. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	0.8282	0.9972
3	Uniforms allow an equal and fair social status only based on personality and not looks. I do half to admit wearing what you want is fun and creative but its only fun if everyone can do it and for some children thats not the case and they cant afford to live up to their peers standards so uniforms would make social life much easier and it would give a more mature look to the school.	0.7954	0.9976
4	They prepare people for the clothes they may have to wear later on in life. They ensure that no unsuitable clothing is worn.	0.7724	0.9747

Table A.3 Ranking of top 5 arguments by TFR-BERT model using Pairwise Logistic Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.9943	1.0000
1	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.9730	0.9978
2	I think school uniform is a good idea. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	0.9669	0.9972
3	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.9521	0.9974
4	I believe that the wearing of the school uniform should be encouraged because it reminds each child that they are equal (at least in school). It also shows unity and children feel included and it helps them to work as a team.	0.9143	0.9960

Table A.4 Ranking of top 5 arguments by TFR-BERT model using Softmax Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	1.0000	0.9978
1	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.9403	0.9974
2	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.9108	1.0000
3	year,i support this view. when i was studying in the school,for me wasn't a problem what to wear . In the university....i spend too much time in choosing clothes. firstly, it is wasting a time. secondly, when all students wear one uniform, there wouldn't be any discrimination, dividing into social status groups. A uniform shows students equality.	0.8662	0.9889
4	Uniforms allow an equal and fair social status only based on personality and not looks. I do half to admit wearing what you want is fun and creative but its only fun if everyone can do it and for some children thats not the case and they cant afford to live up to their peers standards so uniforms would make social life much easier and it would give a more mature look to the school.	0.7111	0.9976

Table A.5 Ranking of top 5 arguments by TFR-BERT model using Approx NDCG Loss on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	Uniforms allow an equal and fair social status only based on personality and not looks. I do half to admit wearing what you want is fun and creative but its only fun if everyone can do it and for some children thats not the case and they cant afford to live up to their peers standards so uniforms would make social life much easier and it would give a more mature look to the school.	0.9925	0.9976
1	**1. It makes everyone equal - if children can wear what they want some children will teased and feel less equal to their peers around them vs. uniforms 2. Okay Look school is for learning not how you look and dress but maybe in some levels it matters and most people that go to schools that don't have uniforms take like about 1 hr just to find their pants or shirt I mean really?? When you have a uniform it takes less than 10 min just to take it out and put it on and**	0.9750	0.9984
2	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.9705	0.9978
3	I believe that the wearing of the school uniform should be encouraged because it reminds each child that they are equal (at least in school). It also shows unity and children feel included and it helps them to work as a team.	0.9333	0.9960
4	i think it's a good idea. so the student don't have to worry about what would they wear. they'll become more concentrate with their study so it's more efficient. if other says they can't express them self, oh please, there's so many things you can do to express yourself. and i think, uniform indirectly give thought how to dress correctly. if we let them dress theirself, they could wear Inappropriate clothes such as hot pants, rebel jeans or sexy clothes. that's not good for their mind.	0.9282	0.9972

Table A.6 Ranking of top 5 arguments by Ensemble TFR-BERT on the topic *Is the school uniform a good or bad idea* with the stance *good* on *UKP Rank* dataset. Arguments are shown in bold if they are part of the top 5 according to gold standard.

	Top N Arguments	Predicted Score	Score
0	In a school all the students may not belong to the same financial status . Some may be rich , some may not be that rich . So uniform provides equal status to all the students so that there is no gap among them . If there is no uniform , then the rich students will wear new dresses everyday which the other students cannot afford and may lead to resentment among them . Some insensitive children may also mock other students wear old cloths . So I beleive that uniform is essential especially in developing countries .	0.9206	0.9978
1	**According to the legacy educational resources, as fashion and trends change, students become more concerned with how they look and how they are perceived than they do with their academic success and achievement. The fashion of low rise jeans, bagging jeans, large trench coats, low cut shirts, and many others contribute to behavior problems and safety issues in the classrooms and in the hallways of schools today.**	0.8872	1.0000
2	School uniform is a great idea, just because it makes impossible to hold the race for the fashion among pupils. let it be, one pupil is richer than another. rich can begin to show off in front of those who are poorer. this action will create a negative atmosphere in the school and can start row between both pupils. As a rule, As a rule, it often occurs between the girls, although it is not rare between the boys. that's why school uniform liquidates all conflicts.	0.8727	0.9974
3	Uniforms allow an equal and fair social status only based on personality and not looks. I do half to admit wearing what you want is fun and creative but its only fun if everyone can do it and for some children thats not the case and they cant afford to live up to their peers standards so uniforms would make social life much easier and it would give a more mature look to the school.	0.8155	0.9976
4	year,i support this view. when i was studying in the school,for me wasn't a problem what to wear . In the university....i spend too much time in choosing clothes. firstly, it is wasting a time. secondly, when all students wear one uniform, there wouldn't be any discrimination, dividing into social status groups. A uniform shows students equality.	0.8133	0.9889