# RESOLUTION OF TURKISH SCALAR IMPLICATURES BY LARGE LANGUAGE MODELS

MUSTAFA KÜRŞAT HALAT

BOĞAZİÇİ UNIVERSITY

2024

# RESOLUTION OF TURKISH SCALAR IMPLICATURES BY LARGE LANGUAGE MODELS

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Linguistics

by

Mustafa Kürşat Halat

Boğaziçi University

2024

Resolution of Turkish Scalar Implicatures by Large Language Models

This thesis of Mustafa Kürşat Halat

has been approved by:

Assist. Prof. Ümit Atlamaz _____
(Thesis Advisor)

Assist. Prof. Ömer Demirok _____

Prof. Hüseyin Cem Bozşahin _____
(External Member)

June 2024

# DECLARATION OF ORIGINALITY

I, Mustafa Kürşat Halat, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature: ……

Date: ……

ABSTRACT

Resolution of Turkish Scalar Implicatures by Large Language Models

As for a scalar pair consisting of a weaker term and a stronger term, scalar implicatures are the pragmatic inferences drawn when the use of the weaker term denotes the negation of the stronger term. This study investigates the pragmatic capabilities of Large Language Models (LLM) in the resolution of scalar implicatures. As part of the study, a Natural Language Inference (NLI) dataset in Turkish, *ImplicaTR*, was developed, which is a four-class NLI dataset that enriches the conventional entailment-neutral-contradiction classification with a new class, implicature. With a total of 2,340 sets of sentences spanning five different linguistics categories, ImplicaTR has a size of 20,340 rows. Two experiments were conducted on ImplicaTR. In Experiment 1, various BERT models and generative models such as Gemma, Llama-2, and Mistral were inspected on pragmatic reasoning, and we found that LLMs can reason about scalar inferences with an accuracy over 98% on test dataset. As a result, we obtained an NLI model that does more fine-grained analysis. In Experiment 2, we carried out a linguistic inquiry within an ablation study to reveal the factors influencing the entailment and implicature resolution of the models. Our findings showed that the frequency of individual scalar items is positively correlated with the model's ability to resolve the pragmatic inferences. From the linguistic perspective, our study demonstrated that scalar reasoning is not solely a pragmatic process nor a lexical one; both mechanisms seem to play a role in implicature inferencing.

# ÖZET

Türkçedeki Skaler Sezdirimlerin Büyük Dil Modelleri tarafından Çözümlenmesi

Skaler sezdirimler, bir zayıf terim ve bir güçlü terimden oluşan bir skaler çifti için zayıf terim kullanımının güçlü terimin doğruluğunun reddini ifade ettiği edimsel çıkarımlardır. Bu çalışmada büyük dil modellerinin (LLM) skaler sezdirim çözümlenmesindeki yetenekleri araştırılmaktadır. Çalışmanın bir parçası olarak Doğal Dil Çıkarımı'ndaki (NLI) geleneksel gerektirim-nötr-çelişki sınıflarına sezdirim sınıfını ekleyerek dört sınıflı bir sınıflandırma sunan Türkçe NLI veri seti ImplicaTR geliştirilmiştir. Beş farklı dilbilimsel kategori içeren toplamda 2.340 cümle kümesinden oluşan ImplicaTR 20.340 veri satırına sahip olup bu çalışmada bu veri seti üzerinde iki deney yapılmıştır. Deney 1'de BERT modelleri ve Gemma, Llama-2 ve Mistral gibi üretici modellerin skaler sezdirim muhakemeleri incelenmiş ve test veri seti üzerindeki 98% doğruluk oranı ile bu modellerin edimsel muhakeme yapabildiği gösterilmiştir. Bu deney sonucunda daha detaylı analiz yapabilen bir NLI modeli elde edilmiştir. Deney 2'de modellerin gerektirim ve sezdirim yeteneklerine etki eden faktörleri ortaya çıkarmaya yönelik olarak ablasyon deneyi içerisinde dilbilimsel tahkikat yapılmıştır. Deney sonuçlarına göre tekil skaler terimlerin derlemdeki sıklıklarının dil modelinin edimsel çıkarım çözümlemesiyle pozitif korelasyon gösterdiği görülmüştür. Dilbilimsel açıdan bu çalışma, skaler muhakemenin sadece edimsel veya sözlüksel bir süreç olmadığını ve iki mekanizmanın da sezdirim çıkarımında rol oynadığını ortaya koymuştur.

# ACKNOWLEDGMENTS

It is a must for me to begin this section by extending my deepest gratitude to my advisor, Ümit Atlamaz, for his unwavering support, guidance, and companionship throughout this challenging yet rewarding journey. His patience and encouragement have been invaluable in building this thesis from the ground up and he bore with me through iteration over iteration to bring it to perfection. I sometimes found myself stepping on some unassuming thoughts and considerations, where his expertise and stance enriched the quality and depth of this research. I believe I gained some perpetual skills from him, such as putting the linguist mindset to use where needed and reconstructing my ideas continuously for even better.

I would like to express my appreciation to my committee members, Assist. Prof Ömer Demirok and Prof. Hüseyin Cem Bozşahin. Furthermore, I benefited highly from the knowledge and wisdom of every professor and colleague I encountered in my academic life. I wish to convey my gratitude to dear Assoc. Prof. Mine Nakipoğlu for imparting to me a science-full outlook on life and to Assoc. Prof. Elena Guerzoni for joyful conversations and support as an advisor. I would like to extend my thanks to Assist. Prof. Pavel Logačev for his invaluable insights into how meaningful the meaningless numbers are and to Assist. Prof. Nazik Dinçtopal Deniz for aiding me in improving myself in academia and psycholinguistics.

I was very lucky to get to know David Berenstein and to spend time together. I acquired significant perspectives and viewpoints from him regarding not only the industry but also how to keep a large mechanism working. It is imperative for me to thank all the friends at Argilla for giving me this opportunity and for the learning and life experience there.

Endurance is not something that can be preserved without one's dearest companions. Serkan and Aybike were always with me on this journey; I had the noble opportunity to tell them about my problems and also not to talk about them if I do not want to. Kadir, Tarık, Samet, and Adar were those supporters whose friendship has been a source of strength and comfort, for which I am profoundly grateful. Ayşe and Rüstem were the people with whom I laughed, the indispensable component of this journey.

I would like to thank my family, who never ceased to believe me and always supported me in any decision I took. Their belief in my abilities has been a driving force behind my pursuit of excellence and I am deeply grateful for their faith in me. Lastly, I want to thank Bulut for cheerful chirps and for gracefully flying over me.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION ..... 1

CHAPTER 2 LITERATURE REVIEW ..... 4

2.1 Implicatures within the Gricean framework ..... 4

2.2 Scalar implicatures ..... 7

2.3 Natural language inference (NLI) ..... 19

CHAPTER 3 DATASET: ImplicaTR ..... 27

3.1 Making of ImplicaTR ..... 28

3.2 The structure of ImplicaTR ..... 29

3.3 The structure of scales ..... 30

3.4 Linguistic categories ..... 37

CHAPTER 4 EXPERIMENT 1: TRAINING LLMS ON IMPLICATR ..... 54

4.1 Splitting the dataset ..... 54

4.2 BERT models ..... 55

4.3 Generative models ..... 64

4.4 Discussion ..... 70

CHAPTER 5 EXPERIMENT 2: LINGUISTIC INQUIRY ..... 72

5.1 Splitting dataset ..... 73

5.2 Training and testing the models ..... 76

5.3 Results ..... 84

5.4 Feature importance analysis ..... 93

5.5 Further linguistic inquiries ..... 102

5.6 Discussion ..... 108

CHAPTER 6 CONCLUSION ..... 112

APPENDIX A SCALES BY LINGUISTIC CATEGORY ..... 115

APPENDIX B SAMPLE LLAMA-2 COMPLETIONS ..... 120

APPENDIX C CONFUSION MATRICES ..... 121

APPENDIX D FEATURE ANALYSES ..... 123

REFERENCES ..... 126

LIST OF TABLES

Table 1. Base Sentences and the Square of Opposition ..... 33

Table 2. Inferences between Premise and Hypothesis (For Adjectives, Verbs, Quantifiers, and Modals) ..... 34

Table 3. Implicature Tests for a Set of Sentences ..... 36

Table 4. Scale, Set, Row Statistics per Linguistic Categories ..... 40

Table 5. Adjective Scales in ImplicaTR ..... 43

Table 6. Verb Scales in ImplicaTR ..... 44

Table 7. Quantifier Scales in ImplicaTR ..... 45

Table 8. Tenses and Aspects in Modal Scales ..... 47

Table 9. Modal Scales in ImplicaTR ..... 48

Table 10. Numeral and Partitive Scales in ImplicaTR ..... 52

Table 11. Train, Validation, and Test Datasets for Models in Experiment 1 ..... 55

Table 12. Accuracy Scores of BERT Models on Test Dataset before Finetuning ..... 57

Table 13. Training Hyperparameters for BERT Models ..... 58

Table 14. Training Results of BERT Models ..... 60

Table 15. Accuracies of Finetuned BERT-NLI with Different Labeling on XNLI, and Original BERT-NLI Score ..... 63

Table 16. Accuracies of Finetuned BERT-NLI with Different Labeling on MNLI, and Original BERT-NLI Score ..... 64

Table 17. Train, Validation, and Test Datasets for Generative Models ..... 66

Table 18. Accuracy Scores of Generative Models Before Training ..... 67

Table 19. Model Descriptions of Generative Models Used in Experiment ..... 67

Table 20. Train Loss for Generative Models ..... 68

Table 21. Linguistic Categories Used for Training and Testing for Each Model..... 73

Table 22. Train, Validation, and Test Datasets for Models 1-5 ..... 74

Table 23. Train, Validation, and Test Datasets for Model-6..... 75

Table 24. Scales Tested in Model-6 ..... 76

Table 25. Train and Validation Loss Values of Models 1-5 ..... 78

Table 26. Test Results of Models 1-5..... 80

Table 27. Train Loss, Val Loss, and Val Accuracy for Model-6 ..... 83

Table 28. Test Results of Model-6 ..... 83

Table 29. Chi-Square and Cramer's V Results..... 85

Table 30. Accuracies and Statistics for Models 1-5 ..... 86

Table 31. Accuracy Results by Label Classes for Models 1-5 ..... 88

Table 32. Accuracy Results by Sentence Types for Models 1-5 ..... 89

Table 33. Precision by Label Class for Models 1-5 ..... 90

Table 34. Features Extracted for Analysis..... 98

Table 35. Classification Results of Logistic Regression Model ..... 99

Table 36. Classification Results of Random Forest Model..... 101

Table 37. Accuracy Scores by Verb Aspect in Modals..... 104

Table 38. Accuracy Scores by Quantificational Scales..... 104

Table 39. Accuracy Scores of Partitives by Label (Predicted by Model-1)..... 106

Table 40. Accuracy Scores of Partitives by Sentence (Predicted by Model-1)..... 106

Table 41. Accuracy Scores of C Implicatures per Sentence..... 107

LIST OF FIGURES

Figure 1. Square of opposition ..... 31

Figure 2. BERT models' train loss ..... 58

Figure 3. BERT models' validation loss ..... 59

Figure 4. BERT models' accuracy scores ..... 59

Figure 5. Model performance metrics on test set ..... 60

Figure 6. Finetuned model performance metrics on test set ..... 69

Figure 7. Train loss for the models 1-5 ..... 77

Figure 8. Validation loss for the models 1-5 ..... 78

Figure 9. Accuracy scores of the Models 1-5 (validation sets) ..... 79

Figure 10. Train loss, val loss, and val accuracy for Model-6 ..... 82

Figure 11. Confusion matrix of predicted and true labels across models 1-5 ..... 89

Figure 12. Confusion matrix of predicted and true labels for verbs ..... 91

Figure 13. Confusion matrix of predicted and true labels for numerals ..... 91

Figure 14. Confusion matrix of predicted and true labels by Model-6 in numerals . 93

Figure 15. Coefficients of features by the logistic regression model ..... 99

Figure 16. Random forest coefficients ..... 101

# CHAPTER 1

# INTRODUCTION

What we mean consists of two components: what we say and what we refrain from saying. We do not know for sure why we refrain, but we have formalized the second component within pragmatics with the exceptional work of Paul Grice, conversational maxims and implicatures. Conversational maxims are the maxims that any conversation between humans obeys if it is a cooperative conversation. An outcome of this line of work is the implicatures, those unuttered sentences that the other party infers from what is said.

This thesis investigates a specific type of implicatures: scalar implicatures. Scalar implicatures arise when the hearer believes that the speaker uttered a sentence because any alternative sentence with a more intense meaning is not true. When Andy tells his friend Jack that he saw *some* of the new *Star Wars* movie, Andy also conveys that he did *not* see *all* of the movie so that they can watch it together. The quantifiers *some* and *all* establish a pair here that we call a scale, which gives rise to a scalar inference.

As interesting as the implicatures are for human conversation, they are equally intriguing for machines. This thesis investigates the scalar implicatures in the realm of Natural Language Inference (NLI). It has been shown that Large Language Models (LLM) can perform pragmatic reasoning over many different linguistic phenomena (Yu et al., 2023). An investigation of the implicature reasoning of LLMs would reveal how they comprehend the implicature relationships. Conversely, implicatures are often subject to discussion in linguistics due to their complex nature

that relates to both semantics and pragmatics. Such a study can shed light on the nature of this debate from the computational perspective.

As the product of this study, we developed *ImplicaTR*, a Turkish NLI dataset specifically constructed for scalar implicature inferences. For ImplicaTR, we first developed base sets of sentences consisting of quadruplets of sentences:

- A sentence denoting the affirmation of the weaker term of a scalar pair,
- A sentence denoting the affirmation of the stronger term,
- A sentence denoting the negation of the weaker term,
- And a sentence denoting the negation of the stronger term.

We created the premise-hypothesis pairs by combining them with the help of the Square of Opposition (Horn, 2006a). We included the conventional NLI classes and a novel implicature class amounting to a four-way classification task: entailment, neutral, contradiction, and implicature. ImplicaTR was systematically created to include several linguistic categories for linguistic inquiry. ImplicaTR is vitally significant in both NLI research and Turkish Natural Language Processing (NLP). There are few scalar implicature datasets even including English, while the existing Turkish NLI datasets are either small or of relatively poor quality, mainly due to the fact that they are machine translations of large English datasets and do not undergo subsequent checks. ImplicaTR fills the gap in both fields and offers opportunities for further research.

The structure of the thesis is as follows: Chapter 2 reviews the literature for every aspect of the study, including the Gricean framework, scalar implicatures, and the NLI research. We first look into the very basis of implicatures: the Gricean framework. Then, we elaborate on the scalar implicatures within the Gricean theories

and discuss the localist-globalist arguments. The literature of NLI research and how scalar implicatures are treated in NLI is the last section of this chapter.

Chapter 3 is about ImplicaTR, the Turkish NLI dataset that features the implicature class. The technical details in dataset creation are touched upon before detailing the structure of the dataset. Scales and scalar pairs, which are the fundamental building blocks of the dataset, are clarified regarding how we treat them in data generation. We also discuss the base ImplicaTR, which has the four sentence types along with implicature tests. Finally, we will amplify how the linguistic categories are selected for the dataset and what their intricate structures are.

The following two chapters cover the experiments. In Chapter 4, ImplicaTR is used to train various LLMs to show how models perform against implicature data. Specifically, three BERT models are trained and evaluated first; subsequently, we look at how several generative models perform with the same dataset. In Chapter 5, we do an ablation study, where some linguistic categories are removed from the training set and subsequently tested. We analyze how generalizable the scalar implicature reasoning is by employing this method for each linguistic category. Another model is trained with a training procedure where some scalar pairs are lacking in order to see how the model generalizes its reasoning within the same linguistic category. We will then look into the detailed analysis of predictions given by each model and continue with a feature analysis, where we examine how the linguistic structure affects the model performance. This chapter is finalized with four small experiments that are conducted through limited manipulations within ImplicaTR. Lastly, in the last chapter, we discuss the general results and reflect our concluding remarks before shedding some light on further research.

# CHAPTER 2

# LITERATURE REVIEW

This chapter aims to introduce the implicatures within Grice's work and inspect various implicature types discussed in the literature. Then, the scalar implicatures are founded on the grounds of these implicatures and are elaborated on. Finally, this chapter concludes with the complementary part of this study, Natural Language Inference (NLI), in terms of its methods and significance for the study.

## 2.1 Implicatures within the Gricean framework

The meaning of an utterance has been the topic of excessive debate within semantics and pragmatics, contributing especially to the semantics-pragmatics interface. Specifically, whether the utterance of a sentence is the literal meaning of the proposition denoted by the sentence or there are other layers of meaning behind a sole utterance has been a long-standing discussion in philosophy and linguistics. In this field of work, Paul Grice has been thought by many to be the pioneer of the job achieved. He put forward the Conversational Maxims that are obeyed in conversation and proposed the concept of implicatures, which are the meaning not conveyed in what is said but conveyed in what is implicated (Grice, 1975).

Within the Gricean framework, implicatures are mainly classified into two types: conventional and conversational, which stem from different sources. Conventional implicatures are not based on conversational maxims, and they occur "conventionally" with the use of certain implicature-raising lexemes. Such an implicature is context-independent, so the utterance of the same sentence in any context will cause the same implicature. In sentence (1), the use of the word *but*

gives rise to an implicature: “There is a negative correlation between being poor and being happy.” (Gazdar, 1983).

(1) Emily is poor *but* happy.

Conventional implicatures cannot be canceled if they are ever implicated by an utterance. So, a sentence such as “She is poor but happy; however, I do not consider that being poor is negatively correlated with being happy” would sound odd in the same context. However, such an implicature is detachable (Horn, 2006a). Being detachable signifies that the truth-conditional sentence can be formed in the same context without arising the same implicature. Thus, the sentence “Emily is poor and happy” carries the same truth-conditional value as (1) but does not raise the implicature mentioned (Culpeper & Haugh, 2014).

On the other hand, conversational implicatures are context-dependent in that the particular implicature may arise in a given context with the given utterance, while it will not in different contexts with the same utterance. The example taken from Horn (2006a) shows how such an implicature arises. Here, (2b) is a potential implicature of the first sentence in that the speaker might actually mean (2b) by stating how good her personality is.

(2) a. She has a good personality.

b. She’s not that attractive.

Unlike conventional implicatures, conversational implicatures can be canceled with further sentences and contexts. Therefore, as the sentence “She has a good personality, and she is really beautiful” does not result in any incorrect or odd meaning, (2b) can be regarded as a conversational implicature. Unlike conventional implicatures, which are detachable, conversational implicatures are not detachable from the utterance. This is to say that if a particular implicature arises from a

sentence, that implicature will always arise by means of the sentences with the literal meaning and truth-conditional value in the same context. Consequently, it is seen that the conventional implicatures are non-cancelable and detachable, whereas conversational ones are cancelable but non-detachable.

Grice makes a further distinction between conversational implicatures: particularized conversational implicatures and generalized conversational implicatures. The example in (2a) is an exemplar of the particularized ones. This type of implicature is quite dependent on the context in which the first sentence is uttered, and the hearer must have this contextual information to infer the implicature. On the other hand, generalized conversational implicatures do not depend on contextual knowledge as much, and a similar implicature arises in many different contexts. Here is an example of such a generalized conversational implicature taken by Horn (2006a):

(3) a. The cat is either in the hamper or under the bed.

b. I don't know for a fact that the cat is under the bed.

Upon the utterance of the sentence (3a), the implicature in (3b) can be inferred without any prior knowledge regarding the context. Unlike particularized conversational implicatures, generalized ones can arise in any context with the same utterance, which makes us conclude that the difference between the two is the prior knowledge about context.

The following section introduces the scales and scalar implicatures and discusses the position of scalar implicatures as a type of generalized conversational implicatures.

## 2.2 Scalar implicatures

Scalar implicatures have been examined in the literature from a pragmatic point of view starting from Grice's framework of cooperative principle and implicatures, as we have discussed so far (Grice, 1975). As the pragmatic view fails to account for some cases where scalar implicatures do not seem to stem from the cooperative principle, different approaches have been proposed to govern (thoroughly or at least partially) how scalar implicatures arise (Chierchia, 2004; Levinson, 2000). As we will go through in this chapter, any framework or school of thought is likely to fail while explaining the whole phenomenon, which is an indication that the subject of scalar implicatures (or implicatures in general) is sailing between different lands, specifically between pragmatics and semantics.

### 2.2.1 Scalars and Horn scales

Before constructing a basic explanation for scalar implicatures, let us look at the scales and scalars. A scale (or a Horn scale) is a set of ordered words where the words are sorted according to their semantic intensity (Horn, 1972). Such an example would be the scale of <*warm, hot*>, where the first element of the set is a weak term compared to the second element. In other words, the second element *hot* is a word stronger than *warm* in the context where they are related to each other (e.g., the temperature of a room or a cup of coffee). When other meanings of the individual words are at hand, the scale relationship between the two might not hold, which is the case most of the time. The sense of *hot*, when it is used in sentences like *We have some hot topics to discuss*, is clearly out of a scalar relation with *warm*. Adjectival scales, in particular, are argued to have three aspects that create the basis for the

scalar relationship: degrees, dimension, and an ordering relation (Kennedy & McNally, 2005).

One of the primary examples of scales is the scale of quantifiers *some* and *all*. Following from logic, *all* is a term that is stronger than *some* in the dimension they both share. Thus, a scale of the pair can be constructed as <*some, all*>. A scale can be increased in the number of elements it has given that the relationship among them still holds. In our case, a possible scale with three elements can be <*some, many, all*>. An essential caveat while building the scales is that the stronger element must entail the element(s) in the scale weaker than itself when the scale is positively constructed. The sentences below are an example for the scale of <*some, many, all*>:

(4) a. Some people watched the show yesterday.

b. Many people watched the show yesterday.

c. All of the people watched the show yesterday.

In a context where (4c) is true, it is undoubtedly the case that (4b) is true as well, which is also the case for (4a). Similarly, the existential *some* must hold whenever we utter *Many people watched the show*. One problem with this definition is the case of antonyms on the same scale, an apparent weakness of the system (Horn, 1972). Intuitively, such a scale can include *none*, forming the scale of <*none, some, all*>. This holds without a problem from a logical point of view with the reasoning that *some* is a term with a more intense sense than *none* and *all* is similarly the most intense of the three (Hirschberg, 1985). However, the condition that the stronger terms in the scale must entail the weaker ones causes problems. Let us adjust the sentences in (4) with the new scale:

(5) a. None of the people watched the show yesterday.

b. Some people watched the show yesterday.

c. All of the people watched the show yesterday.

What one should expect from the sentences in (5) is that (5c) should entail (5b) and (5a), and (5b) should entail (5a). *None* (or *no*) refers to the non-existence of something, whereas *some* is the exact opposite of *none* and means the existence of something. In a sense, then, *some* comes to the meaning "*not none*", which is constructed by negation. As Hirschberg (1985) states, if we are to maintain that there must be an entailment relationship between the scalar elements, then we have to eliminate negative poles in the same scales to prevent the contradiction.

While the *none* and *some* case can be eliminated from the scene with the assertion that they are each other's negated versions, the issue is a bit more intricate for other scales, such as adjectival scales. <*pretty, beautiful*> can be considered a scale where beautiful entails its weaker version *pretty*. The same intuition above might result in the fact that <*ugly, pretty, beautiful*> is another valid scale where the intensity of each word increases to the right. To clarify why this triple might possibly create a scale, imagine that someone is scoring people, unfortunately, according to the features that each person has. A score of two out of ten might qualify as *ugly*, while five is *pretty*, and nine is *beautiful*. Now, the very first intuition that *beautiful* and *ugly* cannot be on the same scale is gone because the two terms are assembled on a single line of numerals or on a single *dimension* (Kennedy, 2007). When examined from the semantics of two words, they are clear antonyms. Being *prettier* will result in being *beautiful*, while being *uglier* will never lead to a transformation into being *beautiful*. However, gathering them together on the same line (0-10) might give rise to the perception that *ugly* is a concept that differs from *beautiful* only by intensity.

A potential solution for this discrepancy might be the use of middle terms with both sides of the pole rather than creating one single scale. Thus, we can create the scales of <*pretty, ugly*> and <*pretty, beautiful*>, in both of which the intensity increases going from left to right. One question here might be whether *pretty* and *ugly* are in such a scalar relationship. In a potential context where beautiful examples of something are elected, saying that something is *pretty* comes to the meaning that it is not *beautiful* and even maybe *ugly*. Therefore, we can hold the scalar relationship between the two elements while also preserving the entailment condition that Horn (1972) proposes.

### 2.2.2 Structure of scalar implicatures

In the previous section, we have established the concept of scalars, which are the basis of the scalar implicatures. Scalar implicatures arise when a speaker utters the weaker element of the scale, and the hearer infers that the stronger element is not true. Formally, a definition in (6) can be proposed for scalar implicatures (Russell, 2006; Van Rooij & Schulz, 2004):

(6) Given the scale <α, β>, if Speaker A utters that α is true, Hearer B infers that β is false under Grice's maxims and the Cooperative Principle.

Acknowledging first the scale definition that the stronger element entails the weaker one, the definition given in (6) is formed on the basis of Grice's principle of cooperativeness (Grice, 1989). The Hearer B must assume that Speaker A is in cooperative communication and obeys the maxims. If so, B also assumes that A will utter that β is true in any case where β is true. Upon hearing that A has said that α is true, B will deem that it is not the case that β is true. In another sense, the utterance

of the weaker item leads to the implication that the negated version of the stronger item is true, again when the scalar items are not negated.

This approach towards scalar implicatures relates to the Maxim of Quantity by Grice (Sauerland, 2012):

(7) The first Maxim of Quantity:

Make your contribution as informative as is required. (Grice, 1989)

Given the scale of <*some, all*>, it can be stated that *all* is a word that is more informative than the word *some*, thus stronger. In obeyance to the Gricean maxims, the more informative alternative must be uttered where applicable when it is true. In cases where this is not observed, an automatic conclusion would be that the more informative alternative is unavailable in the particular context. Thus, quantity implicatures also seem appropriate for defining those implicatures formed out of scales, along with the term scalar implicatures (Horn, 2006a).

Additionally, it might be the case that the speaker lacks the information to utter a more robust alternative, which would be another depiction of how a scalar implicature is formed. This exemplifies a case where a scalar implicature emerges based on the second maxim of quality (in (8)). In this case, the speaker might choose to refrain from asserting the alternative *all* as they lack information with respect to whether it is true or not.

(8) The second maxim of quality:

Do not say that for which you lack evidence. (Grice, 1989)

To illustrate the case in Turkish, let us examine the case of the scale <*bazı, bütün*> (<some, all>):

(9) a. Bazı müzeler kapatıldı.
*Some museums have been closed.*

b. Bütün müzeler kapatıldı.

*All of the museums have been closed.*

c. Bütün müzeler kapatılmadı.
*Not all of the museums have been closed.*

Just like in English, *bazı* means the existence of something, while *bütün* is a stronger term in the same dimension. When the bidirectional relationship between the two sentences is examined, (9b) entails (9a). On the other hand, the use of (9a) creates the implication that (9b) is not true. More specifically, the use of (9a) implicates (9c).

### 2.2.3 Scalar implicatures within generalized conversational implicatures

The positioning of scalar implicatures within the hierarchy of implicatures proposed by Grice (1989) and later developed by Levinson (2000), among many others, seems relatively clear among many different theories. The functioning of scalar implicatures within a sentence or a discourse resembles how generalized conversational implicatures work in similar contexts, which is why scalar implicatures are classified within generalized conversational implicatures in general (Potts, 2015). This tendency has also led scholars to examine the concept of scalar implicatures by means of the general pragmatic theories that hold for implicatures in general. These theories heavily depend on the context where the utterance is made, thus warranting a pragmatic view towards the issue (Sperber & Wilson, 1995). However, later developments have shown that the issue of scalar implicatures might not be solely pragmatic and a semantic-based theory or even a grammaticist theory (Chierchia et al., 2012) might be required. Theoretical and experimental work has shown that any approach to the issue is likely to fail in a particular case, so a hybrid theory might satisfy as many disputes as possible (Katsos & Cummins, 2010; Sauerland, 2012).

Before advancing with generalized implicatures, let us note again how particularized conversational implicatures are constructed. Below is an example taken from Carston (2004):

(10) a. Ann: Does Bill have a girlfriend these days?

b. Bob: He flies to New York every weekend.

c. *Implicature*: Bill (probably) has a girlfriend in New York.

The response that Bob gives Ann does not answer the question when the semantics is taken into account solely. As the response does not have a yes/no response in the first place, Ann will need to make some other processes to find out what Bob actually *said* (compared to what has been uttered). Similar to the case with generalized conversational implicatures and scalar implicatures, this response is not in an entailment relationship with another form of response that is semantically complete. This is to say that the response at (10b) is not obligated to mean either yes or no to the question Ann asked. In either case, after uttering (10b), Bob has the opportunity to state that Bill has a girlfriend or Bill does not have a girlfriend. This signifies that, even if an inference whose semantic meaning is different from the uttered version has been derived, the speaker has the chance to cancel it, which actually makes this inference an implicature. As we do not depend on any lexical or syntactic clue while deriving it, it can be named particularized conversational implicature (Grice, 1989).

When the context arguably plays little to no role and it is some lexemes (or some other syntactic structures) that automatically derive the implicatures, they are called generalized conversational implicatures (Horn, 2006b). The example below demonstrates a generalized conversational implicature (Carston, 2004):

(11) a. Tim turned the key and the engine started.

b. *GCI*: Tim turned the key thereby causing the engine to start.

The semantic resolution of (11a) will generate a context where two different actions happen: Tim turning the key and the engine starting. Any hearer will probably consider that the reason why the engine started is because Tim turned the key. This causal relationship is what the word provokes in the sentence, and it cannot be proposed that the relationship is peculiar to the context of this sentence. When two propositions are combined with the conjunction, the hearer is inclined to infer that the second event is the result of the first, while this might not always be the case. As (10c) can be eliminated from the meaning of (10a), a sentence such as *Tim turned the key, and the engine started; however, it was because of a short-circuit that* will cancel out the implicature in (11b).

With this example's idea that we can cancel the implicature meaning, we can consolidate the place of scalar implicatures within conversational implicatures. Scalar implicatures are also subject to conversational implicatures of cancelability, suspension, or re-enforceability (Potts, 2015). These properties have been used to test if a particular inference is the implicature of a sentence or not (Capone, 2006; Sauerland, 2012). These tests are critical in assigning the inference type of a sentence as they have been shown to be quite reliable by the authors. The sentences below exemplify a simple case of a test:

(12) a. Premise: Some of the students are here.

b. Implicature: Not all of the students are here.

c. Cancelation: Some of the students are here, even all of them.

d. Suspension: Some of the students are here, maybe all of them.

e. Reinforcement: Some of the students are here, but not all of them.

The scalar term *some* in the premise gives rise to the implicature stated in (12b). As it is a conversational one, this implicature can be canceled by stating that the stronger term is actually true as shown in (12c). In (12d), the speaker utters their suspension regarding the truthfulness of the implicature raised. These two tests are similar to each other in that they denote a suspension on the part of the speaker, while one test is stronger than the other. Similarly, the fact that implicatures are not a part of the semantic meaning can also be seen in the reinforcement test in (12e). Here, the implicature that all is not true, hence not all, is further reinforced with the conjunction but without creating any problem in the sentence.

These tests can be considered evidence regarding the position of scalar implicatures within the implicatures or Gricean framework. However, these perspectives over scalar implicatures do not go without counterarguments that question the place of scalar implicatures. This debate can be called the localists vs. globalists dispute, which has its roots in how they are derived.

### 2.2.4 How scalar implicatures are derived: Localists vs. Globalists

The basic examples so far demonstrated how we derive scalar implicatures from sentences as in the example (13) below, whose implicature would be *John did not watch all of the movies*:

(13) John watched some of the movies.

This implicature is Gricean in the sense that it is generated post-propositionally, meaning that the pragmatic system assessed the whole proposition before generating the implicature. A sentence such as (14), however, creates a problem for this view:

(14) Emily believes that John watched some of the movies.

Following the Gricean way (Geurts, 2009), we encounter an incorrect implicature: *The speaker believes it is not the case that Emily believes that John watched some of the movies* instead of *Emily believes that John watched not all of the movies*. This discrepancy regarding the scalar implicatures created some lines of work that deviated from the Gricean pragmatics. As Geurts (2009) states, some diverged to some extent from the original Gricean proposal to account for such cases (Chierchia, 2004; Landman, 2000; Levinson, 2000) while some others revised the pragmatic view on the issue (Russell, 2006; Sauerland, 2012; Van Rooij & Schulz, 2004). As a result, we encounter several theories over the nature of implicature derivation stemming from how scalar implicatures are derived in different contexts.

The approach towards the derivation of scalar implicature that has been discussed here so far is generally in line with the globalist view (Lee, 2008). Sauerland (2012) calls the basic globalist view as the pragmatic theory, though this title might be used to refer to some other ideas in other contexts. According to the globalist view or pragmatic theory, in implicature derivation, commonsense reasoning plays the key role: depending on the Gricean maxims, the hearer continuously analyzes the alternatives of an uttered sentence and makes conclusions about why any specific alternative has not been said. The conclusions the hearer makes are the scalar implicatures of that particular sentence. Globalists, as well as those from other perspectives, recognize the existence of scales that are formed by lexical items. When the weaker term is uttered, the pragmatic processing concludes that the stronger term is either false, unknown, or something in a similar fashion, which would be described by Gricean maxims.

This view is globalist in the sense that it needs entire speech acts to run the reasoning process. This aspect has commenced debates over the applicability of

assuming totally pragmatic causes in implicature derivation, as its inability to generalize to other cases is shown in (14).

Several authors proposed a different line in implicature derivation, which is known as the localist view (Chierchia, 2004; Levinson, 2000). This view argues that pragmatic processing always happens when a scalar item is confronted, and the result is always the same, that the stronger version of the scalar item is not true. Then, they question the validity of additional and tiresome processing for scalars, instead of which they propose that the result of this processing is saved and used afterward every time the particular scalar item is encountered (Sauerland, 2012). Therefore, scalar words are stored in the lexicon with their scalar meaning, and they are activated every time when the scalar word is encountered. This is the reason why the localist view can overcome the problems demonstrated in (14). Sauerland (2012) proposes a hybrid they named pragmatic + lexical theory that accounts for the shortcomings of both theories.

Among other theories, a relatively radical one is the grammatical theory, which is even more divergent than the pragmatic principles (Chierchia et al., 2012; Fox, 2007). According to this view, there are Exh (Exhaustivity) and O (Only) operators that are responsible for implicature derivation. While it can predict both local and global implicatures, Sauerland (2012) asserts that this view needs empirical evidence since the previous theories rely on commonsense reasoning, unlike the grammatical one. Moreover, a globalist and a harsh opponent of the grammatical view Geurts (2009) finds this “unfortunate” perspective wrong and “pointless”, stating that there is no need for any inclusion of syntax while pragmatics can account for any case.

Though not on the same line of discussion, a related proposal is defaultism, which basically discusses that the upper-bounded meanings (some but not all) of non-maximal scalar items (some) are the default ones (Breheny, 2019). According to this view, generalized scalar implicatures (thus scalar implicatures) are generated automatically without any effort because they are stored in the lexicon. What is effortful is the cancelation of the derived implicatures. On the other hand, particularized conversational implicatures, which are called *ad hoc* implicatures, are derived with extra pragmatic processing. This view is seen generally hand in hand with localism, thought they do not necessarily entail each other (Katsos & Cummins, 2012).

These theoretical developments have also been subject to different experiments. As complex as the literature is, the experiments to test the diverse set of theories are complicated, too. Some studies showed that the resolution of scalar implicatures is not totally pragmatic (Papafragou & Musolino, 2003; Papafragou & Schwarz, 2005). In a series of experiments done with children on numeral scalars, numerals did not show 'at least' meaning in general, and children generally evaluated it with the upper-bound 'exact' meaning. While there is a difference between numerals and other scalar items, the role of semantics is shown in implicature derivation. In another study, Shetreet et al. (2014) conducted an fMRI study on scalar implicatures and found that the IFG (BA 47) area of the brain is involved in implicature derivation. In previous studies, this area is linked to semantic processing rather than pragmatic processing, which seems consistent with the grammatical theory the most, as well as other theories to some extent.

While these studies do not rule out the inclusion of pragmatic processes in scalar implicature derivation, some other studies argue a larger effect of pragmatic

reasoning for implicatures. In a sentence verification task, participants judge sentences as true or false depending on the implicature inferences they have (De Neys & Schaeken, 2007). It was seen that implicature derivation is not automatic and it is effortful, and participants were shown to use pragmatic reasoning more than logical reasoning, which led them to correct answers. Huang and Snedeker (2009) conducted an eye movement experiment and found that participants showed both semantic and pragmatic reasoning, the latter causing more delay. Lastly, in psycholinguistic research, Katsos (2005) conducted both offline and online judgment and comprehension experiments. Their finding was that pragmatic reasoning is highly used by the human parser in scalar implicature derivation. Consequently, studies do not clearly favor any side in the debate, and further experiments are needed.

## 2.3 Natural language inference (NLI)

Natural language inference can be defined as a classification problem where the task is to infer the relationship between two propositions/sentences (Gubelmann et al., 2023). These two sentences are mostly referred to as premise and hypothesis in the literature, which are exemplified below:

(15) a. Premise: I saw the man entering the green house.

b. Hypothesis: The man entered a house.

The classes into which the inference between the premise and hypothesis is to be classified are either entailment (if P is true, then H must be true as well in any circumstance), contradiction (if P is true, H cannot be true under any circumstance), and neutral (neither entailment nor contradiction). Because of this classification structure, NLI is also viewed as Recognizing Textual Entailment (RTE) (Dagan et al.,

2013; Poliak, 2020). Recognizing these inferences is an inevitable aspect of human linguistic communication, which makes the field an indispensable part of other NLP tasks and problems such as information retrieval, question answering, summarization, or translation. In this regard, NLI becomes a multidisciplinary field contributed to by logic and philosophy as well as NLP and linguistics (Korman et al., 2018).

Yu et al. (2023) view NLI as part of a more general task called natural language reasoning, for which they offer several definitions. Reasoning is something more than a mere understanding of the text; it includes generating a new assertion based on the given assertions. In addition, according to them, NLI has three main entailment relationships in which the task is validating a paraphrase of the premise, semantic understanding of compound sentences, or reasoning over the implicit meaning within the premise. Thus, the task of inferencing the entailment relationship involves comprehending the implicit meaning dimensions, such as pragmatics.

With the spread of Transformers (Vaswani et al., 2017) in the NLP realm, NLI research has been mostly carried out with LLMs such as bidirectional models like BERT (Devlin et al., 2019) or generative models like GPT (Radford et al., 2018). The main reasons for this include the representation of language by these models in a high-dimensional space, their ability to understand the implicit knowledge within the text, and their ability to adapt many different tasks easily with few-shot or zero-shot methods (Kojima et al., 2023; Yin et al., 2019; Yu et al., 2023). What Transformers changed in NLP and in artificial intelligence research in general was the use of self-attention mechanism so effectively that the models then accomplished to retain much larger spans of knowledge in their ‘attention’, which made them comprehend and produce larger contexts.

As a Transformer-based architecture, BERT was developed by implementing a masked language modeling task. In this task, the sentences are given to the model with a random word in it being masked. Then, the model predicts the masked word by taking the left and right contexts of the masked word. Considering that words are represented in high-dimensional spaces with word embeddings, BERT has been extensively used in NLP research for its deeper understanding of the text at hand (Cho et al., 2021).

A typical method in NLP workflows after Transformers has been to use the pre-trained models in fine-tuning the task for downstream tasks. Called the *pretrain-finetune paradigm* (Hupkes et al., 2023), this method offers the advantage of employing the general language understanding of a model, obtained in the pretraining phase through massive unsupervised learning on language, by subsequently adjusting the weights according to the task at hand. Therefore, this process includes cross-task generalization. This paradigm also offers a greater ability to inquire about the linguistic capabilities of the model. It has been shown that the weights learned during pretraining help BERT or other models to generalize better, especially for linguistic data (Hao et al., 2019).

### 2.3.1 NLI datasets

Upon the proposal of the textual entailment paradigm (Glickman & Dagan, 2005), where the truth of hypothesis *h* is inferred from the text *t*, various benchmarks were proposed to assess this novel NLP task. Several datasets were gathered for benchmarking with PASCAL, a challenge for textual entailment datasets (Dagan et al., 2006). These datasets have two classes: *entailment* and *not entailment*. Those

datasets that had been originally created with three classes were revised into two classes by combining *neutral* and *contradiction* into a single *not entailment* class.

Later on, the SNLI dataset (Stanford Natural Language Inference) (Bowman et al., 2015) was created for NLI research and can be considered the first large-scale NLI dataset with its 570k data size. Generated via image captions, this dataset involves three classes entailment, contradiction, and neutral. Crowdworkers are given a premise sentence, for which they are requested to write a true description sentence, a possibly true description sentence, and a definitely false description sentence, which constitute the hypothesis part. However, SNLI has a narrow set of styles and genres and may not be generalized to many different contexts.

A more advanced dataset is the MNLI (Multi Genre Natural Language Inference) dataset (Williams et al., 2018), which consists of 433k of data in the same vein as SNLI but covers more versatile genres and styles. Both MNLI and SNLI were deductively created since the hypotheses are generated on the basis of 'necessarily true or appropriate' valid inferences when the premise is true (Gubelmann et al., 2023). One of the problems with these datasets, in addition to having narrow genres, is the fact that the labels can be predicted with certain structures within hypotheses, such as negation for contradiction, superlatives for neutral, or generic nouns for entailment (Gururangan et al., 2018). It was shown that more than half of MNLI and around two third of SNLI can be correctly predicted with these constructions inside hypotheses.

To escalate NLI benchmarking to different languages, XNLI (Cross-Lingual Natural Language Inference Corpus) (Conneau et al., 2018) was created based on MNLI. The resulting dataset covers 15 languages, including Turkish. A second NLI dataset is NLI-TR (Budur et al., 2020), which would be the largest NLI dataset in

Turkish. It both involves SNLI and MNLI, which were translated into Turkish by Amazon Translate.

Pragmatics and implicatures are not represented in NLI research as much the entailment relationships; nonetheless, it is still possible to find several datasets featuring the implicature relationships. One such dataset is the Implicature dataset created by George and Mamidi (2020). In this dataset, the implicatures are given within a dialogue, unlike the previous datasets mentioned, where the label is the yes or no answer given to the first question. All of the implicatures in this dataset are examples of particularized conversational implicatures that are dependent on the context they are in. In a similar fashion, BIG-Bench (Srivastava et al., 2022) is a collection of datasets that includes fields ranging from child development to linguistics. As a part of this effort, a dataset of implicatures is generated, where the two-way dialogue involves a question about the particularized conversational implicature.

Another dataset involving conversational implicatures is the GRICE dataset for implicatures and conversational reasoning (Zheng et al., 2021). Authors use automated grammar to create open dialogues that also involve questions regarding each dialogue. As well as featuring contexts for the questions, this dataset is significant in that it includes generalized and scalar implicatures as well as the particularized ones. Jeretic et al. (2020) proposed the IMPPRESsive dataset that included both presuppositions and implicatures. Implicatures in this dataset are all scalar implicatures. In addition, they categorized the linguistic categories of the scalar items to further investigate the differences between categories.

### 2.3.2 Linguistic investigation with NLI

A number of NLP tasks have been used to investigate various linguistic phenomena, and NLI is one of the most used tasks among these with its adaptability to many different problems. Many subjects from syntax or semantics have been examined under the NLI roof or with other tasks (see Poliak (2020) for a list of various phenomena researched). Naik et al. (2018) carried out an extensive study on how the labels in NLI are predicted by models and what factors affect the predictions. They found that lexical similarity between premise and hypothesis is positively correlated with the model predicting entailment for the sentence pair. In addition, in one of the tests, one word from the premise is changed with its antonym and this new sentence is used as a hypothesis; and the model predicts the inference between the two. It is seen that the models overpredict the entailment label for these cases, which is explained by the fact that the sentences are quite similar to each other with only one word changing. The authors propose stress-tests to validate NLI datasets as they do with MultiNLI (Williams et al., 2018).

In another study, where monotonicity and negative polarity items (NPI) licensing is investigated, it is found that LLMs are successful in licensing the NPI items in different environments (Jumelet et al., 2021). Moreover, they are shown to process NPI items and monotonicity in similar manners in terms of the representational cues they use while there is also a difference between the resolution of environments with words from different parts of speech like adjectives or determiners. This study is important in that it shows that the models have generalization capabilities across different linguistic phenomena. In a similar study by Kalouli et al. (2022), authors examine how models treat function words and content words. Authors find that function words might not be captured well in terms

of their semantics. Models are argued to get distracted when the words in the sentence are semantically similar to each other, which in turn results in the fact that function words are not grasped to a higher extent by the model.

Rather than simply feeding the input data and making the model infer on it, other forms of inferencing have been suggested in the literature. In one of those studies, where the inferencing is supervised with human explanations, the authors found that the NLI models perform better with explanations (Stacey et al., 2022). With this method, prediction for a specific class is accompanied with an explanation over why the class can be or cannot be correct. They find that the words that receive the most attention in explanations are nouns.

Although NLP and NLI research have been relatively extensive on the subject of textual entailment, we see that pragmatics and implicature within NLI literature are scarce. As discussed earlier, implicatures, specifically scalar implicatures, are more granular comprehensions of the implicational relations among sentences, and their study would yield better representations for the models. In one study, authors question the LLMs in their capability to infer conversational implicatures (Kim et al., 2023). The experiment is structured around chain-of-thought prompting, where the inferencing is done step-by-step. This method explicitly demonstrates the inferential reasoning to the model and the model is expected to use the same answering style while labeling. They found that the chain-of-thought method increased the pragmatic capabilities of the model, and the model performed quite well on the conversational implicature dataset, even outperforming humans on the same task. Authors also claimed that models can create the lacking background information in the premise-hypothesis pairs by means of this method.

Lastly, in another study, where scalar implicatures and presuppositions are studied, authors found that models can perform pragmatic reasoning (Jeretic et al., 2020). In this study, they investigated both presuppositions and scalar implicatures by developing diagnostic datasets. One important point is that they categorized the dataset into different parts of speech, namely verbs, connectives, numerals, adjectives, modals, and determiners. This way, they were able to reveal whether different categories are treated differently in pragmatic terms. They found that the BERT model trained on MultiNLI was successful in reasoning over scalar implicatures formed by *some* and *all*. However, they could not find significant evidence to find out whether this is also the case for other categories or not. This can be accounted for because other scalar items do not create contrast between themselves with little to no change in the intensity in the meaning. As the authors state, models treat most of the scales other than *some-all* as synonymous in their study. Therefore, datasets that feature well-defined entailment and implicature relationships along with highly acceptable scales are required for implicature research within NLI.

# CHAPTER 3

# DATASET: ImplicaTR

This chapter provides the details of *ImplicaTR*, a granular NLI dataset in Turkish that is specifically designed for scalar implicatures. This dataset is an NLI dataset that consists of the conventional entailment, neutral, and contradiction classes, on top of which the current study adds another class of inference: implicature. Each row in the dataset includes a premise sentence and a hypothesis sentence. The inference between these two sentences is labeled as one of the four categories mentioned.

ImplicaTR has been developed to fulfill different needs from the related fields. Primarily, it can be seen that NLI literature lacks the theoretical work that has been developing in the field of semantics and pragmatics. Implicatures, one of the prominent subjects in the semantics-pragmatics interface, presents a distinct line of work for the NLI research with its more granular comprehension of the pragmatic inferences. NLI research conventionally includes three main inferences: entailment, contradiction, and neutral. However, as is shown in the following sections, scalar lexemes give rise to implicatures, which derive from the same foundations as the inferences mentioned. By adding another dimension to the regular inference types, we aim for a more granular and precise understanding of pragmatic inferences.

Our second aim in developing ImplicaTR pertains to the limited scope of NLI research in Turkish, which is characterized by a scarcity of studies and datasets. Turkish is different from English in many ways such as its rich morphology or its syntax, which mandates a special treatment of the linguistic phenomena instead of mere replications from English. In one such study, the fundamental datasets SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) are translated into Turkish

through Amazon Translate (Budur et al., 2020). Though a remarkable step in Turkish NLI studies, the resulting dataset might not reach the expected quality, which would decrease model performance. We consider that a dataset particularly curated for the linguistic characteristics of Turkish is significant in this sense.

In the next sections, we look at the structure of the dataset and the scales. After explaining how the scales are constructed, we further look into the linguistic categories employed in the dataset.

## 3.1 Making of ImplicaTR

ImplicaTR is a dataset created through a process that does not include any traditional data collection methods such as collecting from participants or annotators. Rather, it has been created by specific methodologies employed by those who have linguistic proficiency. In consequence, the dataset has a methodical and systematized structure within itself and it can be used for linguistic research as well as NLP studies.

Principally, each of the premise-hypothesis pairs is built on a particular scale, which is to say that the premise and hypothesis for the same item include the same scale. While either the stronger or the weaker scalar term is used in the premise, the other term is used for the hypothesis sentence. Also, the resulting set of sentences is created by manipulating the polarity of the sentence. This process can be considered semi-synthetic. All data items are created based on four sentences, which we call A, B, C, and D (to be elaborated on in the next section). Among these, only sentence A is manually generated. The sentences B, C, and D are the transformations of sentence A by means of the scalar term and negation. In this sense, after creating sentence A manually, we used GitHub Copilot, an autocomplete tool for code and natural text, to generate the rest of the sentences. GitHub Copilot uses GPT-4 (OpenAI et al., 2024)

as its LLM in the background. This process allowed us to very quickly create the sentence sets as well as offer more time to examine the resulting data to prevent any issues in the dataset.

The data generation process was conducted with two expert linguists. At each step of dataset creation, experts reviewed the resulting data, and a revision stage took place after every review. The aspects that were heavily paid attention to during these reviews were the acceptability of individual sentences and the inferences between proposed premise-hypothesis pairs. Due to the semi-synthetic generation method, it was possible that some pairs might lead to unexpected inferences out of our control, which is why we controlled every stage in data production. After the data creation process was completed, we checked a sample from the dataset to ensure the reliability of the resulting dataset. The resulting dataset has 20,340 data rows. With a 95% confidence level and a 2% margin of error, the sample consists of 2,148 randomly selected data items. After annotating the sample, we found that our dataset has a reliability score of 97.81% with the confidence level and margin of error said.

### 3.2 The structure of ImplicaTR

The dataset is built on 5 different scalar categories: adjectives, verbs, quantifiers, modals, and numerals. The scalar category is the linguistic category that determines the type of the scalar item so that we can differentiate between the effects of the different categories on the resolution of implicature understanding (Baker et al., 2009). We generally followed the work of Jeretic et al. (2020) to create the basis for categories (see section 3.4). Unlike the categorization in their work, we have not included the connectives in our studies. This is mainly because the connectives such as *veya* (or) or *ve* (and) are ambiguous in that the logical operator *or* carries the

inclusive and exclusive meaning (Hurford, 1974). The sentences below demonstrate the case of inclusive (16) and exclusive-or (16) meanings.

(16) a. There is milk or honey at home, maybe both.

b. There is milk or honey at home, but not both.

In the first sentence, both objects might exist with the inclusive meaning of *or*, whereas the second sentence rules out the cases where both of these objects exist at the same time. The use of exclusive meaning contradicts our aim in this study, which is basically presenting both the lower and upper bounds in the sentence and investigating if the use of a weaker/stronger scalar item leads to an inference of implicature. As Chierchia et al. (2001) asserts, implicature introduction should lead to a stronger statement without the implicature itself. In (16b), this cannot be true as the intrinsic meaning cancels the implicature itself. We have decided to exclude connectives from ImplicaTR to eliminate this mediating variable.

### 3.3 The structure of scales

Within each of the categories defined above, we created several scales that include scalar pairs in that particular category. A scale (or a Horn Scale) (Horn, 1972) is a set of two or more lexemes that are in a relationship of strength or intensity. In other terms, a scale consists of a weak and strong term. A well-known example of this is <*some, all*>, where *some* is the weaker term while *all* is a term stronger than *some* in the same dimension of meaning. The Turkish equivalent to this scale would be <*bazı, bütün*>.

In the data production stage, we first created two sentences with scalar terms of this kind, where both of the sentences are syntactically quite similar to each other and only differ in the scalar term they involve:

(17) a. Bazı oyuncular sakat.
*Some players are injured.*

b. Bütün oyuncular sakat.
*All of the players are injured.*

Then, we created the negated versions of the same sentences to be able to continue with the formation of the dataset:

(18) a. Hiçbir oyuncu sakat değil. = (17a')
*No player is injured.*

b. Bütün oyuncular sakat değil. = (17b')
*Not all players are injured.*

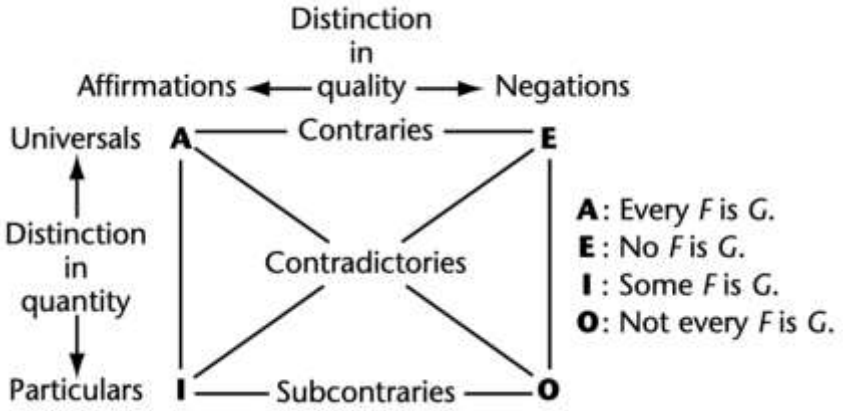



Figure 1. Square of opposition
Source: Horn (2006a)

The square of opposition, as shown in Figure 1, above plays a key role in the creation of the set of sentences. It is based on the logical relationships between different quantificational lexemes. A and O ends (as well as E and I ends) are in a contradictory relationship: it is not possible for these two pairs to be true at the same time or false at the same time. On the other hand, the contraries in this schema are the A and E ends of the square, where they cannot be true at the same time. There is a relationship of subcontrariety between I and O ends. It is possible that these two

cases are true at the same time, but they cannot be false simultaneously. Lastly, there is a relationship of subalternity between the nodes that are quantitatively distinct from each other: A and I, E and O. Subalternation relation refers to the cases where the superaltern being true also entails the truth of its subaltern whereas it is not the case in the opposite direction. While the superalterns A and E entail their respective subalterns, it is not a bilateral relationship. Thus, I and O do not entail A and E, respectively.

Used for research on entailment relationships as well as implicatures (Newstead & Griggs, 1983), the square of opposition offers a simple schema for implicature reasoning: the assertion of any of the subcontraries implies the other subcontrary. Considering that contradictory ends are just the negated version of each other, uttering one subcontrary leads to the inference in Gricean terms that the stronger (or more universal) version is not applicable in the context, which means that the negation of it is true. As Horn (2006b) argues, this square of opposition demonstrates interesting patterns cross-linguistically. While the A, I, and E ends of the square can be freely found to be lexicalized in many languages, the O node is almost never lexicalized, except for some cases like negated modality (De Haan, 2006). In English, for instance, *all*, *some*, and *no* exist while there is no direct lexeme **nall* which would mean *not all*. Apart from quantifiers, this pattern is also true for other categories such as **nalways, *noth,* or **nand.*

Based on the square of opposition and the relationships within it, we first created a scalar pair such as <*some, all*>. Then, we generated four sentences out of one scale: a sentence with the weaker item in the scale, another sentence, with the stronger item in the scale, another sentence with the weaker item in the scale under negation, and another sentence with the stronger item in the scale under negation.

Each of these four sentences falls on one corner of the square. We named the sentences A, B, C, and D for simplicity, which differs from the figure shown above. The resulting schema is as follows. As seen, the sentences are created in line with those in the square, and they preserve the same relationships as above. From now on, the sentences will be referred to with their letters in the current study: with A, B, C, and D.

Table 1. Base Sentences and the Square of Opposition

| Sentence | Equivalent in S.O. | Example Scale | Turkish Sentence | English Sentence |
|---|---|---|---|---|
| A | I | some | Bazı oyuncular sakat. | Some players are injured. |
| B | A | all | Bütün oyuncular sakat. | All players are injured. |
| C | E | none | Hiçbir oyuncu sakat değil. | No player is injured. |
| D | O | not all | Bütün oyuncular sakat değil. | Not all players are injured. |

Afterward, we created the premise-hypothesis pairs and their respective labels. As the relationships between the sentences show, B and C are more universal or stronger than A and D. Whenever B or C is true, then A or D must be true as well, respectively, hence an entailment relationship. On the other hand, B and C are contraries, meaning that when one is true, the other must be false, which creates a contradictory relationship between them. The truthfulness of weaker terms in the square (A and D) do not entail or contradict their stronger counterparts (B and C), thus creating a neutral relationship with them. With these relationships at hand, we are able to create the conventional NLI inference classes.

Finally, implicatures, the main focus of our study, are created by the relationship between the subalterns: A and D. As previously explained, the use of either one of them does not entail or contradict the other; rather, it ‘implies’ the truthfulness of the other. Thus, we named the relationships between A and D, and D

and A implicatures. Table 2 below demonstrates the relationships among each combination of sentences with examples. This list is applicable for adjectives, verbs, quantifiers, and modals, whereas numerals will require a different set of interpretations and labeling, which will be covered later.

Table 2. Inferences between Premise and Hypothesis (For Adjectives, Verbs, Quantifiers, and Modals)

| Premise Type | Hypothesis Type | Premise Example | Hypothesis Example | Inference Type/Label |
|---|---|---|---|---|
| A | D | Bazı oyuncular sakat. | Bütün oyuncular sakat değil. | implicature |
| D | A | Bütün oyuncular sakat değil. | Bazı oyuncular sakat. | implicature |
| C | D | Hiçbir oyuncu sakat değil. | Bütün oyuncular sakat değil. | entailment |
| B | A | Bütün oyuncular sakat. | Bazı oyuncular sakat. | entailment |
| D | C | Bütün oyuncular sakat değil. | Hiçbir oyuncu sakat değil. | neutral |
| A | B | Bazı oyuncular sakat. | Bütün oyuncular sakat. | neutral |
| B | C | Bütün oyuncular sakat. | Hiçbir oyuncu sakat değil. | contradiction |
| C | B | Hiçbir oyuncu sakat değil. | Bütün oyuncular sakat. | contradiction |

For some scalars, the second contradiction pair (C-B) is also an example of metalinguistic negation (Horn, 1985) such as *It is not warm, it is hot!* This sentence includes the use of negation metalinguistically which means that it is not the propositional content that is negated. Therefore, although the pair together forms a quite acceptable sentence, the relationship between them is semantically contradiction.

To confirm the methodology, the first implicature is formed by the utterance of the weaker term in the affirmative. We would expect that the hearer infers the negation of the stronger term. Similarly, when the negation of the stronger term is present in the context, the hearer is expected to infer that the affirmative weaker term is true. Thus, this method results in the expected data items.

Apart from the sentence combinations and resulting labels, we also featured tests for implicature meanings: cancelation, suspension, and reinforcement (Capone, 2006; Potts, 2015). These tests are to be taken as properties of conversational implicatures in general. Therefore, they are used, in the literature as tools to test if an inference is implicature or not. The cancellation test is based on the defeasibility of conversational implicatures. After the implication is inferred from the semantics of the sentence, one can try to cancel the implication meaning. If the implicature meaning can be eliminated from the sentence without any problem, then we can say that the inferred meaning is actually an implicature. We can exemplify the A-D and D-A implicature sentences below:

(19) a. Some players are injured; indeed, all players are injured.

b. Not all players are injured; indeed, no player is injured.

While the utterance of the first part creates the inference that not all players are injured, we are able to cancel that implicature by uttering that the stronger item in the same scale is actually true. Therefore, in the cancelation test, the first part is sentence A while the second part is sentence B for A-D implicature. On the other hand, for the D-A implicature, we utter the D sentence in the first part and C sentence in the second part.

The other test, suspension, works on the same principles as the cancelation test, differing only in the commitment of the speaker to the truthfulness of the stronger sentence. Here, the implicature meaning is not canceled directly but it is suspended as the speaker is not sure about whether the more intense form holds or not. The examples below demonstrate how the suspension tests are applied for conversational implicatures. The sentences to create the test are the same as the cancelation test; only the adverb changes.

(20) a. Some players are injured; maybe all of them are injured.

b. Not all players are injured; maybe none of them are injured.

Finally, the reinforcement test is grounded on the fact that the repetition of the conversational implicature does not create redundancy or any other problem because they are not part of the truth-conditional meaning. For this test, the first part is the A and second part is D for A-D implicatures, and it is D in the first part and A in the second part for D-A implicatures.

(21) a. Some players are injured but not all of them.

b. Not all players are injured but some of them are.

ImplicaTR, in its base version, includes all three tests for each of the sets for every category. These tests were primarily used verify the implicature meaning arising from the sets but they will not be used in following experiments in this study. These test items might offer a great deal of data of quality for future studies. Along with the tests, a complete set is given in Table 3 below along with the tests.

Table 3. Implicature Tests for a Set of Sentences

| Column | Turkish | English |
|---|---|---|
| scale | bazı-bütün | some-all |
| A | Bazı oyuncular sakat. | Some players are injured. |
| B | Bütün oyuncular sakat. | All players are injured. |
| C | Hiçbir oyuncu sakat değil. | No player is injured. |
| D | Bütün oyuncular sakat değil. | Not all players are injured. |
| A-D Cancelation Test | Bazı oyuncular sakat, hatta bütün oyuncular sakat. | Some players are injured; indeed, all players are injured. |
| A-D Suspension Test | Bazı oyuncular sakat, belki de bütün oyuncular sakat. | Some players are injured; maybe all of them are injured. |
| A-D Reinforcement Test | Bazı oyuncular sakat ama bütün oyuncular sakat değil. | Some players are injured but not all of them. |
| D-A Cancelation Test | Bütün oyuncular sakat değil, hatta bazısı bile sakat değil. | Not all players are injured; indeed, no player is injured. |
| D-A Suspension Test | Bütün oyuncular sakat değil, belki de bazısı bile sakat değil. | Not all players are injured; maybe none of them are injured. |
| D-A Reinforcement Test | Bütün oyuncular sakat değil ama bazı oyuncular sakat. | Not all players are injured but some of them are. |

## 3.4 Linguistic categories

The top segmentation within ImplicaTR is the linguistic categories it encompasses: adjectives, verbs, quantifiers, modals, and numerals. In the literature, scalar expressions specifically from these categories have been proposed to lead to an implication inference, among others (Baker et al., 2009; Jeretic et al., 2020, among others). In addition, these categorical differences may lead to differences in inferencing scalar implicatures (Breheny, 2019; Katsos & Cummins, 2010). This is why we have implemented a distinction in the dataset in terms of the parts of speech of particular scales. In this section, we look further into how these categories are treated in ImplicaTR.

The key difference between the linguistic categories are the frequencies they have in a given corpus. The open-class categories, adjectives and verbs, would be expected to have a large number of individual terms in the corpora while their frequencies are relatively low. For quantifiers and modals, it should be the case that the number of lexical items from these categories are small while they have a higher frequency in the corpora. Numerals, similarly of a closed-class, theoretically have an unlimited number of lexical items while it should be noted that a smaller class of them must be in use practically. Studies on word frequency in corpora reveal that this is the case for numerals, where the most frequent ones are those smaller than a hundred or under a thousand (Leech et al., 2002). On the other hand, as expected, Davies and Gardner (2010) show that verbs and adjectives have higher number of words (992 and 893, respectively) in the top words list while numbers and determiners have a smaller size (36 and 34, respectively). This is accompanied by the fact the use of individual lexical items from these closed-class categories comes in

the first places. This shows that type frequency is low for closed class categories while token frequency is higher.

Before moving forward with particular categories, a preliminary look into what would be the properties of scales in each one of them would be preferable. Among the categories we have, adjectives and verbs are open-class categories, meaning that they have relatively large numbers of lexical items compared to closed-class categories. This high number of lexical alternatives might first offer many candidates; however, we were in a prudent mindset to make sure that the proposed words are really in a scalar relationship for the most part. For instance, the intensity scales used in De Melo and Bansal (2013) are too lenient in the sense that most of the scalar items there need to be used in very particularized contexts and meanings to reveal the scalar meaning. For instance, the scale <*available, accessible, visible*> gives the impression that the intensity increases with every step in the scale; however, *visible* should be used in a very precise way to establish an entailment relationship with *available*. *The internet settings are visible in the menu* might entail that the settings are *available* (let us rule out the very possible cases such as if a setting is visible, then it can be grayed out, hence non-available). In terms of implicatures, it is even more challenging because of the fact *availability* does not really implicate *invisibility*. Therefore, these lexemes do not bear the scalar relationship within the proposed scales apart from some peculiar cases. Contrarily, the scales in ImplicaTR were devised to display the scalar relationships in as many different contexts as possible.

It should also be noted that each scale is composed of two items <*weaker term, stronger term*>, which would normally conclude that each linguistic category

should have $2*n_{scale}$ lexical items, which is not always the case. This is because some categories use one lexical item in two different scales, as will be shown.

Another important point is the case of partitives. They share the same set of scales with numerals, although they are classified as a distinct linguistic category in the current study. They differ from regular numeral constructions in that they do not create the implicature inferences in similar conditions as numerals. Additionally, regarding numerals, there is also the case of *C implicatures*, whose premise sentences are the C sentences, hence the name C implicatures. These were created in limited numbers with the aim to test different types of implicatures for the numeral case. We will elaborate on the case of partitives and C implicatures under the numeral section below.

Considering these caveats for each of the categories, the statistics regarding the scales are given in Table 4 below. The number of scales for open-class adjectives is the highest, while the other open-class verbs are similar in the number of scales. However, as seen in the number of distinct terms column, these two scales have the highest number of distinctive terms. Numerals and partitives have the same number of distinct terms as verbs, but it should be noted that numerals are theoretically unlimited in number of potential terms. Quantifiers and modals are the other closed-class categories, and they have the least number of distinctive terms.

In the number of sets per scale column, values denote the number of unique sets for each scale. For instance, <*vasat, iyi*>, an adjectival scale, has 30 different sets of sentences, each of which has the A, B, C, and D sentences. On the other hand, the quantifier set <*bazı, bütün*> has 50 scales, each of which has 50 sets of sentences, again with the sentences A, B, C, and D.

As the total set column shows, our aim was to have the same number of sets for each category, which we attained by producing 450 sets of sentences for each category. As a test category within numerals, partitives differ from the pattern, which is not a problem as it will not be used in the training part of the experiments. For reference, a set consists of the scale name, the A, B, C, and D sentences, and the A-D and D-A implicature tests as shown in Table 3.

On the other hand, the total row column shows the number of rows per category in total. As we mentioned earlier, a set yields 8 rows of data for adjectives, verbs, quantifiers, and modals (see Table 2). Each row is the combination of A, B, C, and D sentences within the set, along with the label of the premise-hypothesis pair. For numerals and partitives, we have 11 rows per set rather than 8, which is caused by the extra C implicatures we have, the extra three implicatures whose premises are the sentence C. Total row number of numerals data is consequently higher than the other scales.

Table 4. Scale, Set, Row Statistics per Linguistic Categories

| | N of scales | N of distinct terms | N of sets per scale | Total Set | Total Row |
|---|---|---|---|---|---|
| Adjectives | 15 | 28 | 30 | 450 | 3600 |
| Verbs | 9 | 18 | 50 | 450 | 3600 |
| Quantifiers | 9 | 7 | 50 | 450 | 3600 |
| Modals | 2 | 4 | 225 | 450 | 3600 |
| Numerals | 9 | 18 | 50 | 450 | 4950 |
| Partitives | 9 | 18 | 10 | 90 | 990 |
| Total | | | | 2340 | 20340 |

### 3.4.1 Adjectives

Adjectival scalars seem to be one of the categories that are extensively researched in the literature. They are studied not only in the scope of implicatures or scalar implicatures but also (maybe even more) in the scope of scalar and degree

relationships between different adjectival terms (Kennedy, 1999; Kennedy & McNally, 2005; Rotstein & Winter, 2004). It is argued that gradable adjectives denote measure functions that map individuals on the degrees of scales. For instance, the function below is the denotation of the adjective tall (taken from Breheny (2019)):

(22) $[[tall_{\langle e,d\rangle}]] = \lambda x.\mu_{HEIGHT}(x)$

Accordingly, adjectives do not differ in the function they map; rather, what is different is whether the scales created by the adjective have a maximal element or a minimal element, or both. The adjective *full*, for instance, forms the maximal point of the scale <empty, full>, while *empty* is the minimal point. Such scales as this one, which possess both minimal and maximal elements, are called closed scales. On the other hand, adjectives such as *tall* or *short* are neither maximal nor minimal elements. One test for this would be whether the adjective can be modified by *slightly* or *perfectly* (Breheny, 2019). While *perfectly full* is an acceptable modification, *#perfectly tall* is not an acceptable construction. The scales composed of non-minimal and non-maximal elements such as <short, tall> are called open scales. Combining minimal and maximal elements, we obtain lower-closed scales such as <bent, straight> and upper-closed scales such as <safe, dangerous> (Kennedy & McNally, 2005). In ImplicaTR, we include all four types of scales to prevent any bias against the adjective types, if any.

For adjectives, we have gathered 15 sets of Turkish adjectival scalars, each of which consists of a weaker and a stronger term. Our aim for the adjectival dataset is that the pairs are closely related to each other, and they somewhat easily lead to implicature inference. As Kennedy and McNally (2005) put forward, scales have three important aspects: degree, dimension, and ordering relation. While degree is the intensity of the adjective compared to the other adjective, dimension can be

considered the meaning line the adjectives are positioned on. The ordering relation determines the ordering between adjectives that are on the same dimension. In studies such as De Melo and Bansal (2013), a significant part of adjectival pairs does not reflect the implicature relationship we want to extract. This is mainly because the dimension that they must share in order to reveal the implicature meaning is so particularized that the contexts that will authenticate the dimension are quite limited. Similar to the given example above, the terms in the scale <*good, real, authentic*> must be used in a very particular way to make the hearer infer that *if x is good, it is not authentic*. With an aim to involve as many contexts as possible, we have come up with a relatively limited number of scalar pairs in ImplicaTR, which will display the scalar meaning in most contexts.

Another point is that the scalar terms are sometimes taken from the opposite ends of a dimension so that the scale is composed of antonyms. However, this creates another problem: if both of the antonyms are at the opposite ends of the dimension, then it is not always the case that the negation of one implicates the truthfulness of the other. This can be seen in the case <*ugly, beautiful*>. When the utterance is *She is not beautiful*, it is not commonly understood that *She is ugly* is true. Rather, there is a mediocre value, such as *fine* or *nice*, which is implicated when the stronger term *beautiful* is negated. This is why we include mediocre terms in the scales to correctly construct the implicature inference.

Table 5 below provides a comprehensive list of the adjectival scales that ImplicaTR has. To see example implicature sentence pairs, please see Appendix A.

Table 5. Adjective Scales in ImplicaTR

| | | |
|---|---|---|
| benzer-aynı | yeterli-tam | zararlı-ölümcül |
| *similar-same* | *sufficient-complete* | *harmful-lethal* |
| vasat-kötü | yakın-bitişik | makul-harikulade |
| *mediocre-poor* | *close-adjacent* | *reasonable-marvelous* |
| uygun fiyatlı-ucuz | talihsiz-vahim | yanıltıcı-yalan |
| *affordable-cheap* | *unfortunate-disastrous* | *misleading-false* |
| samimiyetsiz-düşmanca | iyi-mükemmel | ilgili-takıntılı |
| *insincere-hostile* | *good-excellent* | *interested-obsessed* |
| vasat-iyi | muhtemel-kesin | idare eder-harika |
| *average-good* | *likely-certain* | *acceptable-great* |

### 3.4.2 Verbs

Similar to adjectives, verbs are types of words that can denote degrees in certain dimensions. A scale of verbs can be created with different degrees in the same dimension. Unlike adjectives, verbs are not generally studied in the scope of scalars, so the literature is relatively limited, though verbs have been shown to appear in many different meaning dimensions (Jackendoff, 1996). In ImplicaTR, the verbal scales are chosen among the ones that exhibit the scalar relationships in as many contexts as possible, as we did for adjective scales.

Inceptive verbs, which are among the verb types employed in the dataset, are the verbs that denote the inception and the development of the action. In this sense, they can be naturally considered be establishing the scalar meaning between themselves (Pedersen, 2014). For instance, in the scale <start, finish>, which possesses a ‘weaker-stronger’ relationship in a sense, the utterance of the weaker term implicates in many contexts that the proposition of the second term is not yet true. With these points in mind, we included nine verbal scales in ImplicaTR, each with 50 sets of sentences. Table 6 below lists the scales, while Appendix A includes the scales and example implicature premise-hypothesis pairs.

Table 6. Verb Scales in ImplicaTR

| | | |
|---|---|---|
| yönelmek-varmak | denemek-başarmak | öğrenmek-ustalaşmak |
| *approach-arrive* | *attempt-succeed* | *learn-master* |
| algılamak-kavramak | başlamak-bitirmek | katılmak-benimsemek |
| *perceive-grasp* | *start-finish* | *agree-adopt* |
| seslenmek-bağırmak | gülmek-kahkaha atmak | serinlemek-üşümek |
| *call out-yell* | *laugh-laugh out loud* | *cool-shiver* |

### 3.4.3 Quantifiers

Quantifiers are probably the foremost examples of scalars and scalar implicatures, with <some, all> being the introductory scale into the realm of scalars. This is mainly because quantifiers denote quantified expressions within which there is the inherent presence of degrees of meaning. This is to say that quantifiers such as *some*, *most*, or *all* are immediate displayers of the existence of intensity or scope of some measurement. As required by the Horn scales, the stronger term must entail the weaker term in the scale, which is how the quantificational scales were created. For this, we used the maximal value *all*, which entails anything lower as well as logical determiner *some* and non-logical determiner *most* (Tomioka, 2021).

Turkish is a language that expresses the existential and universal quantifiers with many different words and structures like English. However, these languages differ in how and where they employ the mass or countable quantifiers (Özyıldız, 2017). This is why we included many different quantifiers in Turkish which give the existential meaning of *some*, the universal meaning of *all*, and the non-logical *most*.

The minimal terms included in ImplicaTR are *birkaç* ‘a few’, *biraz* ‘a little’, *bazı* ‘some’ while the maximal terms are *bütün* ‘whole’, *hepsi* ‘every’, and *tüm* ‘all’. There is also *çok*, which means ‘most, many, much’. With the combinations of these quantifiers, we produced 9 different quantificational scales each of which possess 50 sentences. The table below is the list of the scales used while one can refer to

Appendix A to see scales with example implicature sentences. In these tables, English glosses are given the most appropriate version of the quantifier, while the translations of premise-hypothesis pairs use different scalar terms. As mentioned above, this is because quantifiers in both languages have different use cases.

Table 7. Quantifier Scales in ImplicaTR

| | | |
|---|---|---|
| birkaç-bütün | birkaç-tüm | birkaç-hepsi |
| *a few-whole* | *a few-all* | *a few-every* |
| bazı-bütün | bazı-tüm | bazı-hepsi |
| *some-whole* | *some-all* | *some-every* |
| biraz-çok | birkaç-çok | bazı-çok |
| *a little-much* | *a few-most* | *some-most* |

### 3.4.4 Modals

Modality has been an area of research that is investigated by different fields such as linguistics, logic, and philosophy. In semantic research, a general distinction between different modalities is the distinction between *deontic modality* and *epistemic modality* (Kaufmann et al., 2006). The former refers to the cases of obligation, permission, or necessity where the subject is understood to be under a force to differing extents. On the other hand, epistemic modality includes cases where the certainty of the knowledge that the subject has is at stake (De Haan, 2006). In ImplicaTR, we opted for epistemic modality rather than deontic ones for two reasons: the convenient nature of epistemic modals for scalar relationships and the structure of Turkish modalities.

First, a further classification for epistemic modalities is the case of *gradable epistemic modals* (Lassiter, 2010). Influenced by the work by Kennedy and McNally (2005) on gradable adjectives, this type of modals embodies the modals such as *possible*, *likely*, *probable* or *certain*, which obviously denote the different levels of

certainties, hence creating order within themselves. With tests on these modal lexemes, Lassiter shows that *possible* shows the characteristics of a minimum-standard adjective while *certain* is surely a maximal element, which is significant in that these modal types will not lead to any ambiguous use cases.

Second, Turkish is a language that expresses modality with several structures such as lexemes, adjectives, adverbs, or inflectional suffixes (Kerimoğlu, 2010). These modality constructions sometimes have overlapping uses, which creates ambiguity. For instance, in Turkish, as well as in many other Turkic languages, we encounter the use of the ability marker *-Abil* in for epistemic modality (Johanson, 2009). In (23a) below, this marker is used in the epistemic meaning, which means that it is possible that Elif will sing a song tomorrow. In (23b), however, the same marker is in deontic meaning, which refers to the fact that Elif has the ability to sing a song.

(23) a. *Elif yarın bir şarkı söyle-yebil-ir.*
Elif tomorrow a song sing-PSB-AOR.
'Elif might sing a/one song tomorrow.'

b. Elif yarın bir şarkı söyle-yebil-ir.
Elif tomorrow a song sing-ABILITY-AOR.
'Elif can sing a/one song tomorrow.'

If such a marker were placed in a scale as a minimal element along with a maximal element like *kesin* 'certain', there would be cases where the deontic meaning interferes with the entailment and implicature relationship between the scale items, thus hindering scalar inference. Consequently, on top of our desire to abstain from such ambiguities, the fact that epistemic modalities are good examples of gradable meanings led us to opt for epistemic modality instead of deontic in the modal scales in ImplicaTR.

A caveat for the modal scales is that we also employed a structured selection of certain tenses and aspects in the dataset. This is to say that some scales are constructed with certain tenses and aspects while we used another tense and aspect for other scales. With this manipulation, it might be possible to see if tense and aspect have any influence on the scalar inference. Table 8 below shows the specific tense and aspects with example sentences.

Table 8. Tenses and Aspects in Modal Scales

| Turkish Sentence | English Equivalent | Tense/Aspect |
|---|---|---|
| Muhtemelen yapacak. | S/he will probably do it. | Future |
| Muhtemelen yapacakmış. | (I heard that) s/he will probably do it. | Presumptive Future |
| Muhtemelen yapacaktır. | S/he will probably do it. | Emphatic Future |
| Muhtemelen yapar. | S/he will probably do it. | Aorist |
| Muhtemelen yaparmış. | (I heard that) s/he will probably do it. | Presumptive Aorist |
| Muhtemelen yaptı. | S/he probably did it. | Past |
| Muhtemelen yapmış. | (I heard that) s/he probably did it. | Presumptive Past |
| Muhtemelen yapmıştır. | S/he probably did it. | Emphatic Presumptive Past |
| Muhtemelen yapıyor. | S/he is probably doing it. | Continuative |

These different constructions might arguably create differences in the modal nuances, as they differ in specific aspects that they denote (Sebüktekin, 1971). Among these, emphatic can be considered an extra layer of possibility since the emphatic -DIr is a marking for different degrees of epistemic certainty (Arslan, 2020).

Lastly, we have used both adverbial and adjectival modals for different sentences in the very same set to prevent the negation from denoting the other extreme end of meaning. To illustrate, *Kesinlikle evden çıkarım* (I will surely leave the house) denotes the stronger case, while *Kesinlikle çıkmam* (I will surely not leave) denotes another stronger case which is not a part of our scale at hand. To prevent this, we use *Evden çıkacağım kesin değil* (It is not certain that I will leave the house), where the modality is expressed by an adjective.

As a result, ImplicaTR has two different scales for modals, each of which has 225 sets of sentences. Within each scale, the modal sentences are created with different tenses and aspects to further investigate the effect of tense and aspect on the inference of scalar implicatures, as shown above. The list below gives the modal scales while Appendix A has example implicature meanings.

Table 9. Modal Scales in ImplicaTR

| | |
|---|---|
| muhtemelen-kesin | herhalde-yüzde yüz |
| *probably-certain* | *possibly-one hundred percent (certain)* |

### 3.4.5 Numerals and partitives

The final linguistic category in ImplicaTR is numerals and partitives. Partitive constructions containing numerals differ from ordinary numerals in how they give rise to implicature inferences, and they will be examined at the end of this section.

Numerals are fundamentally different from other categories in that they inherently possess the scalar meaning. This is because they automatically give the 'ordered relation' meaning as they are all actually placed on the number line. Therefore, the number line becomes the dimension for numeral scales. In the literature, however, numerals are subject to debates about whether they possess an 'exactly' meaning or an 'at least' meaning in their semantics or whether they are ambiguous.

Those who advocate that numerals have the 'at least' meaning in their semantics (Levinson, 2000) propose that the 'exactly' meaning is computed in pragmatics. Thus, they support that numerals have a lower-bounded meaning, whose implicatures are the 'exactly' meaning (for upper-bounded meaning and ambiguity arguments, see Panizza et al. (2009)). Combining Gricean reasoning with lower-

bounded numeral arguments, we can infer that (24a) below means that the number of her cars is not more than three. However, as implicatures can be reinforced, we can reinforce this implicature as in (24b).

(24) a. She has three cars.

b. She has three cars, but not four/and she wants to buy a fourth one.

Related to this, there are *upward entailing* and *downward entailing contexts*. In upward entailing contexts, the superset inferences are entailed from the subsets. The examples in (24) are upward entailing contexts, where a sentence with superset, such as *She has three vehicles* is entailed by (24a) or (24b). This entailment pattern can be inverted with downward entailing contexts such as conditional antecedents (Panizza et al., 2009). The sentence below exemplifies one such context:

(25) If she has three cars, she won't be able to pay her rent.

Here, the speaker infers that the conditional statement should be true whenever the number of cars she has is *at least* three rather than a somewhat strange inference that it will not be the case if she has four cars, five cars, etc. As Panizza and colleagues put forward, the 'exactly' inference is generally preferred in upward entailing contexts while the 'at least' inference is preferred in downward entailing ones. In this study, with a Neo-Gricean perspective, we take numerals as referring to the 'at least 'meaning in their semantics and as giving rise to the 'exactly' implicature by means of pragmatics. Therefore, in ImplicaTR, numerals are always in upward entailing contexts where 'exactly' will be an implicature, and there is no downward entailing context in data items.

Apart from the semantic-pragmatic meanings of numerals, the scalar nature of numerals is argued in the literature as well. Papafragou and Musolino (2003) argue that numerals can give rise to 'at most' implicature, unlike other scalars. For instance,

a sentence like *This car can go 500 miles without refueling* generates the implicature that the car can run ‘at most’ 500 miles. On the other hand, the Gricean view that numerals are mostly in ‘at least’ meaning is challenged by the fact that numerals are actually used with an ‘exact’ meaning. These are also significant in displaying the distinct nature of numerals compared to others.

In ImplicaTR, the first peculiarity of numerals is the construction of the second implicature. What is significant regarding numerals compared to other scalars is that, for a scale <three, five>, the negation of the stronger value does not directly implicate the affirmative of the weaker value. Instead, it refers to the inference that any value below it can be true. This directed us towards implementing a small change in implicature sentences for numeral scales. In other linguistic categories, A-D implicature is the affirmative weaker term and negated stronger term (Some friends danced +> Not all friends danced). And the second type of implicature, D-A implicature is just swapping the subalterns (Not all friends danced – Some friends danced). In numerals, however, the premise of the second type of implicature is changed to a proposition that denotes ‘more than what the weaker value denotes does not exist (is not applicable, etc.)’. Sentences below demonstrate example numeral implicatures in the dataset:

(26) a. A-D: *Üç kişi dans etti +> Beş kişi dans etmedi.*
Three people danced +> Five people did not dance.

b. $D_{numeral}$-A: *Üç kişiden fazlası dans etmedi +> Üç kişi dans etti.*
No more than three people danced +> Three people danced.

This way, we ensured that the second implicature correctly established the implicature meaning.

The second point regarding numerals is three more implicature premise-hypothesis pairs for each set. We called these three extra implicatures *C implicatures*

because of the fact that the premise sentence of these implicatures is the C sentences in each set. To recall, C sentences are created by the negation of the weaker term in a scale like *I have not started my homework*. Such sentences denote that even the lesser/weaker value in a scale is not true. Sentences below exemplify each C implicatures for the scale <three, five>:

(27) a. Premise (Sentence C): She does not have 3 cars.

b. C implicature 1: She has less than 3 cars.

c. C implicature 2: She has at most 2 cars.

d. C implicature 3: She has at least 1 car.

C implicature 1 is quite similar to how the ordinary D-A implicatures are constructed. The C implicature 2 exhibits the aspects of what Papafragou and Musolino (2003) find strange in conventional scalar implicature perspective by (Neo)-Griceans. This implicature has the 'at most' implicature. On the other hand, the C implicature 3 is an example of the existential implicature within the scope of generalized conversational implicatures but not within scalars. We consider that these extra implicatures might reveal important outcomes regarding the nature of numeral scalars.

For numerals, 9 different sets were created, each of which is scale of pairs of two integers smaller than 100. We included numbers that are close to each other as well as number pairs which are not commonly found together. Table 10 below gives the comprehensive list of the scales.

Table 10. Numeral and Partitive Scales in ImplicaTR

| | | |
|---|---|---|
| iki-üç | üç-beş | on-on iki |
| *two-three* | *three-five* | *ten-twelve* |
| on beş-yirmi | on yedi-yirmi dokuz | yirmi dört-otuz altı |
| *fifteen-twenty* | *seventeen-twenty-nine* | *twenty-four-thirty-six* |
| yirmi beş-kırk | otuz-altmış | elli-yetmiş |
| *twenty-five-forty* | *thirty-sixty* | *fifty-seventy* |

Another important case for numerals is the case of partitives. Partitives are constructions that express the partial or incomplete quantity of something, for which an example would be *three cars* vs. *three of the cars*. The entailment and implicature relationships that we have discussed for the numerals so far curiously do not hold for the structures where the numerals are inside partitive structures. One explanation for this is the fact that the partitives are specific NPs (Enç, 1991). Specificity refers to whether the noun has a specific referent or not in the real world although the referent does not have to exist. For an implicature inference, the scalar items must be in existential sentences where the negation of the stronger term *all* or *three* should implicate the existential weaker terms *some* and *two*. However, when the scalar items are in specific NPs such as *some of the books* or *two of the books*, where the quantity and quality of the books are already known or not important in the context, the scalar dimension can be said to collapse in the sense that there is no *all* or *some of the books*, but there is *the specific books*. Enç (1991) states that partitive constructions generate complex determiners, and complex determiners such as the ones in the sentences below are not acceptable in existential sentences:

(28) a. *There are some of the cars.

b. *There are two of the cars.

As partitivity leads to specificity and specificity removes ordinary entailment and implicature relationships, partitives are included as a separate category within numerals with the same scalar items.

As a result of the discussion regarding numerals and partitives, these categories have different labels in ImplicaTR. In Appendix A, the labels for both categories are listed with example sentences.

# CHAPTER 4

# EXPERIMENT 1: TRAINING LLMS ON IMPLICATR

To investigate the pragmatic abilities of LLMs, we conducted a series of experiments using the ImplicaTR dataset that we have developed. Many linguistic studies demonstrated that there is pragmatic reasoning in the derivation of scalar implicatures (De Neys & Schaeken, 2007; Huang & Snedeker, 2009; Katsos, 2005). In NLI research, language models have been shown to reason over pragmatic processes happening implicitly in the background in the human parser (Kim et al., 2023; Stacey et al., 2022). Based on these findings, the aim of Experiment 1 is to investigate if LLMs have pragmatic abilities and resolve scalar implicatures. For the models to train, we chose three BERT models: the base BERT (Devlin et al., 2019), BERT-NLI (Laurer et al., 2023), and BERTurk (Schweter, 2020), which will be elaborated on in the upcoming sections. Upon training the BERT models, we further experimented with three decoder-only LLMs: Llama-2 by Meta (Touvron et al., 2023), Gemma by Google (Gemma Team et al., 2024), and Mistral (Jiang et al., 2023).

Before discussing the training details and results, let us look into how the data is split for our purposes.

## 4.1 Splitting the dataset

As the dataset is semi-automatically created by producing the sets and extracting the sentences out of each set, it is necessary to make sure the data is properly split so that the label distribution is balanced and any set (with is quadruplet of sentences) is included in one and only one split. This is to say that, for instance, a single

quadruplet from the 50 <*birkaç, tüm*> quadruplets in total within quantifiers gives way to eight inferences (two for each entailment, neutral, contradiction, and implicature case). We aimed to involve all these eight inferences from one quadruplet in either train, test, or validation to prevent the model from seeing the same sentence in train or test. Thus, we used stratified sampling by first splitting the quadruplet sets into different splits and then populating each split with the eight sentences from the set. As a result, train, test, and validation splits were created with those numbers expressed in Table 11 below.

Table 11. Train, Validation, and Test Datasets for Models in Experiment 1

| | N of Scalar Sets | N of Rows | % |
|---|---|---|---|
| Train | 1440 | 12309 | 64% |
| Validation | 360 | 3153 | 16% |
| Test | 450 | 3888 | 20% |
| Total | 2250 | 19350 | 100% |

## 4.2 BERT models

BERT is an encoder-decoder model developed by the Google Team (Devlin et al., 2019) based on the transformers architecture (Vaswani et al., 2017). With their bidirectional architecture, BERT-family models take into account the left and the right context of a masked element within a sentence, which increases their ability to grasp the meaning of a sentence further. This is why they are highly employed in academic settings by NLP researchers to work on the linguistic understanding of NLP models with their contextualized representations (Cho et al., 2021). Though they are not preferred for next-word prediction tasks, they offer remarkable next-sentence prediction abilities, which is why they are utilized in sentence-level NLP tasks like NLI. They are also shown to demonstrate notable crosslinguistic

knowledge with the help of cross-lingual word embeddings, which makes them good candidates for multilingual studies (Ruder et al., 2019; Wu & Dredze, 2019).

BERT has given rise to a large number of similar models where the architecture is slightly or reasonably changed, or the model is predominantly finetuned for a specific task. One model of this kind is BERT-NLI (Laurer et al., 2023), where the base DeBERTaV3 (He et al., 2023) model is finetuned on XNLI (Conneau et al., 2018) and MNLI (Williams et al., 2018) datasets. This model is a multi-lingual zero-shot classifier where the authors transformed any classification task into an NLI task (Yin et al., 2019). For instance, to classify whether a piece of news is in the category of "sports" or "politics", they create premise-hypothesis pairs where the premise is the document and the hypothesis is some sentence like "This is a sports-related text", which go into a usual NLI pipeline which is a three-way classification task between entailment, neutral, and contradiction. Based on the classes chosen by the model and the scores, the model finally chooses a label for the document, hence completing the task. Finally, BERTurk (Schweter, 2020) is a model that was trained on different Turkish datasets such as Wikipedia dumps. This is a family of models from different BERT architectures with high accuracy scores for Turkish tasks. Among the different versions, we trained the uncased version with a 32k vocabulary size.

### 4.2.1 Preliminary tests

Before training the models, we first conducted a test on how these models perform on our test dataset. It is important to note that BERT-NLI has a classifier head with 3 classes, entailment, neutral, and contradiction, while the two others have not been finetuned on a downstream task. This is why the classifier heads of all three models

are reset when we add another class, implicature. Now that they have a classifier head with 4 classes, all the weights in this head are newly initialized. Accordingly, we expect random-like accuracies from the models. Table 12 below shows the accuracy scores of the models, which are in line with our expectations.

Table 12. Accuracy Scores of BERT Models on Test Dataset before Finetuning

| | BERT-Base | BERT-NLI | BERTurk |
|---|---|---|---|
| Accuracy | 0.23148148 | 0.23636831 | 0,30632716 |

### 4.2.2 Training details

After we obtained the train, validation, and test datasets and made a preliminary test on the models, we continued with the training. For training, we have devised several hyperparameter-finetuning stages to increase the accuracy on the validation dataset and decrease the training and validation loss. To deal with any overfitting issues, dropout regularization techniques were applied to the models during training (Lim, 2021). The dropout probabilities of hidden layers were set to 0.3 while the dropout of attention probabilities was set to 0.25 for BERT-Base and BERT-NLI, whereas 0.05 lower ones for both values were applied for the BERTurk model. Instead of smaller numbers, these relatively larger values were selected because the task whose dataset was systematically curated might be somewhat easy for such complex models. As it is seen, no overfitting or underfitting issues arose during training.

Adam optimizer (Kingma & Ba, 2017) from PyTorch library (Ansel et al., 2024) was used along with a learning rate of 0.00001 as a starting point. We also used a learning rate scheduler to prevent stagnation in the loss value. Again, from the PyTorch library, we implemented the type of scheduler where the learning rate is reduced when a parameter is on a plateau and stops improving. The factor for

learning rate reduction is 0.5, which is applied when there is no change in the loss value larger than 0.2. Other hyperparameters used in the training procedure for all of the phases were batch sizes for both train and validation as 64. Every model was trained for 10 epochs. The comprehensive list of hyperparameters used during the training of BERT models is given in Table 13 below.

Table 13. Training Hyperparameters for BERT Models

| Hyperparameter | Value |
|---|---|
| hidden dropout value | 0.3 |
| attention dropout prob | 0.25 |
| number of epochs | 10 |
| gradient accumulation steps | 2 |
| warmup ratio | 0.01 |
| batch size | 64 |
| weight decay | 0.05 |
| learning rate | 0.00001 |
| lr reduction factor | 0.5 |
| lr reduction threshold | 0.2 |

While training the models, we did not detect any overfitting or underfitting. Train and validation loss are given in the figures below.

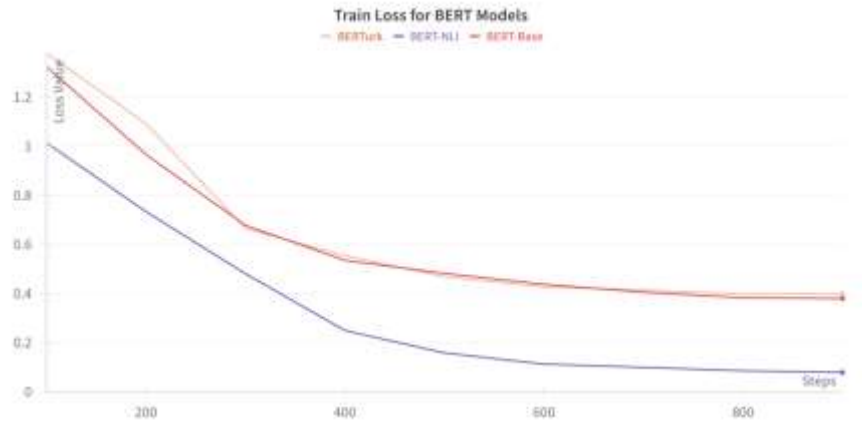



Figure 2. BERT models' train loss


Figure 3. BERT models' validation loss

The accuracy scores increase for each model which shows that the models learn the patterns between the classes, which is visualized in Figure 4 below.

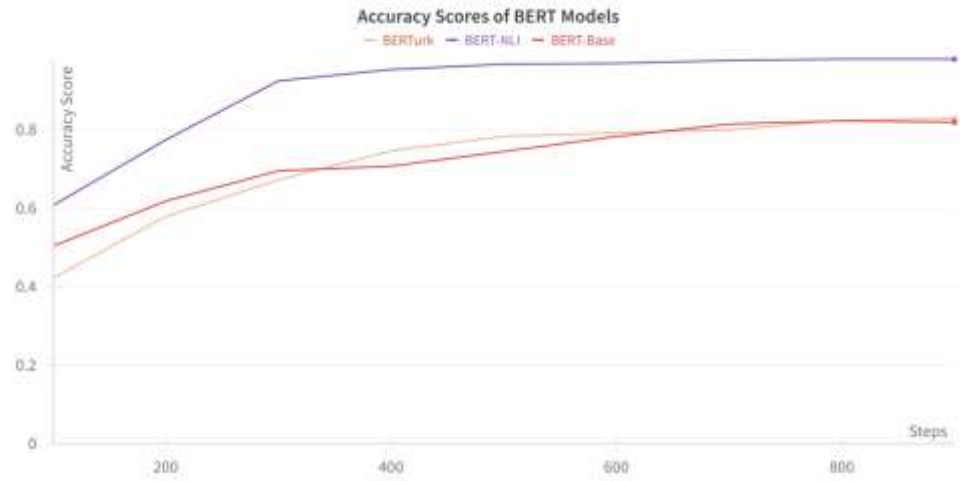



Figure 4. BERT models' accuracy scores

For reference, loss and other metrics are given in the Table 14 below. The fine-tuned BERT-NLI seems to be the best model among the three, with an accuracy score of 0.98 on the validation dataset. On the other hand, BERT-Base and BERTurk demonstrate a good performance though being relatively worse than BERT-NLI. It is also interesting that the base model and the BERTurk model do not show a great

difference; BERTurk has a higher validation accuracy just by 0.01. Additionally, the fact that precision, recall, and F1 scores are close to each other indicates the success of the learning process.

Table 14. Training Results of BERT Models

| | BERT-Base | BERT-NLI | BERTurk |
|---|---|---|---|
| Training Loss | 0.379500 | 0.078900 | 0.397300 |
| Validation Loss | 0.355760 | 0.085844 | 0.335098 |
| Accuracy | 0.818903 | 0.979385 | 0.827149 |
| F1 Score | 0.815816 | 0.979294 | 0.825047 |
| Precision | 0.821169 | 0.979569 | 0.830822 |
| Recall | 0.818903 | 0.979385 | 0.827149 |

### 4.2.3 Test results

After training the models, they were tested on the test dataset to see how successful they were and detect if there was any discrepancy between the scores. The test results are given in Figure 5 below.

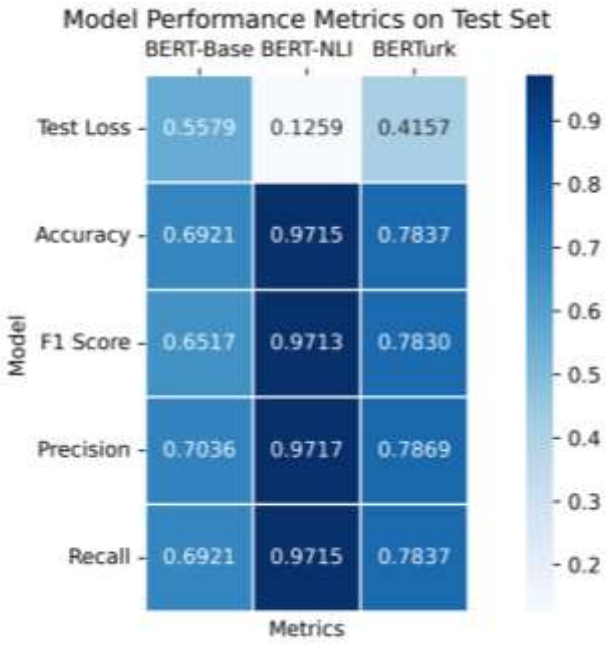



Figure 5. Model performance metrics on test set

Accordingly, the best model is BERT-NLI, with an accuracy score of 0.97. It is followed by the BERTurk model with 0.78 and by the base BERT with a score of 0.69. One thing to notice here is the difference between the validation accuracy and test accuracy of the BERT-Base and the BERTurk models, both of which declined in performance by 0.11 and 0.04, respectively. This indicates an overfitting between the validation and test. In their study, Hao et al. (2019) state that the BERT model is over-parametrized (345M parameters) for the datasets, which will likely lead to overfitting. However, the pre-training process before fine-tuning with the target dataset decreases the likelihood of overfitting or at least minimizes it, which is what we see in our case.

Other than that, BERTurk, which was trained on extra corpus such as OSCAR (Ortiz Suárez et al., 2019) or OPUS (Tiedemann, 2012) corpora, was able to retain a higher accuracy, which shows that this model does a better reasoning than the base model and it learns the patterns within the data to a better extent.

BERT-NLI model was trained on several NLI datasets, which makes it specialized in recognizing the NLI patterns within the data. The fact that we add a new label into the pipeline does not seem to deteriorate its performance, at least on our own data. Indeed, our dataset is a synthetically produced dataset where the labels generated are based on an inner structure. Thukral et al. (2021) show in their study, where they create a synthetic dataset to challenge the temporal reasoning of NLI models, that BERT or similar LLMs are prosperous in detecting label creation procedures and are likely to show near-perfect results in these cases, which is what we observe in our tests with BERT-NLI.

### 4.2.4 Benchmark on XNLI and MNLI

After training the models, we continued evaluating how successful our training was. For this aim, we wanted to do benchmarking on XNLI and MNLI with the finetuned BERT-NLI model. The reason we chose this model is that the authors tested the original BERT-NLI model on the test datasets of these benchmarks and shared the scores (Laurer et al., 2023).

The original model was a three-way classifier with entailment, neutral, and contradiction labels. With finetuning, we increased the class size to four by adding the implicature class, thereby adjusting the model weights and reinitializing the classifier head. In addition, ImplicaTR has ~20k rows of data, which is small compared to the model size. This is why we expected a decrease in model accuracy on XNLI and MNLI; nonetheless, it would be a more granular inferencing among the four classes.

We employed four different metric evaluation methods. The reason for this is that XNLI and MNLI are three-class datasets while our model is a four-class classifier, and it might not be representative as much as possible to evaluate it on three classes. Therefore, firstly, we accepted the implicature predictions as implicatures and computed the metrics as is. As a second method, we converted the implicature predictions of the finetuned BERT-NLI model into entailment and computed the metric scores. In the third step, implicature predictions were converted into neutral, and in the last step, similarly, the implicature predictions were converted into contradiction labels before computing the metrics. This way, we will see which class is predicted as implicatures the most. The results of the XNLI benchmark are given in Table 15 below with all four methods and also with the original BERT-NLI score.

As seen, our model performed worse than the original BERT-NLI model in all cases, with 61% accuracy being the least accuracy among all, which is for the implicature case where the model predictions are not converted to any other label. For the cases where the implicature predictions are converted into neutral and

Table 15. Accuracies of Finetuned BERT-NLI with Different Labeling on XNLI, and Original BERT-NLI Score
Implicature: when the predicted implicature labels are evaluated as they are, Entailment: after converting them into entailment, Neutral: after converting them into entailment, Contradiction: after converting them into contradiction

| XNLI dataset | implicature | entailment | neutral | contradiction | BERT-NLI |
|---|---|---|---|---|---|
| ar | 0.60479 | 0.623952 | 0.670459 | 0.676048 | 0.794 |
| bg | 0.649102 | 0.664671 | 0.71477 | 0.711377 | 0.822 |
| de | 0.632535 | 0.648703 | 0.702395 | 0.7 | 0.824 |
| el | 0.632335 | 0.647106 | 0.696208 | 0.697804 | 0.809 |
| en | 0.672655 | 0.68483 | 0.746307 | 0.740519 | 0.871 |
| es | 0.652894 | 0.669661 | 0.724351 | 0.720958 | 0.832 |
| fr | 0.640719 | 0.656287 | 0.71018 | 0.708583 | 0.823 |
| hi | 0.586028 | 0.60978 | 0.6499 | 0.657884 | 0.769 |
| ru | 0.626148 | 0.640519 | 0.696806 | 0.693214 | 0.803 |
| sw | 0.574451 | 0.596407 | 0.636926 | 0.644511 | 0.746 |
| th | 0.591816 | 0.613573 | 0.660679 | 0.658483 | 0.786 |
| tr | 0.579641 | 0.602994 | 0.654092 | 0.658283 | 0.792 |
| ur | 0.553693 | 0.576846 | 0.612974 | 0.626547 | 0.744 |
| vi | 0.615569 | 0.632335 | 0.678044 | 0.686427 | 0.793 |
| zh | 0.618164 | 0.633932 | 0.682236 | 0.684631 | 0.803 |
| Average | 0.61536933 | 0.63343973 | 0.6824218 | 0.68435127 | 0.80073333 |

contradiction, we see that the accuracy score goes up to 68% while the entailment case is 63%. Consequently, taking the contradiction case into account, our model demonstrated a performance drop of 12%, which we think is not bad considering that our model's analysis is more granular than the original three-way classifier model. The interesting case is why neutral and contradiction cases have higher accuracies than the entailment case. Considering that these models are positively biased towards the contradiction and neutral class because of the existence of negations words like 'not' and superlative words such as those denoting the maximal values (Gururangan et al., 2018).

On the other hand, for the MNLI benchmark, we have both matched and mismatched test sets. The table below again shows our model's performance on these sets, along with the original BERT-NLI model's scores.

Table 16. Accuracies of Finetuned BERT-NLI with Different Labeling on MNLI, and Original BERT-NLI Score
Implicature: when the predicted implicature labels are evaluated as they are, Entailment: after converting them into entailment, Neutral: after converting them into entailment, Contradiction: after converting them into contradiction

| MNLI Dataset | implicature | entailment | neutral | contradiction | BERT-NLI |
|---|---|---|---|---|---|
| Matched | 0.654712 | 0.665206 | 0.72216 | 0.71839 | 0.857 |
| Mismatched | 0.663344 | 0.670566 | 0.724776 | 0.731082 | 0.856 |

The results are in line with the XNLI results above: implicature cases yielded the least successful performance, while contradiction and neutral cases improved the performance. The contradiction case is similarly 12% less accurate. We again see the effects of hypothesis creation procedures of these large datasets, where the crowdworkers used similar heuristics to create these hypotheses that favor neutral and contradiction cases. It is plausible to confer again that the more granular analysis of the model performed the performance in all cases but it is still at good values with 73%.

## 4.3 Generative models

In our experiment, the dataset was also employed to train some generative models, which are decoder-only models trained with an autoregressive modeling objective. This way, we can see how they perform with such an NLI task while also having the opportunity to make a basic comparison between the models. In this sense, we will experiment with Llama-2, Gemma, and Mistral.

Unlike BERTs, which use encoder-decoder architecture, these models rely only on the decoder component. They are trained on seq2seq tasks, where they take the input sequence and generate an output sequence (Fu et al., 2023). Thus, these models learn and generate output by performing next-word prediction. While their performance is argued to be worse (Fu et al., 2023) in understanding the meaning of language compared to encoder-decoder models, they have been used in many different NLP tasks such as translation, generation, summarization, or even classification (Raffel et al., 2023). In the next sections, data processing, preliminary tests, and training of the models will be presented.

### 4.3.1 Data processing

The models we experiment on in this section are text generation models, which make next-word prediction. The models use the tokens in the input sequence to predict the next word that comes after the sequence and continue this behavior until a desired piece of text is produced. Thus, the data we have, which is a text classification dataset, needs to be properly transformed into inputs that these models accept: prompts. We created prompts that guide the model about what it should do, NLI classification. Within each prompt, we also included the premise-hypothesis pair along with the ground label. A prompt as in (29) below was created:

(29)

Below is an instruction that describes a classification task. Give a label in your response that appropriately completes the request.
You will give only the label.
### Instruction:
The labels are:
**Labels:** entailment, neutral, contradiction, implicature
The two sentences that you will classify are:
**Sentences:** A: *Yeni kullanmaya başladığı ilaçlar zararlı değil.* B: *Yeni kullanmaya başladığı ilaçlar ölümcül.*
**Question:** What is the correct label that describes the relationship of B to A?
### Response:
*contradiction*

The parts in italic text are the parts that change with each prompt created. In our prompt testing process, we see that short but precise prompts tend to give better results, so the final prompt, as given in (29), is the best one among the ones we tested. What we spotted is that longer and detailed prompts (such as giving the labels twice, explaining the subject of implicatures) distracted the models and they gave random responses like 'You have 20 minutes to complete the questions…'.

The test dataset prompts are the same as (29); only the label is not given to the model this time, hence the prompt below.

(30) (The prompt above in (29))

### Response:

As we did not use any validation set in the training phases of these models, the data was split into two sets: test set and train set, which is a combination of the train and validation sets in the previous section. The respective number of sets and rows is given in Table 17 below. With these train and test prompts, we first test the models without any finetuning and see how they perform, before training the model.

Table 17. Train, Validation, and Test Datasets for Generative Models

| | N of Scales | N of Rows | % |
|---|---|---|---|
| Train | 1800 | 15462 | 80% |
| Test | 450 | 3888 | 20% |
| Total | 2250 | 19350 | 100% |

### 4.3.2 Preliminary tests

With the prompts generated, the first phase was testing the models without any finetuning or extra prompting. The three models were run to complete the test prompts. One apparent problem is the occasional inability to get the desired output from the model. Their responses varied from giving unrelated responses to

paraphrasing the input prompt. In those cases, we rerun the prediction for the problematic prompt to get a response eventually. One tendency we observe is their tendency to explain why they classify the sentences as they do, although it is a wrong classification and explanation. We extracted the labels from of their completions and calculated the metrics against the ground labels. As expected, accuracy scores indicate the randomness of the responses, while F1 scores are worse than accuracy. The scores obtained are given below.

Table 18. Accuracy Scores of Generative Models Before Training

| | Llama-2 | Gemma | Mistral |
|---|---|---|---|
| Accuracy | 0.240802 | 0.267558 | 0.304347 |
| F1 Score | 0.160455 | 0.189167 | 0.202525 |
| Precision | 0.220722 | 0.253031 | 0.183037 |
| Recall | 0.240802 | 0.267558 | 0.304347 |

### 4.3.3 Training details

To train the generative models, we used the HuggingFace platform, as it hosts many open-source models as an ML hub. All 3 models are 7B parameter versions of their family. Besides, we did not use the base models; all 3 models were the finetuned versions by instruct datasets, hence the chat version for Llama-2 and instruct versions for Gemma and Mistral. While Gemma was trained on 6T tokens and Llama-2 on 2T tokens, the token size is unknown for Mistral as they have not disclosed the info. The table below summarizes the model descriptions.

Table 19. Model Descriptions of Generative Models Used in Experiment

| | Parameters | Version | Training Token Size |
|---|---|---|---|
| Llama-2 | 7B | Chat | 2T |
| Gemma | 7B | Instruct | 6T |
| Mistral | 7B | Instruct | NA |

For training, transformers (Wolf et al., 2020) and trl (von Werra et al., 2020) packages from HuggingFace were used. Since these models have a huge number of parameters, we employed the quantization technique and LoRA configuration to train the model with lower resources (Hu et al., 2021). A supervised-finetuning pipeline was applied to the models by using the training dataset. The hyperparameters used in training are 2e-4 for the learning rate and 4 for the accumulation steps for gradients. Table 20 below lists the hyperparameters used for training. As a result of the training, we obtained low train loss, which shows potentially successful learning. The details of the losses are given below.

Table 20. Train Loss for Generative Models

| | Llama-2 | Gemma | Mistral |
|---|---|---|---|
| Train Loss | 0.122000 | 0.122600 | 0.087600 |

### 4.3.4 Test results

After the training part was completed, we tested all 3 models on the test dataset that we created. Regarding the completions of the models to the prompt given, Gemma and Mistral learned the pattern so accurately that their response was only the label they predicted. For Llama-2, it performed almost the same as the other two regarding how it should form its response, but it very rarely produced no label in its response, for which we ran the inference another time. Without any further issue, it produced the prediction in its completion, and we were able to collect all the responses. It is also remarkable that Llama-2 almost always continued its prediction with the reasoning it used while classifying the relationship between sentence A and sentence B. For reference, some responses given by Llama-2 to the test pairs are included in Appendix B. The metric results of 3 models are given in Figure 6 below.

According to the scores above, Gemma and Mistral excelled in the task with an accuracy score of 0.98, which is just 0.01 above the BERT-NLI model developed in the previous section, although it has been also been shown that decoder-only models might not be as good as encoder-decoder ones (Deng et al., 2023). On the other hand, Llama-2 shows a relatively poor performance with an accuracy score of

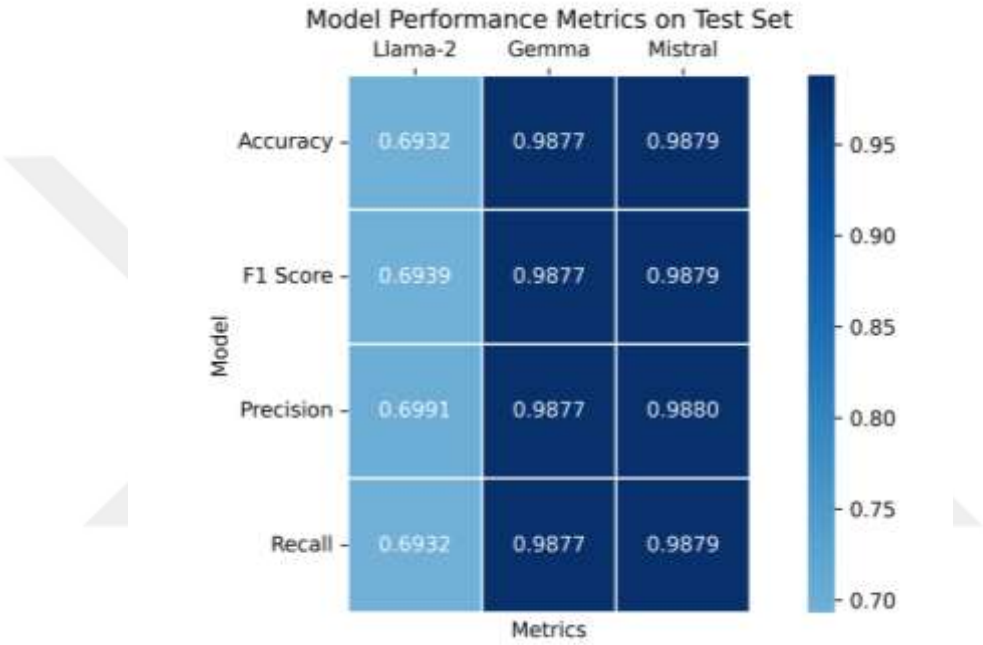



Figure 6. Finetuned model performance metrics on test set

0.69. The primary reason for this might be the models themselves: Gemma was trained on a three times larger corpus with 6T tokens than Llama-2, which is also the case for the Mistral model although we do not know its training details. A larger training corpus might have resulted in better reasoning abilities. Second, the training procedure might affect the performance, though it is unlikely because the models are generally trained similarly, and we did not include any extreme hyperparameter. Finally, the reason for the worse performance could be the prompting. This is a possible scenario because each model necessitates the prompts similarly to how it

was trained. While our prompt is a generic one, it is actually not in the exact way that Mistral authors want it to be (with [INST] tokens). Therefore, how much prompting affects the model performance in such a task should be investigated with several alternatives.

## 4.4 Discussion

In Experiment 1, we have put our dataset in use by training several models and evaluating their performance. BERT-like encoder-decoder LLMs are bidirectional models that take into account the left context and right contexts of a masked word in given sentences. This translates to their deeper understanding of the linguistic structures in data. In this regard, in the first section, the BERT models performed quite well, given that they were only given a premise and hypothesis. As implicature reasoning in humans increase with more context, we might have expected a worse score by the models which are given no context; however, they, especially BERT-NLI, acquired the patterns within the data. BERT-NLI had been trained on several NLI datasets, from which we can infer that the previous training on textual inferences was transferred to the slightly more complex task of identifying implicatures besides entailment. However, it is noteworthy to further investigate the influence of NLI training on more complex entailment reasoning.

In the SNLI and MNLI benchmark, we saw that our finetuned BERT-NLI performed worse than the original BERT-NLI with around 12% decrease in accuracy. Converting the predicted implicature labels into contradiction resulted in a better performance than the other cases. We inferred, as discussed in the literature, that the way we constructed the neutral and contradiction sentence hypotheses, which have maximal scalars and negation, confirmed the fact that these datasets are biased towards certain constructions and items in hypotheses sentences.

On the other hand, generative models seem to perform equally with the models developed in the first section. The models used are state-of-the-art models that proved their reasoning skills in different settings. Thus, the results they obtained here may be considered expected. The training showed that prompting is crucial in getting accurate completions out of the models. Moreover, converting a text classification task into a seq2seq task, as we did, does not seem to cause poor performance. The task can be converted in some other ways to scrutinize the model's abilities. For instance, in a study by Kumar and Talukdar (2020), the authors develop a pipeline that they call "NILE" where a premise-hypothesis pair is fed into a model that returns explanations for each of the candidate labels. Another model analyzes the explanations and chooses the correct label accordingly, which increases the model's performance. Consequently, the model's ability to classify labels into categories accurately in the specific test environment we created does not necessarily generalize to the model's absolute understanding of implicature relationships. For those cases, further studies should be employed with differing settings.

# CHAPTER 5

# EXPERIMENT 2: LINGUISTIC INQUIRY

Upon examining LLMs' abilities in identifying scalar implicatures and pragmatic reasoning in the previous experiment, in this chapter, we aim to investigate the effect of the various linguistic categories on implicature reasoning. For this aim, at the first stage, we developed 5 different models: each model is trained by leaving out exactly one linguistic category that ImplicaTR covers (adjectives, verbs, quantifiers, modals, and numerals) in its training data. Then, each of these 5 models was tested on the linguistic category they were not trained on. The first stage aims to see whether models can generalize the implicature reasoning they obtained to a novel linguistic category. In the second stage, one more model, Model-6, was developed. This model was trained on all of the linguistic categories that ImplicaTR has with one difference: we eliminated some of the scalar pairs from each of the linguistic categories (except modals, more on this later) and trained the Model-6 on the remaining ones. Then, this model was tested on the scalar pairs it did not see during the training procedure. The aim of this second stage is to further investigate if models can generalize the scalar reasoning that they gather from one specific linguistic category to the other lexemes in the same category. Each of these models will be called with their respective numbers throughout the study (e.g., Model-1).

In this chapter, we will first see how the dataset was split to train, validation, and test sets for all the models by elaborating on the first stage, which consists of the first 5 models, and then on the second stage, which covers the Model-6. The training procedures and loss-accuracy values will be given for each of the models. Then, we will investigate the test results per each of the models. What follows will be a further

examination by class and by linguistic category. After that, we will conduct a feature analysis to see how the linguistic structure affects the model performance. Lastly, we will do four small experiments for different linguistic categories of ImplicaTR before concluding the chapter with a general discussion section.

## 5.1 Splitting dataset

### 5.1.1 Models 1-5

Unlike Experiment 1, where the train, validation, and test datasets are created by means of stratified sampling, we will eliminate one of the linguistic categories entirely from the training and validation dataset. It will be ensured that the model will not see any of the samples from each category. Table 21 below shows the categories used for training and testing per each model.

Table 21. Linguistic Categories Used for Training and Testing for Each Model

| | Train | Test |
|---|---|---|
| Model-1 | Adjectives<br>Verbs<br>Quantifiers<br>Modals | Numerals |
| Model-2 | Adjectives<br>Verbs<br>Quantifiers<br>Numerals | Modals |
| Model-3 | Adjectives<br>Verbs<br>Modals<br>Numerals | Quantifiers |
| Model-4 | Adjectives<br>Quantifiers<br>Modals<br>Numerals | Verbs |
| Model-5 | Verbs<br>Quantifiers<br>Modals<br>Numerals | Adjectives |

As discussed previously, ImplicaTR features numerals and partitives; partitives use the same scales as numerals do. However, partitives do not show regular entailment and implicature relationships similar to other linguistic categories, which is why they are considered to hinder the model in learning the numeral constructions. Therefore, partitives are not included along with the numeral data in order to correctly evaluate the model's predictive power for the numeral dataset. The same case is also applicable to the C implicatures where the sentence C from each set of sentences leads to two extra scalar implicatures and one existential implicature.

Similar to how the dataset was split in the previous experiment, the train and validation sets were sampled with stratified sampling. Again, we first randomized and divided the sets (a set consists of 8 sentences) from each category into train and validation sets. Afterwards, the train and validation splits were populated with the 8 sentence pairs for each set they had. As a result of this process, a test set that includes the entirety of a single linguistic category has been created along with training and validation sets by means of stratified sampling. Thanks to the sampling method, the resulting datasets are ensured to have a balanced distribution across the four target variables we have. The following is the respective number of data rows for each dataset.

Table 22. Train, Validation, and Test Datasets for Models 1-5

| | N of rows | % |
|---|---|---|
| Train | 11520 | 64% |
| Validation | 2880 | 16% |
| Test | 3600 | 20% |
| Total | 18000 | 100% |

### 5.1.2 Model-6

As a successive study to the first 5 models where the whole linguistic category is tested, in this model, only some scalar pairs from each category were eliminated from the training procedure. The resulting model was then tested on the scaler pairs that were removed during the training procedure. Table 24 gives the list of scalar pairs that are removed from each linguistic category.

This removal was done for the adjective, verb, quantifier, and numeral categories and not for the modals. The reason for this is the fact that modals have only two scales, each of which consists of 225 sets, and eliminating one scale would mean eradicating half of the dataset. As it will be inconsistent with other categories, we have decided to retain the whole category of modals in the training dataset.

For each of the other four categories, the first 150 sets out of 450 sets were eliminated from the training dataset. This led to the fact that a total of 5 scales from adjectival scales were entirely excluded in the training procedure, whereas this number is 3 for verbs, quantifiers, and numerals.

The resulting sets for training and validation were again stratified across the sets within each linguistic category. By populating the stratified sample, we have created the final dataset. Similar to the previous models, the balanced distribution of entailment, neutral, contradiction, and implicature labels are ensured by the sampling method. Below is the resulting dataset's size.

Table 23. Train, Validation, and Test Datasets for Model-6

| | N of sets | N of rows | % |
|---|---|---|---|
| Train | 1320 | 10560 | 59% |
| Validation | 330 | 2640 | 14% |
| Test | 600 | 4800 | 27% |
| Total | 2250 | 18000 | 100% |

The scales from each category left out in the training procedure and tested subsequently are as follows.

Table 24. Scales Tested in Model-6

| Linguistic Category | Scales |
|---|---|
| Adjectives | benzer-aynı<br>yeterli-tam<br>zararlı-ölümcül<br>vasat-kötü<br>yakın-bitişik |
| Verbs | yönelmek-varmak<br>denemek-başarmak<br>öğrenmek-ustalaşmak |
| Quantifiers | birkaç-bütün<br>birkaç-tüm<br>birkaç-hepsi |
| Numerals | iki-üç<br>beş-yedi<br>on-on iki |

## 5.2 Training and testing the models

The datasets obtained are used to train 6 different models, which will be employed to test the respective datasets. The base model for each of the six models is the BERT-NLI model used in the first experiment (Laurer et al., 2023). As mentioned previously, BERT-NLI is a zero-shot classifier model that was a DeBERTaV3 model (He et al., 2023). Among alternatives such as the base BERT or BERTurk, BERT-NLI is employed throughout this experiment as it has been trained on the main NLI inferences previously. The model has already seen the entailment, neutral, and contradiction relationships which are correlated to the implicature-raising environments. BERTurk could have also been used for this section; however, as these models are multilingual and they are shown to transfer the learning process in one language to another, we have opted for BERT-NLI for this experiment.

The hyperparameters throughout this chapter are the same as those used in Experiment 1. First, we will see how the models 1-5 performed in training and testing, and subsequently, we will look into the details of the Model-6.

### 5.2.1 Training and testing the models 1-5

Train loss and validation loss for the first five models were seen to be decreasing steadily for every model and there was no sign of overfitting or underfitting except for the numerals (Model-1). The figures below show the respective train loss and validation loss of the first 5 models. The categories given in the parentheses next to the model's name refer to the linguistic category that is tested in that particular model:


Figure 7. Train loss for the models 1-5

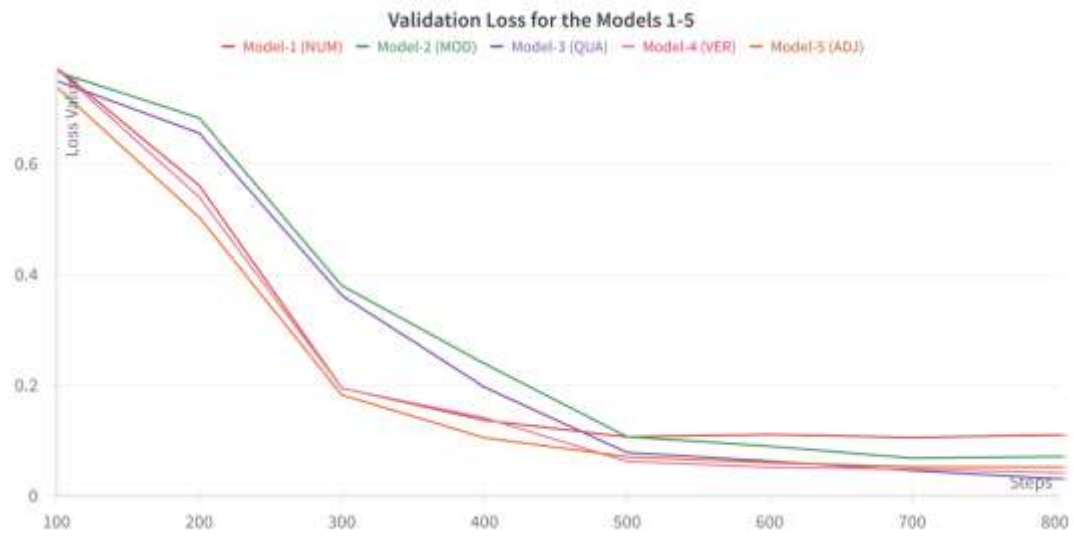



Figure 8. Validation loss for the models 1-5

As seen in the loss charts, the loss values decrease constantly until they hit a plateau and stop improving. There is a very slight sign of overfitting for the numerals, where the train loss is 0.066 and the validation loss is 0.107. Compared to other results and considering the test results, this is not suggestive of a significant overfitting case that hinders the model performance. To illustrate better, Table 25 below shows the respective train and loss values of each model along with the accuracy on the validation dataset.

Table 25. Train and Validation Loss Values of Models 1-5

| | Train Loss | Validation Loss | Validation Accuracy |
|---|---|---|---|
| Model-1 | 0.066500 | 0.107452 | 0.979167 |
| Model-2 | 0.095800 | 0.064293 | 0.987504 |
| Model-3 | 0.078700 | 0.026510 | 0.994553 |
| Model-4 | 0.056700 | 0.040738 | 0.991669 |
| Model-5 | 0.074700 | 0.050688 | 0.991029 |

The accuracy scores on the validation set also increase until reaching maximal values and stopping improving. Figure 9 below demonstrates how the accuracy scores improve by training steps.

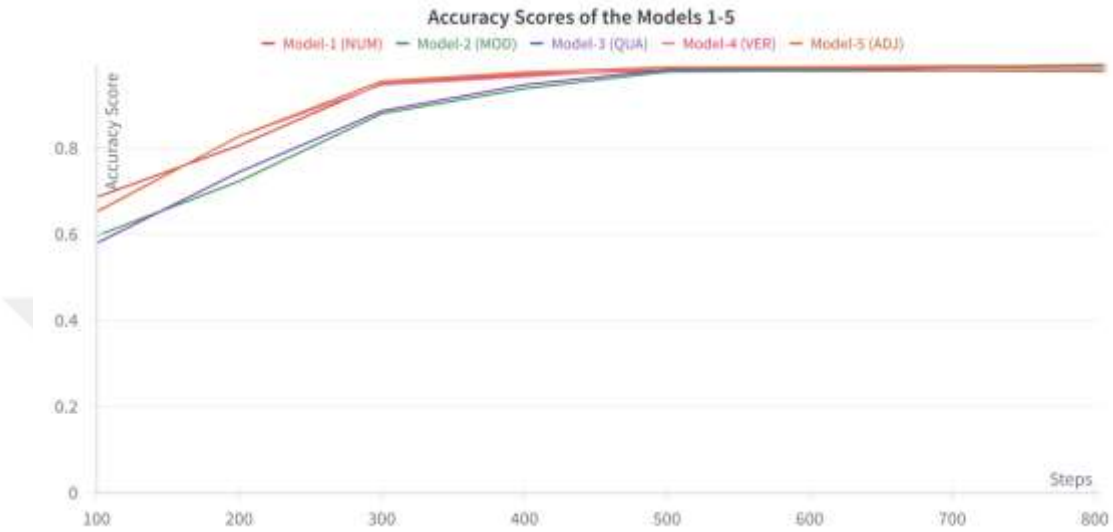



Figure 9. Accuracy scores of the Models 1-5 (validation sets)

The figure above indicates that the modal and quantifier tests converge relatively slower than the other sets. While adjectives, verbs, and numerals conjointly exhibit similar decrease patterns in loss and increase patterns in accuracy, Model-2 and Model-3 values are an indication that the absence of these two categories challenges the model to learn the patterns, which might show that quantifiers and modals are alike in certain respects while they are different from the other linguistic categories in terms of scalar reasoning. On the other hand, it might also be argued that the BERT-NLI model carries a bias in favor of quantifiers and modals, which would result in the fact that their presence accelerates the convergence. However, solely depending on the training loss values would yield quite unsupported claims, which is why further inquiry is needed.

*Testing the Models 1-5*

After each model was developed, the tests were done on the dataset that had been left out during the training procedure. The results are given in Table 26 below.

Table 26. Test Results of Models 1-5

| | Category Tested | Test Loss | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| Model-1 | Numerals | 1.5592 | 0.6036 | 0.5506 | 0.6723 | 0.6036 |
| Model-2 | Modals | 0.1416 | 0.9622 | 0.9621 | 0.9644 | 0.9622 |
| Model-3 | Quantifiers | 0.2860 | 0.9336 | 0.9340 | 0.9396 | 0.9336 |
| Model-4 | Verbs | 0.7137 | 0.7969 | 0.7948 | 0.8153 | 0.7969 |
| Model-5 | Adjectives | 1.3374 | 0.7152 | 0.7150 | 0.7169 | 0.7152 |

Across different linguistic categories, test results demonstrate differing performances for each model. First, the most successful categories are the modals and quantifiers. In Model-2, where the category of modals is left out in training and tested subsequently, the performance of the model reaches 0.96 for the accuracy score, which can be evaluated to be quite an excellent result. The model accomplished to correctly predict the entailment, neutral, contradiction, and implicature labels for modal sentences without seeing them formerly. Quantifiers were similarly predicted with a higher accuracy though having been vanished from the training set.

In the training process, these two categories were the categories with a somewhat divergent path of convergence compared to other models. This is to say that the presence of these categories accelerated the learning process, which might be accepted as a sign that the BERT-NLI model is prone to learn the implicature and entailment relationships for the scalars in these categories. The test results support this idea: quantificational and modular scalars are easier to understand for the model.

On the other hand, adjectives, verbs, and numerals have lower accuracy values compared to quantifiers and modals. In Model-4, where verbs are tested, the

accuracy score is 0.79, indicating that the model was more than moderately successful in predicting the verb inferences without seeing them. Following the verbs, adjectives seem to have moderate accuracy. These two categories, adjectives and verbs, are similar to each other in that they are content words. As they are content words, the lexicon of adjectives or verbs is quite larger than those of quantifiers and modals. Therefore, each lexical item from these content-word categories has lower frequencies than those from function-word categories. As a result, individual content words are seen more infrequently in pre-training, and LLMs have less information regarding scalar content words and their possible scalar spouses. This can explain why we see that scalar reasoning from other categories is generalized to adjectives and verbs, nevertheless, with lower accuracy.

Unlike adjectives and verbs, numerals are the least successful category, with an accuracy score of 0.60. In the same vein, cardinals have an unlimited number of lexical elements in theory, though, in practice, this number is still extremely large. Thus, the number of scalar relationships that a particular numeral can establish is quite large, the majority of which are unknown to the model or not reinforced in pre-training. This again results in the fact that numeral scalars can copy the scalar reasoning from other categories, though with worse performance.

These results require an in-depth look that will reveal the causes of difference, if they are statistically significant. In addition, we will need to scrutinize the predictions of the model in comparison to the ground labels in order to grasp the logic and reasoning of the model during inferencing. Before a detailed investigation of these cases, let us look into the training and test results of Model-6.

### 5.2.2 Training and testing the Model-6

For Model-6, we used the same training setup and the same hyperparameters except for the batch size, which is 32 for this model. The train and validation datasets are different in structure compared to the first 5 models: this model is exposed to every category during the training. What it is supposed to accomplish is correctly predicting other scale pairs from the same categories it has seen during the training. Train loss, validation loss, and accuracy scores of the model is as follows.

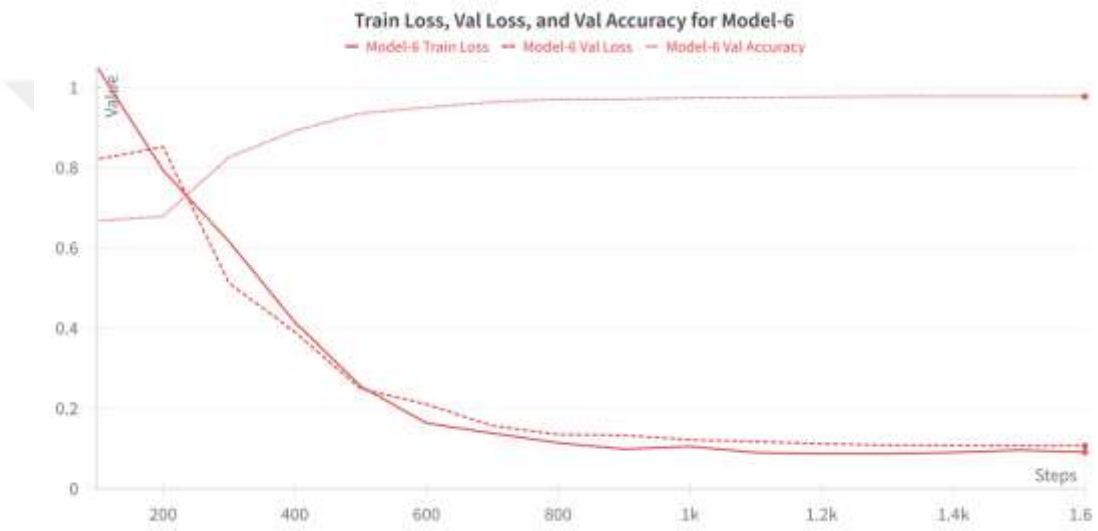



Figure 10. Train loss, val loss, and val accuracy for Model-6

Similar to what we have seen in the previous models, the loss values degrade steadily through training steps. As the loss values decrease, the accuracy score of the model increases and reaches a plateau. Regarding the loss values, similar to Model-1, where numerals were tested, we see an insignificant sign of overfitting where the validation loss is very slightly higher than the training loss, although this tiny difference will not create a serious performance problem for the model. For reference, the final loss and accuracy values are given in Table 27 below.

Table 27. Train Loss, Val Loss, and Val Accuracy for Model-6

| | Train Loss | Validation Loss | Validation Accuracy |
|---|---|---|---|
| Model-6 | 0.091500 | 0.107118 | 0.978030 |

After the model was trained, it was tested on the test data we had created that lacked some of the sets and scales altogether. As mentioned, modals are not included in this test as their scale number is very limited, and the results will not be consistent with the results of other categories. As a result, we have obtained a test result for the whole dataset as well as specific test results for each of the four linguistic categories, as given below.

Table 28. Test Results of Model-6

| | Test Loss | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Adjectives | 0.0903 | 0.9733 | 0.9734 | 0.9743 | 0.9733 |
| Verbs | 1.5226 | 0.6625 | 0.6625 | 0.6626 | 0.6625 |
| Quantifiers | 0.0926 | 0.9866 | 0.9866 | 0.9866 | 0.9866 |
| Numerals | 0.4532 | 0.8541 | 0.8538 | 0.8613 | 0.8541 |
| Average | 0.5397 | 0.8691 | 0.8691 | 0.8697 | 0.8691 |

First, the overall accuracy is relatively higher compared to the 4 of the first 5 models (as modals are not included here). While the numeral and adjective adjectival categories show an increased performance, quantifier predictions are again very successful with a high score. Verbs, on the other hand, showed a decrease in performance, which is contradictory to the other categories. We can basically infer that the scalar reasoning exists within the linguistic categories for adjectives and numerals so that training on similar structures gained the model the scalar implicature reasoning that it used for unseen scalar pairs.

The score of adjectives is significant here as it reaches to 0.97, much higher than Model-5 results (0.71) where the model was tested on the adjective dataset.

Numerals are also noticeable in that they show an increase from 0.60 to 0.85 when they are also included in the training dataset.

The peculiar case of verbs indicates a decline in the accuracy score compared to what we have seen in Model-4, where the test result was 0.79, which does not seem in line with other patterns. One possible explanation might be the verb types tested. As seen in table above, the tested verbs are the inceptive verbs. The difference between the type of verbs in training and test sets might have resulted in a lower performance. On the other hand, verbs are morphologically altered from sentence to sentence, and among the linguistic categories ImplicaTR possesses, verbs are the only category in the sense that they must undergo a change in their morphology. As well as tense and aspect changes, various morphological rules of Turkish also affects the root form, e.g., from *başla-mak* (to start) to *başl-ıyor* (he/she is starting). This decreases the frequency of a particular verb token within the corpus, which might be the reason for the low performance.

## 5.3 Results

After examining the preliminary test results for each of the models, we further analyzed the predictions given by each model to reveal the factors that affected their predictions. We aim to see how these factors affected models' correctness for each linguistic category. We collected all predictions given by each model for their respective test datasets.

As the first step, we conducted a chi-square test in order to attest that the accuracy scores for linguistic categories are statistically different from each other. As our target variable is the prediction of the model as one of the four categorical variables instead of a continuous one, we have used chi-square instead of ANOVA

for accurate results. The table below shows the Pearson Chi-Square value of the tests along with Cramer's V value.

Table 29. Chi-Square and Cramer's V Results

| Pearson Chi-Square | p-value | Cramer's V |
|---|---|---|
| 1256.2951 | <0.0001 | 0.1525 |

The differences between the linguistic categories in their predictions are statistically significant with a chi-square value of <.0001. This value, accompanied by such a high chi-square statistic, shows that the difference is quite large between the categories while the size of our test results might intervene and increase the number even further (McHugh, 2013; Schumacker & Tomek, 2013). Cramer's V is a value to demonstrate the strength among the differences between the categories. Cramer's V above refers to a relatively small effect size in general; however, with the sample size being quite large, the resulting value might not be as important as expected. With the values obtained, we can reject the null hypothesis that there are no significant differences between categories and accept that statistically significant differences exist.

We have also conducted a Tukey's HSD to inspect the pairwise group comparisons between the linguistic categories (see Appendix D). The results show that categories are different from each other with a very low p-value (<.001) except for the quantifier-modal comparison, where the difference is still significant at the level of 0.011.

After verifying the results between the categories, the table below shows the mean accuracies for each category again with other statistical information. As Table 30 shows, Standard of Error (SE) is very low for each of the categories, indicating that the mean accuracies are a trustworthy statistic for the differences.

Table 30. Accuracies and Statistics for Models 1-5

| Linguistic Category | N | Accuracies | SD | SE | 95% Conf. | Interval |
|---|---|---|---|---|---|---|
| Adjectives | 3600 | 0.7153 | 0.4513 | 0.0075 | 0.7005 | 0.7300 |
| Verbs | 3600 | 0.7969 | 0.4023 | 0.0067 | 0.7838 | 0.8101 |
| Quantifiers | 3600 | 0.9336 | 0.2490 | 0.0041 | 0.9255 | 0.9417 |
| Modals | 3600 | 0.9622 | 0.1907 | 0.0032 | 0.9560 | 0.9685 |
| Numerals | 3600 | 0.6036 | 0.4892 | 0.0082 | 0.5876 | 0.6196 |

A very basic interpretation of why we have such differences without doing further analysis requires awareness of how these categories are treated in NLP models. Quantifiers and modals are the categories with the most limited number of members among the linguistic categories in the study. Quantifiers, POS-tagged as DET in NLP literature, are among the most frequent tokens found in corpora (Roland et al., 2007). It is reasonable that the information they possess have already been captured by the LMs. The total statistics of pretraining corpora in terms of word frequencies is unknown, but we would expect a similar pattern there. With regards to the implicature inference that our dataset introduced, it is possible that the model needed to capture the linguistic and logical aspects of implication from other linguistic categories while the conventional entailment-neutral-contradiction reasoning was already existent. Considering this accuracy score, it is no surprise that quantifiers preserve their high score in Model-6 as well, where the training data already has quantifiers.

Modals are the other closed class we feature in the study along with quantifiers and numerals. As we have discussed in the previous experiment, we have two scales within the modal category, both of which denote epistemic modality. While both epistemic and deontic modalities are prevalent in corpora, studies such as Römer (2004) or Millar (2009) show that the epistemic use of modality lexemes is extensive. This is also the case for the words that have possibility, ability, and

permission meanings; they are mostly used in their possibility denotation. We can again make a naïve interpretation that the high frequency of epistemic modality might be a reason why we see such high accuracy in modals compared to other categories.

Two of the remaining three categories, adjectives and verbs, are open classes, which signifies that their total vocabulary size is excessively larger than the others. A natural consequence of this is that the individual frequencies of scale members are comparatively low in the training corpora of BERT or other LLMs. The model might have a relatively lower insight into the syntactic and semantic features of adjectives and verbs. The absence of these categories in the dataset further imposes a challenge on the model in that it would not be able to grasp more information about them. As a result, it is possible that the predictive power of the model regarding these structures results in being weak.

The fifth category, numerals, establishes the bottom line of accuracy scores among all categories. A first look at the case of numerals may reveal their distinct nature compared to other classes. A member from other classes, like adjectives, is in a scalar relation with only a limited portion of adjectives, and it is quite possible that models learn this relationship in training as they encounter those structures. On the other hand, because of their nature, a numeral word is in a scalar relation with all other numbers (theoretically infinite, practically a huge number) larger and smaller than itself. If we assume that this type of inference is based on learning relationships between words, this might explain why the model does worse when it is tested on scalar items that it has not encountered. These naïve interpretations, however, will require further investigation.

### 5.3.1 Results by label classes

After probing into the performance of our model per linguistic category, we also investigated the accuracies of the model to see how the model performed across different label classes, namely entailment, neutral, contradiction, and implicature. The table below shows the results for label classes.

Table 31. Accuracy Results by Label Classes for Models 1-5

| category | N | Accuracy | SD | SE | 95% Conf. | Interval |
|---|---|---|---|---|---|---|
| entailment | 4500 | 0.7229 | 0.4476 | 0.0067 | 0.7098 | 0.7360 |
| neutral | 4500 | 0.8496 | 0.3575 | 0.0053 | 0.8391 | 0.8600 |
| contradiction | 4500 | 0.8673 | 0.3393 | 0.0051 | 0.8574 | 0.8772 |
| implicature | 4500 | 0.7696 | 0.4212 | 0.0063 | 0.7572 | 0.7819 |

The results show that the model behaves differently across different labels, performing better in contradiction and neutral cases, whereas the accuracy score is 0.72 and 0.76 for entailment and implicature classes, respectively. The models we developed in these models seem to have grasped the relationship of contradiction quite well for linguistic structures that it was not trained on explicitly. This can be considered expected as these models were already efficient in NLI label predictions (N. Jiang & De Marneffe, 2019), specifically in contradiction due to the existence of "not" (Gururangan et al., 2018). The neutral cases are similarly easy to grasp for the model, which also seems in line with the literature. However, the case of entailment is surprising as it is the least successful category, with an accuracy score of 0.72.

Additionally, models seem to have reasoning regarding the implicatures in the categories they have not seen previously as the accuracy score is more than moderate with a 0.76. The heatmap below shows how the predictions are scattered across the classes. Table 32 presents a finer analysis regarding how the models performed between the two types of sentences of each label.

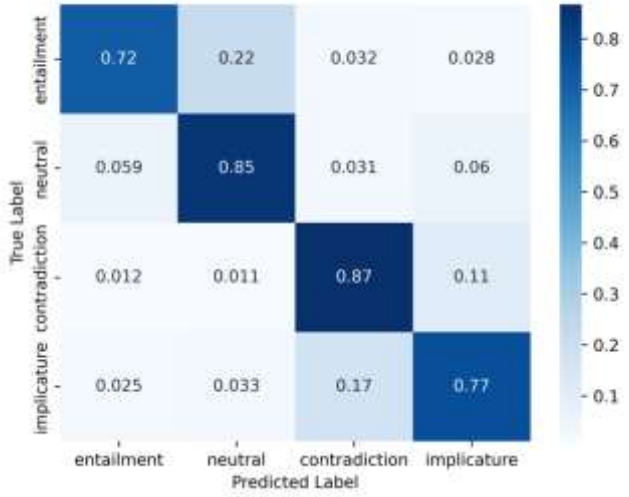



Figure 11. Confusion matrix of predicted and true labels across models 1-5

Table 32. Accuracy Results by Sentence Types for Models 1-5

| Sentence Type | N | Mean |
|---|---|---|
| Implicature 1 | 2250 | 75% |
| Implicature 2 | 2250 | 78% |
| Entailment 1 | 2250 | 72% |
| Entailment 2 | 2250 | 71% |
| Neutral 1 | 2250 | 92% |
| Neutral 2 | 2250 | 77% |
| Contradiction 1 | 2250 | 89% |
| Contradiction 2 | 2250 | 84% |

When the models predict the entailing premises-hypothesis pairs, they mostly confuse them with the neutral class with 22%. To remember, the sentence pairs with an entailment label and those with the neutral label are the exact opposite of each other, meaning that the premise in one is the hypothesis of the other and vice versa. In this sense, the comparison between entailment and neutral sentences is also important for the textual similarity tasks, where the two classes would fall into only one class (Agirre et al., 2012). It seems plausible to assume that the model appeals to the symmetric equivalence between the two sentences when it fails to predict the inference as entailment. As a result, the neutral label is falsely given to the sentence pair. Then, the inquiry should be on why the model fails to predict accurately the

entailment labels. For reference, Table 33 shows the precision of the models by classes. The precision for the entailment class is the highest with an important difference to the other classes, meaning that the model is reliable in the cases where it predicts the pair as entailment, though it is not able to grasp all the information regarding the entailment sentences. Thus, the number of predictions is higher for the other classes and lower for entailment.

Table 33. Precision by Label Class for Models 1-5

| Prediction | N | Precision |
|---|---|---|
| entailment | 3689 | 88% |
| neutral | 5001 | 76% |
| contradiction | 4959 | 78% |
| implicature | 4351 | 79% |

In the Table 32 above, the different sentence types for each class do not differ from each other except for the neutral case where the first sentence is more accurately predicted than the second one. When we investigate the confusion matrix for the second sentence of the neutral class (see Appendix C), we can see that the predictions are made as entailment when the model fails to interpret the sentence as neutral. This pattern is what we saw in cases where the model fails to predict entailments correctly, so we can further verify that the model uses sentence similarity for prediction as well.

### 5.3.2 Results by linguistic category

In addition to an analysis by the class labels, we further looked into how specific categories behaved across all class labels to have an insight where the linguistic categories create differences. The rest of the five confusion matrices are given in the Appendix C for reference (as values are rounded in heatmaps, they might not add up

to 1 or be only slightly over 1). We include the heatmaps of verbs and numerals here in Figure 12 and Figure 13 for discussion.

Adjectives show a consistent pattern across classes, where each class is predicted with an accuracy of around 0.7. This is an indication that the features of the adjectives behave more or less the same, and classes do not create an extra challenge for the model. It is still worth noticing that implicatures are relatively challenging as the accuracy is lower. This pattern of adjectives is not seen in verb prediction, where entailment and contradiction are predicted with high accuracy. For the neutral class,

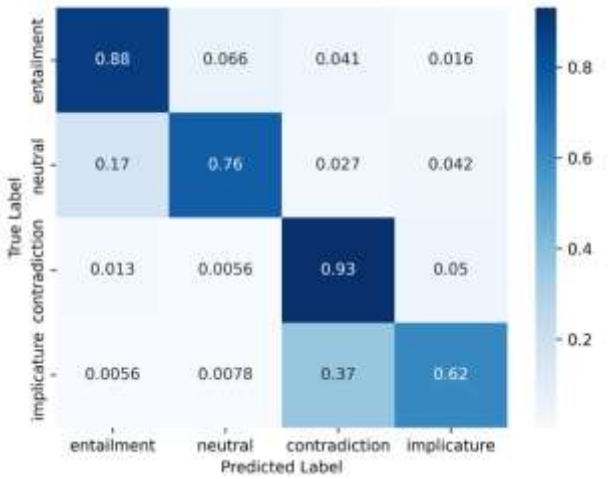



Figure 12. Confusion matrix of predicted and true labels for verbs

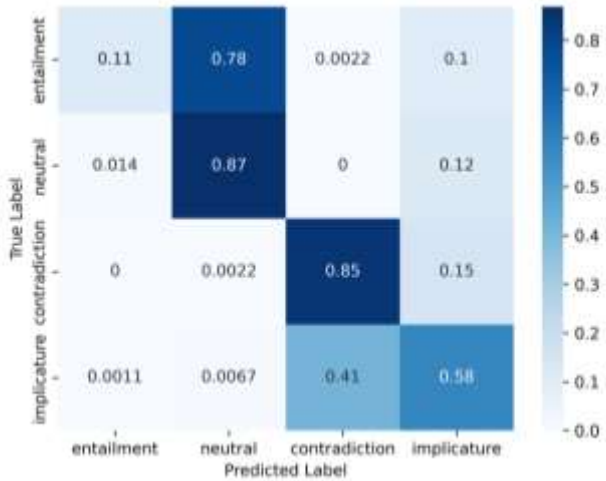



Figure 13. Confusion matrix of predicted and true labels for numerals

we see that when neutrals are predicted incorrectly, they are mostly taken as entailments, which again refers to the tasks of textual similarity. However, we do not see that implicatures are predicted correctly as much as other classes. Moreover, those cases are predicted as contradictions with a percentage of 37%. This exhibits a different pattern, similar to Experiment 1 results.

Regarding quantifiers and modals, which are the categories with the highest accuracies, we see that they demonstrate their high accuracy trend again across all classes. For implicature reasoning, modals show almost a perfect score near 1 (rounded to 1 in the heatmap). This might be an indication that the high frequency of these closed-class words positively influences the predictive power of the model.

What differs from the previous results is the category of numerals, where the neutral and contradiction accuracies are high while entailment and implicatures are quite low. Implicatures are almost divided between contradiction and implicature predictions with 41% and 58% of predictions, respectively. What is more curious is that the entailment cases are classified as neutral with a ratio of 78%. This again demonstrates that models refer to the similarity between the sentences and incorrectly assign the neutral label. For the incorrect implicature cases, we see that the incorrect classifications are for the first implicature sentence where the premise is the smaller cardinal and the hypothesis is the negation of the larger cardinal (e.g., *I read five books* and *I did not read six books*). This is an indication that such specific scalar numeral pairs (e.g., five-six, seventeen-twenty-nine) do not create scales automatically as other categories do. Figure 14 shows that implicature relationships between the larger and smaller values can be taught to the model. The implicature performance increases in Model-6, which saw some other numeral scales in training.

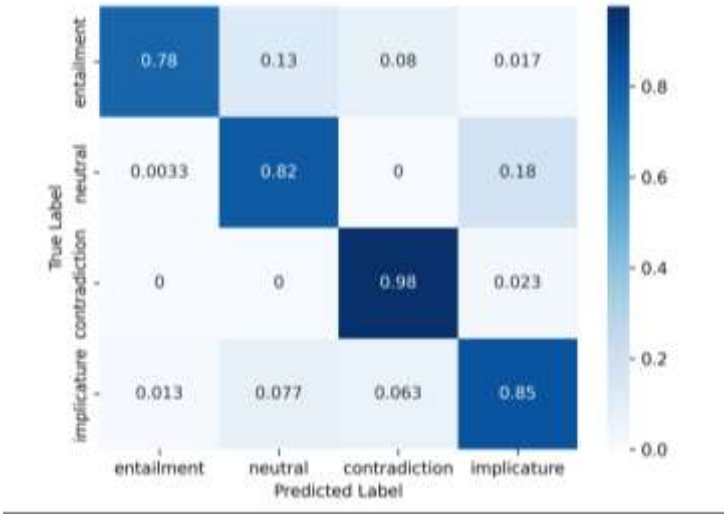



Figure 14. Confusion matrix of predicted and true labels by Model-6 in numerals

After we obtained and analyzed the category and class results, we further conducted a feature importance analysis. This analysis is significant in that it will highlight the linguistic features in data that contribute to the model's reasoning capabilities. In the next section, the features that are extracted are described first and then the related analyses are carried out with two models, logistic regression and random forest.

## 5.4 Feature importance analysis

In order to unveil the potential linguistic triggers in our dataset that lead to the correct or incorrect classification of the premise-hypothesis pairs, we conducted a feature importance analysis. Research has shown that LLMs already learn the linguistic features within the data in the pre-training phase, while linguistic information contributes to a higher performance level in downstream tasks (Radford et al., 2018). In this sense, extracting linguistic features from the sentences in the dataset would be convenient and immensely needed to see which structures contributed to the model reasoning. In literature, the variety of linguistic features

extracted from the corpus is quite extensive, ranging from position of specific POS-tagged tokens and the behavior of the root word to the depth and length of the dependency trees computed for each sentence (Miaschi et al., 2020). Besides, other highly relevant features are the statistics regarding the sentence or token length or average values of each sentence.

We have extracted some features from premises, hypothesis, and their combinations to carry out the importance analysis. In this regard, we used several tools, including spaCy, NLTK, HuggingFace transformers, sklearn, and fasttext. With these features, we tried out some statistical methods to see which ones serve our purposes better and finalize our results with the best ones. In addition, the scope of feature extraction for linguistic data might extend to unrestrained levels. In those cases, it is a good practice to choose among the most appropriate features for the purpose of the task, one of which includes manually combining the features to see which combinations result in better scores (Kriz et al., 2015). By testing different features, we came up with a final list of features that theoretically deliver the best results. In the next section, we first go through the features obtained from the data and then, the results from some statistical methods will be given.

### 5.4.1 Feature extraction

The feature extraction process was first started with the basic features that are commonly used across NLP studies, with token counts per sentence and average token length for each sentence. We obtained these two results for both the premise and hypothesis, which, as would be expected, gave quite similar results because of the data generation method employed. Apart from this, we collected the number of nouns per sentence pair this time instead of both premise and hypothesis since the

study does not manipulate nouns, and it is expected that both have the same number of nouns. The motivation for noun extraction stems from the study of Talman et al. (2021), where they conduct a series of experiments in which they corrupt the dataset in different fashions to see which features contribute less/more to the overall accuracy. Nouns are one of the categories that affect the accuracy of the NLI models in entailing and contradicting cases.

Another categorical variable that we extracted is whether the sentence root is a verbal or a nominal predicate. As many sentences are constructed with *var*, *yok* or an adjective as a predicate, this might be a relevant feature for our dataset. spaCy does not have native support for Turkish for now; therefore, for this task, we use the large spaCy model that Altinok (2023) developed, which works with an assuredly high performance.

In our dataset, the sentences A, B, C, or D from each set are created with the affirmative versions of weak and strong scale items along with the negations of those two sentences. This is why polarity can be considered an important feature of the dataset that might determine the quality of data. To this end, we also extracted features from the premises and hypothesis regarding their polarity in order to see its effect as well as to check the dataset once more. We created 3 different variables out of polarity: the polarity of the premise, the polarity of the hypothesis, and the combination of both (e.g. Positive-Negative). We again used the spaCy model developed by Altinok (2023). As this task requires a morphological analyzer, we used the transformer version instead of the large model this time.

Considering that named entities could interfere with the model's reasoning over sentence pairs, we also included NER tags in our features for inquiry into their potential influence on the results. We tried to include all of the named entities

existing in the dataset regardless of their counts. For this task, we again used the large spaCy model mentioned above. Similar to the polarity classification, we carried out sentiment analysis for each premise and hypothesis as well as their combinations. The classes used for sentiment analysis are positive, negative, and neutral. After examining negation with the morphological analyzer, this feature-extracting process is to verify further if sentences are semantically negative or positive. We used the same BERT-NLI model as a zero-shot classifier model that can easily adapt to this task.

One important statistic extracted is the similarity score between the premise and hypothesis. The model's prediction might be expected to be correlated with how much the input sentences resemble each other. Again, with the large spaCy model above, we compared the word embeddings of both sentences and gave them a score of similarity.

On the other hand, the similarity of the words within a sentence might be of high value for the model, which is why we also collected the similarity scores for each sentence as well. For this, the pairwise similarity of each word in a sentence is collected from a word embedding. Then, the pairwise similarity scores were summed and divided by the number of pairs in each sentence. The resulting score would be a word similarity score for that particular sentence. We used the fastText model for this task, which uses the sub-word information to construct word embeddings (Bojanowski et al., 2017). We also looked into the TF-IDF scores of the individual scale items we have in each sentence and obtained the scores for both premise and hypothesis sentences. TF-IDF is a method that considers the token frequency for a particular word as well as taking into account its presence across all documents. On this account, it gives a reliable score for the importance of the word and has been

shown to be used for feature extraction purposes in the literature (Wendland et al., 2021).

Finally, we collected the relative positions of the scalar items in the premise and hypothesis. This info would reveal if the position of the scalar item within the sentence is a key factor for the model. In addition, we took note of the words on the right and on the left of the scalar item in a given sentence. Being closer to the predicate or root, or the number of tokens between the scalar word and the root might affect the model's reasoning capabilities. While the number of tokens on the right and left are given as integers, position value is given as a number between 0 and 1, where being closer to 0 means being at the beginning and to 1 means being relatively at the end of the sentence. As a result of this stage, we extracted features to analyze further with statistical methods. Table 34 (on the next page) shows the features we have mentioned.

### 5.4.2 Logistic regression model

In the preliminary work on the logistic regression model, we found that some of the features extracted are insignificant in their effect size. Besides, they are not as frequent as the other features. Thus, we decided to remove them from the regression analysis. The features that were removed are those for the NER and sentiment analysis; however, a separate and limited regression analysis is included only for NER in Appendix D to show the effect of particular words. It can be seen that the existence of PERCENT in a sentence is highly correlated with being predicted corrected, as expected by the modal scale *herhalde-yüzde yüz*. On the other hand, the existence of CARDINAL negatively influences the predictive power of the model, as it is supposed by the category of numerals.

Table 34. Features Extracted for Analysis

| Group | Description | Names of Variables |
|---|---|---|
| **Counts and Lengths** | The counts of nouns and tokens, and average length of each token per premise-hypothesis | premise_noun_counts<br>hypothesis_noun_counts<br>premise_token_counts<br>hypothesis_token_counts<br>avg_premise_token_length<br>avg_hypothesis_token_length |
| **Verb** | Whether the predicate is nominal or verbal | premise_is_root_verb<br>hypothesis_is_root_verb |
| **Polarity and Negation** | The polarity of the sentence as obtained from the morphological markers on the root for premise and hypothesis. Also, the possible combinations between premise-hypothesis | premise_polarity<br>hypothesis_polarity<br>isPol_PosPos<br>isPol_PosNeg<br>isPol_NegPos<br>isPol_NegNeg |
| **NER** | The NER tags obtained from both premise and hypothesis | CARDINAL, GPE, PERCENT, ORG, NORP, LOC, MONEY, QUANTITY, DATE, TIME, PERSON, LANGUAGE, EVENT, WORK_OF_ART, FAC, TITLE, ORDINAL |
| **Sentiment** | The sentiment as predicted by zero-shot as one of positive, negative, or neutral | sentiment_premise_negative<br>sentiment_premise_neutral<br>sentiment_premise_positive<br>sentiment_hypothesis_negative<br>sentiment_hypothesis_neutral<br>sentiment_hypothesis_positive |
| **Word Similarity** | The average word similarity for each premise and hypothesis obtained from fastText and the difference between the two | premise_word_similarity<br>hypothesis_word_similarity<br>word_similarity_diff |
| **Sentence Similarity** | The similarity score premise-hypothesis pair calculated by the embeddings | premise_hypothesis_similarity |
| **TF-IDF** | TF-IDF score of the scalar item in each sentence for premises and hypotheses | premise_scale_tfidf<br>hypothesis_scale_tfidf |
| **Scalar Position** | The respective position of the scalar item within a sentence along with the number of tokens to left and to the right for both premise and hypothesis | premise_scale_position<br>premise_words_on_right<br>premise_words_on_left<br>hypothesis_scale_position<br>hypothesis_words_on_right<br>hypothesis_words_on_left |

Utilizing the remaining features, we developed the logistic regression model whose classification results are in Table 35 below.

Table 35. Classification Results of Logistic Regression Model

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| class | | | | |
| 0 | 0.59 | 0.05 | 0.10 | 716 |
| 1 | 0.81 | 0.99 | 0.89 | 2884 |
| | | | | |
| accuracy | | | 0.80 | 3600 |
| macro avg | 0.70 | 0.52 | 0.49 | 3600 |
| weighted avg | 0.77 | 0.80 | 0.73 | 3600 |

The table shows that the model works fine, with an accuracy score of 0.80. However, the precision and recall scores for the class 0 are relatively low. As this is not a model that classifies into one of the four labels but into 0 or 1, we consider that the score is good enough for our purposes. Figure 15 below shows the coefficients of the features fed into the model. Since many features returned a small coefficient, they are omitted in the figure below, and the full table is given in Appendix D.

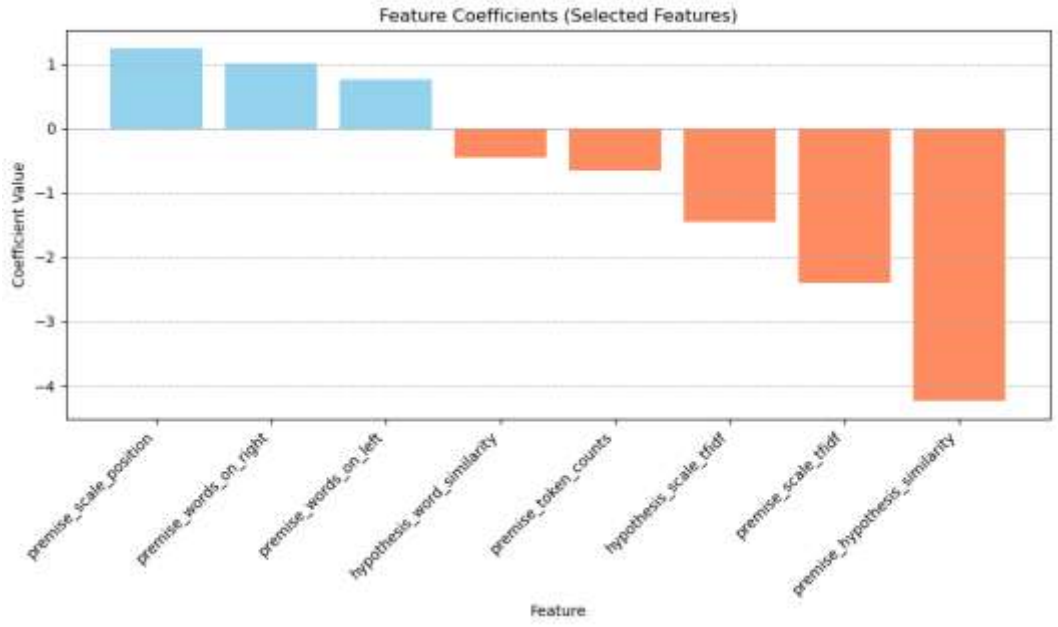



Figure 15. Coefficients of features by the logistic regression model

According to the coefficients, the value that positively affects the correctness of the model prediction the most is the position of the scale. Keep in mind that a higher value for scale position means being closer to the end of the sentence. Therefore, the model predictions show improvement when the scale in the premise sentence is closer to the end. The feature that characterizes the number of words to the right of the scalar item in the premise sentence also seems highly correlated with the correctness of the prediction. When the two are taken together, we can presume that as the scale is near the predicate of the sentence but there are some other words in between, which, for instance, might mean that the verb phrase is populated, the model more correctly analyzes the inference between the two. It is also worth noting that the scale position of the hypothesis does not affect the correctness as much as the position in the premise sentence does.

On the other hand, the TF-IDF score is negatively correlated with correctness. It is important to remember that the TF-IDF score is a score that is crucially correlated with the corpus, which is, in our case, our dataset. Though it might prevent a generalization to other domains, it is still an indication that as the word is more peculiar to the sentence among the documents, it might intervene in the model's performance. Critically, the correctness is negatively correlated the most with the similarity between the premise and hypothesis sentences. When they resemble each other in that their embeddings are similar to each other's, this adversely contributes to the likelihood that the model makes a correct prediction.

### 5.4.3 Random forest model

Upon obtaining results from the logistic regression model, we carried out a random forest classifier to further verify the effects of the features on the model predictions'

correctness. For this, the features whose effect size is low were eliminated, and a new set of features was created. This new set of features does not include any categorical variable, and all of the values are continuous. The list of variables is given in Appendix D. The classification results of the random forest model are given in the table below. Accordingly, the model has an accuracy level of 0.80, which shows that it mostly learned the relationships between the features. Also, the figure of coefficients for the model is given below.

Table 36. Classification Results of Random Forest Model

| class | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.12 | 0.44 | 0.19 | 194 |
| 1 | 0.96 | 0.81 | 0.88 | 3406 |
| | | | | |
| accuracy | | | 0.79 | 3600 |
| macro avg | 0.54 | 0.63 | 0.53 | 3600 |
| weighted avg | 0.92 | 0.79 | 0.84 | 3600 |

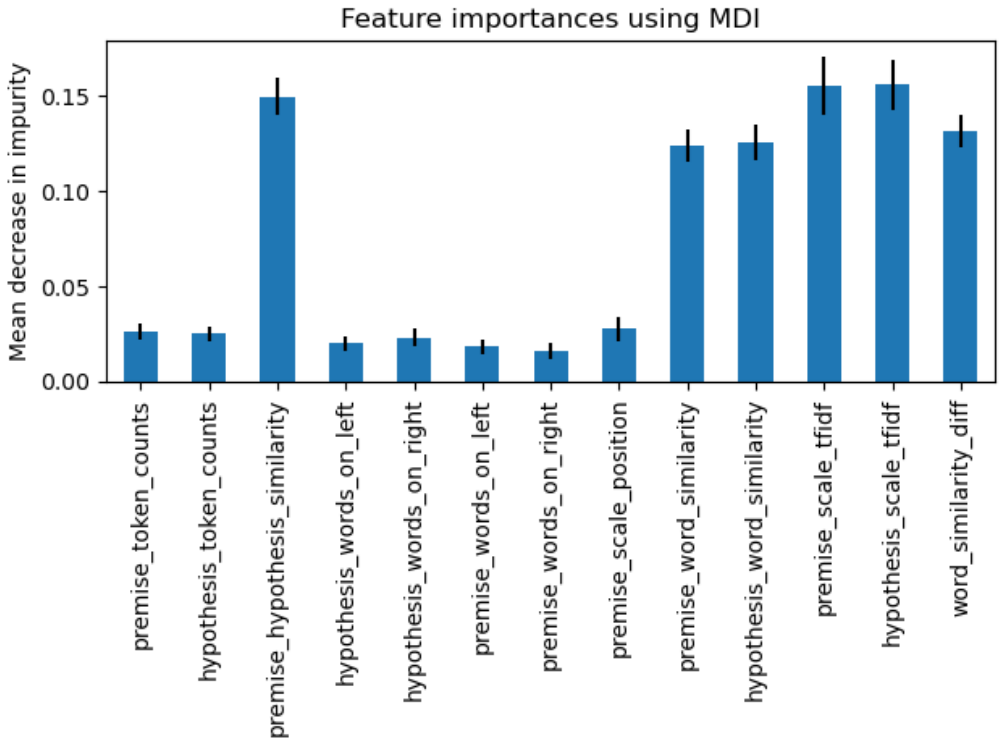



Figure 16. Random forest coefficients

We see that the results of the random forest model are in line with the regression analysis we did. In this model, where the coefficients are calculated by the decrease they cause in the mean accuracy (MDI), the features with the highest decrease are TF-IDF scores of the scalar items within the sentence. This is valid for both premise and hypothesis sentences, as the values of both are the highest. The similarities of the embeddings of the premise-hypothesis pair can be seen to have a negative effect on the correctness of the model prediction. Additionally, the average similarity scores of the words within a sentence are again one of the factors that decrease the score.

## 5.5 Further linguistic inquiries

ImplicaTR includes manipulations to some extent for some of the linguistic categories, as discussed previously. By manipulation, we refer to the systematic data generation that has been done through using certain linguistic phenomena. These manipulations are not at large scale; however, they still shed light on the various characteristics of some phenomena in scalar implicature derivation. This section looks into the details of these manipulations.

The linguistic categories that are systematically manipulated are modals, quantifiers, and numerals. We first look into modals, for which we created the sentences by exploiting the tenses and aspects for each set of sentences. The list of tenses and aspects employed are given in Table 8 in Chapter 3. Second, for quantifiers, the scalar pairs are constructed with combinations of different quantifiers in Turkish. We will see if certain scalar items are learned better or not. Then, we will see two cases for numerals: partitives and C implicatures. Partitives are the numeral structures that substantially change the regular entailment and implicature pattern we

see throughout this study. They will be investigated to see whether they obey the implicature reasoning of numerals or not. Lastly, we will scrutinize the case of C implicatures, which are created by pairing the sentence C from each set of ImplicaTR as the premise and three extra hypothesis sentences created specifically.

### 5.5.1 The case of tenses and aspects

Modal sentences were created by employing certain tenses and aspects for each verb predicate. In total, 9 different aspects exist in 450 different sets of modals. The motivation for this small experiment is the fact that verb aspect has been shown to affect the predictions and performance of BERT models (Cho et al., 2021). As the scalar pairs in modals are epistemic modalities, the aspects must give way to the *possibility* meaning without any extra problem. This is why a limited number of aspects were chosen and tested.

Results show that almost all of the aspects of entailment and implicature reasoning are easily learned by the model. As we have seen previously, modals always had a high accuracy in the models that we developed, which is why there is not much difference between different aspects. One different case is the presumptive future aspect (*Muhtemelen yapacakmış. – (I heard that) s/he will probably do it.*). While the accuracy is still high, it has obviously poorer accuracy than other aspects. This does not demonstrate any conclusion on itself; still, we think that the presumptive meaning might have an influence on this, although other presumptive aspects have higher accuracy. A further study would be on verb aspects in different linguistic categories.

Table 37. Accuracy Scores by Verb Aspect in Modals

| Aspect | N | Accuracy |
|---|---|---|
| Continuative | 400 | 99.75% |
| Emphatic Presumptive | 400 | 99.25% |
| Future | 400 | 98.75% |
| Presumptive Past | 400 | 97.75% |
| Presumptive Aorist | 400 | 97.50% |
| Emphatic Future | 400 | 96.50% |
| Past | 400 | 96.50% |
| Aorist | 400 | 95.50% |
| Presumptive Future | 400 | 84.50% |

### 5.5.2 The case of quantifiers

In NLI research, the syntactic and pragmatic variation among quantifier words is considered a challenge for the models, especially for multilingual models (Cui et al., 2022). Thus, we conducted this small experiment to see how the variation between scalar quantifiers in ImplicaTR might create a challenge for the models developed. The results for each scale are given in Table 38 below.

Table 38. Accuracy Scores by Quantificational Scales

| Scale | N | Accuracy |
|---|---|---|
| birkaç-hepsi | 400 | 99.00% |
| biraz-çok | 400 | 98.50% |
| bazı-hepsi | 400 | 96.50% |
| birkaç-tüm | 400 | 94.50% |
| birkaç-çok | 400 | 94.50% |
| bazı-tüm | 400 | 93.25% |
| birkaç-bütün | 400 | 91.50% |
| bazı-çok | 400 | 86.75% |
| bazı-bütün | 400 | 85.75% |

As the results of each model up to this point demonstrated, quantifiers always had high accuracies, which is shown by the table. It is hard to say that there are large differences between the accuracies of the scalar pairs although pairs containing *bazı* 'some' have a relatively lower accuracy than others. The quantifier *bazı* is an equivalent of some that can be used with plurals and, in limited contexts, with singulars (Özyıldız, 2017). We have not spotted any use difference between *bazı* and

other non-maximal quantifiers we have; therefore, the difference with other scales might be caused by data items, although we have not encountered any unusual case in ImplicaTR.

### 5.5.3 The case of partitives

Partitives are constructions that denote the incompleteness or partialness of any quantity. For instance, while an ordinary numeral construction is *Beş kitap okudum* (I read five books), we can construct partitive numerals by denoting some part of a greater whole such as *Kitapların beşini okudum* (I read five of the books). We saw that partitive constructions remove the regular entailment-implicature relationships that we observe in numerals and in other linguistic categories, which is possibly caused by the specificity of the NP (Enç, 1991) (refer to ImplicaTR discussion in Chapter 3). Besides the numeral dataset, we generated a small partitive dataset to investigate whether the models we developed can spot the differences between numeral and partitive constructions and appropriately predict the irregular entailment-implicature pattern. For this task, we use Model-1, whose test dataset is the numerals, which has an accuracy score of 60%.

Table 39 below shows the performance of Model-1 on the partitive dataset per each label class along with the total accuracy score below as 38%. We see that the inference types that partitives created are not predicted correctly by the model, which was trained on linguistic categories other than numerals. The model overpredicted implicatures, which is why the accuracy is high at 87%, although precision is low.

Table 39. Accuracy Scores of Partitives by Label (Predicted by Model-1)

| Label | N | Accuracy |
|---|---|---|
| implicature | 90 | 87.78% |
| neutral | 810 | 34.94% |
| entailment | 90 | 22.22% |
| Weighted Average | 990 | 38.59% |

For further examination, Table 40 below gives the accuracy score by each sentence along with the ground labels for the partitive dataset and with the regular labels, which is the ground label of that particular sentence for regular numerals (please refer to Chapter 3 for examples for each sentence).

Table 40. Accuracy Scores of Partitives by Sentence (Predicted by Model-1)

| Sentence Number | Ground Label | Regular Label | N | Accuracy |
|---|---|---|---|---|
| 1 | neutral | implicature | 90 | 1.11% |
| 2 | neutral | implicature | 90 | 1.11% |
| 3 | neutral | entailment | 90 | 22.22% |
| 4 | entailment | entailment | 90 | 35.56% |
| 5 | neutral | neutral | 90 | 87.78% |
| 6 | neutral | neutral | 90 | 77.78% |
| 7 | neutral | contradiction | 90 | 22.22% |
| 8 | neutral | contradiction | 90 | 95.56% |
| 9 | neutral | implicature | 90 | 78.89% |
| 10 | neutral | implicature | 90 | 1.11% |
| 11 | implicature | implicature | 90 | 1.11% |

As seen, sentences 5 and 6 are predicted correctly, as expected. However, the model does not cover sentence 11, which is an existential implicature. What is interesting is the case of sentences 8 and 9, where the model identifies that the partitive construction removes the regular inference pattern so that it correctly predicts sentences as neutral. This might signify that models resolve the specificity of the partitive NPs, though this will require more data to verify.

### 5.5.4 The case of C implicatures

The last small experiment is on the C implicatures, which are created by setting the sentence C from each set as the premise and with three different hypotheses specifically created. We saw that numerals give way to extra implicatures when the weaker term (smaller number of the scale) is negated, hence the sentence C. For instance, when the sentence *Not 30 people watched the movie* (which belongs to the scale of <thirty, sixty>) is uttered, three extra implicatures arise: the implicature that less than 30 people watched it, the implicature that at most 29 people watched it, and the implicature that at least 1 person watched it. We tested if these extra implicature inferences were learned as well during the training as well. We saw that the accuracy score of Model-1 on the C implicatures is %47, which seems to be good at first sight. We further looked into each of the 3 extra C implicature's respective accuracies, which is given in Table 41 below.

Table 41. Accuracy Scores of C Implicatures per Sentence

| Sentence Number | N | Accuracy |
|---|---|---|
| C Implicature 1 | 450 | 19.78% |
| C Implicature 2 | 450 | 24.89% |
| C Implicature 3 | 450 | 97.33% |

We see that the model accuracy for sentence 2 is equal to random selection accuracy, while it can even be considered biased against sentence 1. These show that the model did not generalize the implicature reasoning it gained from other categories to the numerals here. The case of sentence 3 is interesting because we named the type of implicature in sentence 3 as existential implicature rather than a scalar one. This is because the hypothesis has the logical meaning of *some* (at least one person did it) while the premise is not a maximal value like *all*; rather, the

premise denotes the negation of any cardinal number. This can be considered evidence that the model can understand and use the denotation of some without in different contexts.

## 5.6 Discussion

In Experiment 2, we further analyzed the LLMs' reasoning capabilities over scalar inferences and pragmatics. For this, we employed an experiment design, which we can call an ablation study, where we eliminate some part of the total dataset to see its effects on the model. We left out one of the linguistic categories wholly at each step and tested the model on the dataset that was left out. What this study showed is that the models can generalize the scalar reasoning that they learn to other linguistic categories. While the least accurate category in this sense is numerals, they are still predicted correctly with 60%. Quantifiers and modals showed a greater performance, whereas adjectives and verbs were mediocre in this regard. These results are partially in line with Baker et al. (2009), where adjectives and quantifiers were found to give rise to the implicature meaning, while numerals did not show implicature reasoning. As they argue, quantifiers have maximal denotations (like *all*) while others do not, which seems to affect the implicature reasoning. Additionally, adjectives and verbs do not denote exact values, and they are vague; numerals are argued to have an 'exactly' meaning, which can deteriorate the implicature reasoning. We also conferred that the low frequency of each word from adjectives and verbs was the reason why the models are not familiar with these words as much as they are with function words such as quantifiers. In Model-6, we tested if scalar inference can be generalized to other scalar pairs from the same linguistic category that the model previously learned in training. Results showed that when the model is exposed to the

scalar implicature data, it learns this inference and can generalize to other scalar pairs that were not explicitly shown.

In our analysis where we looked at the confusion matrices of given predictions by each model, we saw that contradiction and neutral classes are predicted highly correctly. This is in line with Gururangan et al. (2018) and Liu et al. (2022) in that the existence of words like 'not' in hypothesis creates a bias in NLI models towards contradiction. In addition, the superlative forms in hypotheses are also argued to direct the model towards predicting neutral. In our dataset, the hypotheses of neutral and contradiction sentences are B and C. Sentence B denotes the positive form of the stronger term, which is the maximal value such as *all*, while sentence C is the negative form of the weaker term such as *none,* which denotes the non-existence of the weaker term, hence a maximal value on the other end of the scale. Therefore, our results seem to confirm this bias of NLI models.

We further conducted feature analysis to reveal what linguistic features influenced the performance of the models developed. For this, we developed two models: logistic regression and random forest models. First, the logistic regression model showed us that the similarity score between the premise and hypothesis decreases the performance of the model. This score was computed by comparing the word embeddings of premise and hypothesis sentences. This was shown by the random forest model as well. Random forest also showed a greater effect on the inaccuracy of model predictions by within-sentence word similarity. If the words are semantically similar to each other, this hindered the model's entailment and implicature reasoning. These results are in line with what Kalouli et al. (2022) found. If the similarity score between premise and hypothesis is high, then models are generally distracted and show worse performance.

On the other hand, we also collected TF-IDF scores of each scalar item within premises and hypotheses. Both logistic regression and random forest showed that higher TF-IDF scores of scalar items are a factor that decreases the accuracy of the model. A somewhat similar score was obtained by Baker et al. (2009). They found that if the non-maximal scalar items (e.g., *some*) is a salient word, then the implicature reasoning deteriorates. We think that our results can be considered similar to this in that if the TF-IDF score of the scalar item increases, then the reasoning capabilities of the model decrease.

Another result that the feature analysis showed is that the position of the scalar item might be important in scalar inferencing. We found that when the scalar item is close to the predicate word, the model has difficulty in resolving the relationship there. Given that Turkish is an SOV language where predicate is regularly at the end and ImplicaTR does not have any irregular sentence, being close to predicate or being in the verb phrase increases the model's inferencing capabilities.

In the last step, we also conducted four small experiments by using the manipulations in ImplicaTR data sentences. In the first one, we saw that the aspect of the verb does not have an influence on scalar implicature inferencing. While all verb aspects have very high accuracies, we saw that presumptive future has only 0.1 less accuracy score than others. In the second one, we investigated if certain quantifier words in Turkish have an effect on the model capability. We again see that there are no significant differences between quantificational scalar items. Only the word *bazı* performed slightly worse than the others. In the third of these small experiments, we investigated if the implicature reasoning is generalized to partitive constructions, which substantially differ from regular numerals in entailment relationships. We saw

that there is more than a random chance that models can identify that the partitive constructions are not in similar entailment patterns, though the final accuracy results are low. And lastly, we inquired about the C implicatures that ImplicaTR has. We again see that the model cannot generalize to these implicatures while it can diagnose the existential implicature that the last sentence has. These results were preliminary insights on the topic and will require further data and experiments for clarification.

# CHAPTER 6

# CONCLUSION

In this study, we investigated scalar implicatures, which are a type of generalized conversational implicatures within the Gricean framework of implicatures. Our research was on NLI; we investigated how LLMs perform on scalar implicatures with an aim to reveal the pragmatic capabilities of LLMs.

For experiments, an NLI dataset in Turkish, ImplicaTR, was generated. This dataset contains the conventional entailment-neutral-contradiction classes and adds one more class, implicatures, on top of it. The dataset was generated semi-synthetically by creating the set of sentences first. This set has a structure in it: sentence A is the sentence with the weaker term of the scale, sentence B is the sentence with the stronger term of the scale, sentence C is the negation of sentence A, and sentence D is the negation of sentence B. Thus, sentence A signifies that at least the minimal value is true (some and possibly all), sentence B that the maximal value is true (all), sentence C that even the minimal value is not true (not some ~ none), and sentence D that the maximal value is not true but the minimal value might be true (not all). We created the ImplicaTR by combining these sentence pairs.

Our study is significant in many ways. First, ImplicaTR fills the gap for an NLI dataset that involves scalar implicatures, which is very rare in the literature. It also has linguistic categories created specifically, which allows for finer linguistic inquiry. Second, NLI datasets in Turkish are rare and mostly of poor quality due to translation work. Being a dataset originally created in Turkish, ImplicaTR serves as a great tool in Turkish NLI research. Third, the number of studies working on LLMs' scalar implicature capabilities is quite small, and the existing ones use linguistically

poor data. In our research, we developed several models that can do a more granular classification of entailment relationships even though their accuracy is relatively lower. Fourth, we have not encountered any study that does in-depth linguistic inquiry with feature analysis. The current study is also significant in the sense that it compiles NLP and linguistics insights on the issue.

In the literature, the derivation of scalar implicatures has given way to the emergence of new lines of work within the Gricean framework. There are authors supporting the idea that the derivation is fully pragmatic, whereas it has also been argued that it is lexical, and each lexeme also encodes this scalar relationship. While there are also hybrid perspectives of these two lines, it is also possible to encounter fully grammatical theories where the derivation is argued to be fully syntactic. According to our results, a hybrid view between pragmatic and lexical theories seems plausible. As the second experiment shows, the scalar reasoning learned from one linguistic category can be transferred to a new one. Furthermore, any reasoning learned for one specific linguistic category can be transferred to other scalar pairs in the same category. These are signs that there is a pragmatic reasoning going in the background which affects the general pragmatic capabilities of the model across all categories. Most the of scalar items are not exact in their meanings (like adjectives or verbs). We see that supervised learning can teach the model the meaning of the lexeme to a higher extent so that the model infers the scalar implicatures out of that lexeme. This also indicates the lexical aspect of the scalar implicature derivation.

The foremost result of our study is probably the fact that the LLMs can learn scalar implicatures. We saw that, after they are trained on general entailment inferences (i.e., BERT-NLI), they are easily adapting to the scalar type of implicatures.

Without a doubt, our study was not free of limitations. The dataset developed is not a small one, though it can be considered not a large one when it comes to adding a new class to the classification task. For this task, a dataset with many more rows of data would yield more concrete results. On the other hand, the genre and style of the sentences are not versatile, which might hinder the generalization capabilities of the models developed on the dataset. We consider that five linguistic categories are an acceptable variability, but including more categories, such as adverbs, might result in a more comprehensive dataset. In terms of our experiments, we think that the models used in Experiment 1 can be diversified to see the performance and reasoning capabilities of other transformers-based models and architectures. This also applies to the generative models, of which a new one is released almost daily. The linguistic analysis might be limited in the sense that it might explore more linguistic phenomena occurring within each sentence.

Consequently, we believe that this study enlightens the path of scalar implicature reasoning of LLMs both for linguistics and NLI, Turkish and in general. Further studies should explore the above-mentioned limitations to verify the results found here, especially focusing on more versatile datasets. One study can involve contexts for the premise-hypothesis pairs. Thus, the pragmatic abilities of the model can be analyzed better. Besides, the case of embedded speech acts for implicature reasoning is not covered in our dataset and might be of great importance for subsequent investigation. In addition, in Experiment 2, we discussed four small experiments. Further studies can concentrate on these preliminary examinations and augment their scope.

# APPENDIX A

# SCALES BY LINGUISTIC CATEGORY

## Adjectival Scalar Pairs

| Scale | Example Premise | Example Hypothesis |
|---|---|---|
| benzer-aynı | Arabalarımız birbiriyle benzer. | Arabalarımız birbiriyle aynı değil. |
| *similar-same* | *Our cars are similar to each other.* | *Our cars are not the same.* |
| benzer-aynı | Arabalarımız birbiriyle aynı değil. | Arabalarımız birbiriyle benzer. |
| *similar-same* | *Our cars are not the same.* | *Our cars are similar to each other.* |
| yeterli-tam | Ülkedeki ağaç miktarı yeterli. | Ülkedeki ağaç miktarı tam değil. |
| *sufficient-complete* | *The amount of trees in the country is sufficient.* | *The amount of trees in the country is not complete.* |
| yeterli-tam | Ülkedeki ağaç miktarı tam değil. | Ülkedeki ağaç miktarı yeterli. |
| *sufficient-complete* | *The amount of trees in the country is not complete.* | *The amount of trees in the country is sufficient.* |
| zararlı-ölümcül | Bu alandaki hava sizin için zararlı. | Bu alandaki hava sizin için ölümcül değil. |
| *harmful-lethal* | *The air in this area is harmful to you.* | *The air in this area is not lethal to you.* |
| zararlı-ölümcül | Bu alandaki hava sizin için ölümcül değil. | Bu alandaki hava sizin için zararlı. |
| *harmful-lethal* | *The air in this area is not lethal to you.* | *The air in this area is harmful to you.* |
| vasat-kötü | Sahne gösterisi vasattı. | Sahne gösterisi kötü değildi. |
| *mediocre-poor* | *The stage performance was mediocre.* | *The stage performance was not poor.* |
| vasat-kötü | Sahne gösterisi kötü değildi. | Sahne gösterisi vasattı. |
| *mediocre-poor* | *The stage performance was not poor.* | *The stage performance was mediocre.* |
| yakın-bitişik | Eski ve yeni park birbirine yakın. | Eski ve yeni park birbirine bitişik değil. |
| *close-adjacent* | *The old and new parks are close to each other.* | *The old and new parks are not adjacent to each other.* |
| yakın-bitişik | Eski ve yeni park birbirine bitişik değil. | Eski ve yeni park birbirine yakın. |
| *close-adjacent* | *The old and new parks are not adjacent to each other.* | *The old and new parks are close to each other.* |
| makul-harikulade | Yapılan anlaşma bizim için makul. | Yapılan anlaşma bizim için harikulade değil. |
| *reasonable-marvelous* | *The agreement made is reasonable for us.* | *The agreement made is not marvelous for us.* |
| makul-harikulade | Yapılan anlaşma bizim için harikulade değil. | Yapılan anlaşma bizim için makul. |
| *reasonable-marvelous* | *The agreement made is not marvelous for us.* | *The agreement made is reasonable for us.* |
| uygun fiyatlı-ucuz | Yeni elbiseler uygun fiyatlı. | Yeni elbiseler ucuz değil. |
| *affordable-cheap* | *The new clothes are affordable.* | *The new clothes are not cheap.* |
| uygun fiyatlı-ucuz | Yeni elbiseler ucuz değil. | Yeni elbiseler uygun fiyatlı. |
| *affordable-cheap* | *The new clothes are not cheap.* | *The new clothes are affordable.* |
| talihsiz-vahim | Bu ateşkes anlaşması talihsiz. | Bu ateşkes anlaşması vahim değil. |
| *unfortunate-disastrous* | *This ceasefire agreement is unfortunate.* | *This ceasefire agreement is not disastrous.* |
| talihsiz-vahim | Bu ateşkes anlaşması vahim değil. | Bu ateşkes anlaşması talihsiz. |
| *unfortunate-disastrous* | *This ceasefire agreement is not disastrous.* | *This ceasefire agreement is unfortunate.* |
| yanıltıcı-yalan | Otelimizin afişleri yanıltıcı. | Otelimizin afişleri yalan değil. |
| *misleading-false* | *The posters of our hotel are misleading.* | *The posters of our hotel are not false.* |
| yanıltıcı-yalan | Otelimizin afişleri yalan değil. | Otelimizin afişleri yanıltıcı. |
| *misleading-false* | *The posters of our hotel are not false.* | *The posters of our hotel are misleading.* |
| samimiyetsiz-düşmanca | Hastane personeli samimiyetsiz. | Hastane personeli düşmanca değil. |
| *insincere-hostile* | *The hospital staff is insincere.* | *The hospital staff is not hostile.* |
| samimiyetsiz-düşmanca | Hastane personeli düşmanca değil. | Hastane personeli samimiyetsiz. |
| *insincere-hostile* | *The hospital staff is not hostile.* | *The hospital staff is insincere.* |
| iyi-mükemmel | Tarih bilgin iyi. | Tarih bilgin mükemmel değil. |
| *good-excellent* | *Your knowledge of history is good.* | *Your knowledge of history is not excellent.* |
| iyi-mükemmel | Tarih bilgin mükemmel değil. | Tarih bilgin iyi. |
| *good-excellent* | *Your knowledge of history is not excellent.* | *Your knowledge of history is good.* |
| ilgili-takıntılı | Hızlı arabalara ilgili. | Hızlı arabalara takıntılı değil. |
| *interested-obsessed* | *He is interested in fast cars.* | *He is not obsessed with fast cars.* |
| ilgili-takıntılı | Hızlı arabalara takıntılı değil. | Hızlı arabalara ilgili. |
| *interested-obsessed* | *He is not obsessed with fast cars.* | *He is interested in fast cars.* |
| vasat-iyi | Yemeklerin kalitesi vasat. | Yemeklerin kalitesi iyi değil. |
| *average-good* | *The quality of the food is average.* | *The quality of the food is not good.* |
| vasat-iyi | Yemeklerin kalitesi iyi değil. | Yemeklerin kalitesi vasat. |
| *average-good* | *The quality of the food is not good.* | *The quality of the food is average.* |
| muhtemel-kesin | Yeni bir araba almamız muhtemel. | Yeni bir araba almamız kesin değil. |
| *likely-certain* | *It is likely for us to buy a new car.* | *It is not certain for us to buy a new car.* |
| muhtemel-kesin | Yeni bir araba almamız kesin değil. | Yeni bir araba almamız muhtemel. |
| *likely-certain* | *It is not certain for us to buy a new car.* | *It is likely for us to buy a new car.* |
| idare eder-harika | Dış ticaret verileri idare eder. | Dış ticaret verileri harika değil. |
| *acceptable-great* | *The foreign trade data is acceptable.* | *The foreign trade data is not great.* |
| idare eder-harika | Dış ticaret verileri harika değil. | Dış ticaret verileri idare eder. |
| *acceptable-great* | *The foreign trade data is not great.* | *The foreign trade data is acceptable.* |

# Verbal Scalar Pairs

| Scale | Example Premise | Example Hypothesis |
|---|---|---|
| yönelmek-varmak | Aslan yavaşça avına yöneldi. | Aslan yavaşça avına varmadı. |
| *approach-arrive* | *The lion slowly headed towards its prey.* | *The lion did not arrive at its prey slowly.* |
| yönelmek-varmak | Aslan yavaşça avına varmadı. | Aslan yavaşça avına yöneldi. |
| *approach-arrive* | *The lion did not arrive at its prey slowly.* | *The lion slowly headed towards its prey.* |
| denemek-başarmak | Hayvanları korumayı denedik. | Hayvanları korumayı başaramadık. |
| *attempt-succeed* | *We attempted to protect the animals.* | *We did not succeed in protecting the animals.* |
| denemek-başarmak | Hayvanları korumayı başaramadık. | Hayvanları korumayı denedik. |
| *attempt-succeed* | *We did not succeed in protecting the animals.* | *We attempted to protect the animals.* |
| öğrenmek-ustalaşmak | Japon mutfağını öğrendim. | Japon mutfağında ustalaşmadım. |
| *learn-master* | *I learned Japanese cuisine.* | *I did not master Japanese cuisine.* |
| öğrenmek-ustalaşmak | Japon mutfağında ustalaşmadım. | Japon mutfağını öğrendim. |
| *learn-master* | *I did not master Japanese cuisine.* | *I learned Japanese cuisine.* |
| algılamak-kavramak | Bizden istenilenleri algıladık. | Bizden istenilenleri kavramadık. |
| *perceive-grasp* | *We perceived what was asked of us.* | *We did not grasp what was asked of us.* |
| algılamak-kavramak | Bizden istenilenleri kavramadık. | Bizden istenilenleri algıladık. |
| *perceive-grasp* | *We did not grasp what was asked of us.* | *We perceived what was asked of us.* |
| başlamak-bitirmek | Yeni bir dil öğrenmeye başladı. | Yeni bir dil öğrenmeyi bitirmedi. |
| *start-finish* | *He started learning a new language.* | *He did not finish learning a new language.* |
| başlamak-bitirmek | Yeni bir dil öğrenmeyi bitirmedi. | Yeni bir dil öğrenmeye başladı. |
| *start-finish* | *He did not finish learning a new language.* | *He started learning a new language.* |
| katılmak-benimsemek | Toplantıda dile getirilen fikirlere katıldım. | Toplantıda dile getirilen fikirleri benimsemedim. |
| *agree-adopt* | *I agree with the ideas expressed in the meeting.* | *I did not adopt the ideas expressed in the meeting.* |
| katılmak-benimsemek | Toplantıda dile getirilen fikirleri benimsemedim. | Toplantıda dile getirilen fikirlere katıldım. |
| *agree-adopt* | *I did not adopt the ideas expressed in the meeting.* | *I agree with the ideas expressed in the meeting.* |
| seslenmek-bağırmak | Kardeşlerine seslenmiş. | Kardeşlerine bağırmamış. |
| *call out-yell* | *He called out to his brothers.* | *He did not yell at his brothers.* |
| seslenmek-bağırmak | Kardeşlerine bağırmamış. | Kardeşlerine seslenmiş. |
| *call out-yell* | *He did not yell at his brothers.* | *He called out to his brothers.* |
| gülmek-kahkaha atmak | Mübalağamıza gülündü. | Mübalağamıza kahkaha atılmadı. |
| *laugh-laugh out loud* | *(They) laughed at our exaggeration.* | *(They) laughed out loud at our exaggeration.* |
| gülmek-kahkaha atmak | Mübalağamıza kahkaha atılmadı. | Mübalağamıza gülündü. |
| *laugh-laugh out loud* | *(They) laughed out loud at our exaggeration.* | *(They) laughed at our exaggeration.* |
| serinlemek-üşümek | Ormana çıkınca serinlediler. | Ormana çıkınca üşümediler. |
| *cool-shiver* | *They cooled down when they went into the forest.* | *They did not shiver when they went into the forest.* |
| serinlemek-üşümek | Ormana çıkınca üşümediler. | Ormana çıkınca serinlediler. |
| *cool-shiver* | *They did not shiver when they went into the forest.* | *They cooled down when they went into the forest.* |

## Quantificational Scalar Pairs

| Scale | Example Premise | Example Hypothesis |
|---|---|---|
| birkaç-bütün | Birkaç tostu ben yedim. | Bütün tostları ben yemedim. |
| *a few-whole* | *I ate a few of the sandwiches.* | *I did not eat all of the sandwiches.* |
| birkaç-bütün | Bütün tostları ben yemedim. | Birkaç tostu ben yedim. |
| *a few-whole* | *I did not eat all of the sandwiches.* | *I ate a few of the sandwiches.* |
| birkaç-tüm | Birkaç ada turizme açıldı. | Tüm adalar turizme açılmadı. |
| *a few-all* | *Several islands have opened to tourism.* | *Not all islands have opened to tourism.* |
| birkaç-tüm | Tüm adalar turizme açılmadı. | Birkaç ada turizme açıldı. |
| *a few-all* | *Not all islands have opened to tourism.* | *Several islands have opened to tourism.* |
| birkaç-hepsi | Zenginlerin birkaçı fakirleşti. | Zenginlerin hepsi fakirleşmedi. |
| *a few-every* | *Some of the rich became poor.* | *Not all of the rich became poor.* |
| birkaç-hepsi | Zenginlerin hepsi fakirleşmedi. | Zenginlerin birkaçı fakirleşti. |
| *a few-every* | *Not all of the rich became poor.* | *Some of the rich became poor.* |
| bazı-bütün | Çocukların bazısı uyandı. | Bütün çocuklar uyanmadı. |
| *some-whole* | *Some of the children woke up.* | *Not all of the children woke up.* |
| bazı-bütün | Bütün çocuklar uyanmadı. | Çocukların bazısı uyandı. |
| *some-whole* | *Not all of the children woke up.* | *Some of the children woke up.* |
| bazı-tüm | Bazı trafik ışıkları bozuk. | Tüm trafik ışıkları bozuk değil. |
| *some-all* | *Some traffic lights are broken.* | *Not all traffic lights are broken.* |
| bazı-tüm | *Tüm trafik ışıkları bozuk değil.* | *Bazı trafik ışıkları bozuk.* |
| *some-all* | *Not all traffic lights are broken.* | *Some traffic lights are broken.* |
| bazı-hepsi | Köpeklerin bazıları havladı. | Köpeklerin hepsi havlamadı. |
| *some-every* | *Some of the dogs barked.* | *Not all of the dogs barked.* |
| bazı-hepsi | *Köpeklerin hepsi havlamadı.* | *Köpeklerin bazıları havladı.* |
| *some-every* | *Not all of the dogs barked.* | *Some of the dogs barked.* |
| biraz-çok | Kitabın birazını bitireceğim. | Kitabın çoğunu bitirmeyeceğim. |
| *a little-much* | *I will finish some of the book.* | *I will not finish most of the book.* |
| biraz-çok | Kitabın çoğunu bitirmeyeceğim. | Kitabın birazını bitireceğim. |
| *a little-much* | *I will not finish most of the book.* | *I will finish some of the book.* |
| birkaç-çok | Köpeklerin birkaçı hasta. | Köpeklerin çoğu hasta değil. |
| *a few-most* | *Some of the dogs are sick.* | *Most of the dogs are not sick.* |
| birkaç-çok | *Köpeklerin çoğu hasta değil.* | *Köpeklerin birkaçı hasta.* |
| *a few-most* | *Most of the dogs are not sick.* | *Some of the dogs are sick.* |
| bazı-çok | Bazı tabloların rengi solmuş. | Çoğu tablonun rengi solmamış. |
| *some-most* | *Some of the paintings have faded colors.* | *Most of the paintings have not faded colors.* |
| bazı-çok | Çoğu tablonun rengi solmamış. | Bazı tabloların rengi solmuş. |
| *some-most* | *Most of the paintings have not faded colors.* | *Some of the paintings have faded colors.* |

## Modal Scalar Pairs

| Scale | Example Premise | Example Hypothesis |
|---|---|---|
| muhtemelen-kesin | Bu şarkıyı muhtemelen dinlerim. | Bu şarkıyı dinleyeceğim kesin değil. |
| *probably-certain* | *I will probably listen to this song.* | *It's not certain that I will listen to this song.* |
| muhtemelen-kesin | Bu şarkıyı dinleyeceğim kesin değil. | Bu şarkıyı muhtemelen dinlerim. |
| *probably-certain* | *It's not certain that I will listen to this song.* | *I will probably listen to this song.* |
| herhalde-yüzde yüz | Herhalde dörtte uyanırlarmış. | Dörtte uyanacakları yüzde yüz değilmiş. |
| *possibly-one hundred percent (certain)* | *They probably wake up at four.* | *It's not one hundred percent certain that they will wake up at four.* |
| herhalde-yüzde yüz | Herhalde dörtte uyanırlarmış. | Dörtte uyanacakları yüzde yüz değilmiş. |
| *possibly-one hundred percent (certain)* | *It's not one hundred percent certain that they will wake up at four.* | *They probably wake up at four.* |

# Numeral and Partitive Scalar Pair

| Scale | Example Premise | Example Hypothesis |
|---|---|---|
| iki-üç | Şehirde iki hastane var. | Şehirde üç hastane yok. |
| *two-three* | *There are two hospitals in the city.* | *There are not three hospitals in the city.* |
| iki-üç | Şehirde ikiden fazla hastane yok. | Şehirde iki hastane var. |
| *two-three* | *There are not more than two hospitals in the city.* | *There are two hospitals in the city.* |
| üç-beş | Köprüden dün beş araç geçmiş. | Köprüden dün yedi araç geçmemiş. |
| *three-five* | *Five vehicles passed over the bridge yesterday.* | *Seven vehicles did not pass over the bridge yesterday.* |
| üç-beş | Köprüden dün beşten fazla araç geçmemiş. | Köprüden dün beş araç geçmiş. |
| *three-five* | *It was not more than five vehicles that passed over the bridge yesterday.* | *Five vehicles passed over the bridge yesterday.* |
| on-on iki | Sinemada on tane film var. | Sinemada on iki film yok. |
| *ten-twelve* | *There are ten films in the cinema.* | *There are not twelve films in the cinema.* |
| on-on iki | Sinemada ondan fazla film yok. | Sinemada on tane film var. |
| *ten-twelve* | *There are not more than ten films in the cinema.* | *There are ten films in the cinema.* |
| on beş-yirmi | Mutfakta on beş tabak kırdım. | Mutfakta yirmi tabak kırmadım. |
| *fifteen-twenty* | *I broke fifteen plates in the kitchen.* | *I did not break twenty plates in the kitchen.* |
| on beş-yirmi | Mutfakta on beşten fazla tabak kırmadım. | Mutfakta on beş tabak kırdım. |
| *fifteen-twenty* | *It was not more than fifteen plates that I broke in the kitchen.* | *I broke fifteen plates in the kitchen.* |
| on yedi-yirmi dokuz | On yedi defa teşekkür etti. | Yirmi dokuz defa teşekkür etmedi. |
| *seventeen-twenty-nine* | *She thanked seventeen times.* | *She did not thank twenty-nine times.* |
| on yedi-yirmi dokuz | On yedi defadan fazla teşekkür etmedi. | On yedi defa teşekkür etti. |
| *seventeen-twenty-nine* | *It was not more than seventeen times that she thanked.* | *She thanked seventeen times.* |
| yirmi dört-otuz altı | Yirmi dört bayrak var. | Otuz altı bayrak yok. |
| *twenty-four-thirty-six* | *There are twenty-four flags.* | *There are not thirty-six flags.* |
| yirmi dört-otuz altı | Yirmi dört bayraktan fazlası yok. | Yirmi dört bayrak var. |
| *twenty-four-thirty-six* | *There are not more than twenty-four flags.* | *There are twenty-four flags.* |
| yirmi beş-kırk | Yirmi beş ülkeyi ziyaret ettik. | Kırk ülkeyi ziyaret etmedik. |
| *twenty-five-forty* | *We visited twenty-five countries.* | *We did not visit forty countries.* |
| yirmi beş-kırk | Yirmi beşten ülkeden fazlasını ziyaret etmedik. | Yirmi beş ülkeyi ziyaret ettik. |
| *twenty-five-forty* | *It is not more than twenty-five countries that we did not visit.* | *We visited twenty-five countries.* |
| otuz-altmış | Otuz kitabı okudum. | Altmış kitabı okumadım. |
| *thirty-sixty* | *I read thirty books.* | *I did not read sixty books.* |
| otuz-altmış | Otuz kitaptan fazlasını okumadım. | Otuz kitaptan fazlasını okumadım. |
| *thirty-sixty* | *It is not more than thirty books that I did not read.* | *I read thirty books.* |
| elli-yetmiş | Doğruyu elli kişi söyledi. | Doğruyu yetmiş kişi söylemedi. |
| *fifty-seventy* | *Fifty people told the truth.* | *Seventy people did not tell the truth.* |
| elli-yetmiş | Doğruyu elli kişiden fazlası söylemedi. | Doğruyu elli kişi söyledi. |
| *fifty-seventy* | *It was not more than fifty people who told the truth.* | *Fifty people told the truth.* |

## Numeral Labels with Example Sentences

| Premise Type | Hypothesis Type | Premise Example | Hypothesis Example | Inference Type/Label |
|---|---|---|---|---|
| A | D | Bu filmi otuz kişi izledi.<br>*Thirty people watched this movie.* | Bu filmi altmış kişi izlemedi.<br>*Sixty people did not watch this movie.* | implicature |
| D2 | A | Bu filmi otuz kişiden fazlası izlemedi.<br>*No more than thirty people watched this movie.* | Bu filmi otuz kişi izledi.<br>*Thirty people watched this movie.* | implicature |
| C | D | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | Bu filmi altmış kişi izlemedi.<br>Sixty people did not watch this movie. | entailment |
| B | A | Bu filmi altmış kişi izledi.<br>Sixty people watched this movie. | Bu filmi otuz kişi izledi.<br>Thirty people watched this movie. | entailment |
| D | C | Bu filmi altmış kişi izlemedi.<br>Sixty people did not watch this movie. | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | neutral |
| A | B | Bu filmi otuz kişi izledi.<br>Thirty people watched this movie. | Bu filmi altmış kişi izledi.<br>Sixty people watched this movie. | neutral |
| B | C | Bu filmi altmış kişi izledi.<br>Sixty people watched this movie. | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | contradiction |
| C | B | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | Bu filmi altmış kişi izledi.<br>Sixty people watched this movie. | contradiction |
| C | C_implicature_1 | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | Bu filmi otuzdan az kişi izledi.<br>Less than thirty people watched this movie. | implicature |
| C | C_implicature_2 | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | Bu filmi en fazla yirmi dokuz kişi izledi.<br>At most twenty-nine people watched this movie. | implicature |
| C | C_implicature_3 | Bu filmi otuz kişi izlemedi.<br>Thirty people did not watch this movie. | Bu filmi en az bir tane kişi izledi.<br>At least one person watched this movie. | implicature |

## Partitive Labels with Example Sentences

| Premise Type | Hypothesis Type | Premise Example | Hypothesis Example | Inference Type/Label |
|---|---|---|---|---|
| A | D | Ağaçların beşini ben diktim.<br>*I planted five of the trees.* | Ağaçların yedisini ben dikmedim.<br>*I didn't plant seven of the trees.* | neutral |
| D2 | A | Ağaçların beşinden fazlasını ben dikmedim.<br>*I didn't plant more than five of the trees.* | Ağaçların beşini ben diktim.<br>*I planted five of the trees.* | neutral |
| C | D | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | Ağaçların yedisini ben dikmedim.<br>*I didn't plant seven of the trees.* | neutral |
| B | A | Ağaçların yedisini ben diktim.<br>*I planted seven of the trees.* | Ağaçların beşini ben diktim.<br>*I planted five of the trees.* | entailment |
| D | C | Ağaçların yedisini ben dikmedim.<br>*I didn't plant seven of the trees.* | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | neutral |
| A | B | Ağaçların beşini ben diktim.<br>*I planted five of the trees.* | Ağaçların yedisini ben diktim.<br>*I planted seven of the trees.* | neutral |
| B | C | Ağaçların yedisini ben diktim.<br>*I planted seven of the trees.* | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | neutral |
| C | B | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | Ağaçların yedisini ben diktim.<br>*I planted seven of the trees.* | neutral |
| C | C_implicature_1 | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | Ağaçların beşinden azını ben diktim.<br>*I planted less than five of the trees.* | neutral |
| C | C_implicature_2 | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | Ağaçların en fazla dördünü ben diktim.<br>*I planted at most four of the trees.* | neutral |
| C | C_implicature_3 | Ağaçların beşini ben dikmedim.<br>*I didn't plant five of the trees.* | Ağaçların en az birini ben diktim.<br>*I planted at least one of the trees.* | implicature |

# APPENDIX B

## SAMPLE LLAMA-2 COMPLETIONS

A: Son beş yılda çıkan kitaplarınız vasat.
B: Son beş yılda çıkan kitaplarınız kötü değil.
Ground Label: implicature
Predicted Label: implicature
Response: b does not entail a, and b implies that the quality of the recently released books is not good.

A: Kitabımın tüm bölümleri beğenilmedi.
B: Kitabımın birkaç bölümü beğenildi.
Ground Label: implicature
Predicted Label: contradiction
Response: b is in contrast with a, indicating that some of the chapters of the book were read.

A: Ürettiği ürünü nasıl pazarlayacağı konusunu öğrendi.
B: Ürettiği ürünü nasıl pazarlayacağı konusunda ustalaştı.
Ground Label: neutral
Predicted Label: neutral
Response: b is neutral in relation to a. b does not provide any new information that a does not already convey.

A: Pusulanın gösterdiği konuma yöneldi.
B: Pusulanın gösterdiği konuma vardı.
Ground Label: neutral
Predicted Label: entailment
Response: both b and a describe the same relationship.

A: Yorgun şairin evine yönelinmedi.
B: Yorgun şairin evine varılmadı.
Ground Label: entailment
Predicted Label: entailment
Response: b describes a more specific situation than a, as it implies that the poet did not even start walking towards the poet's house.

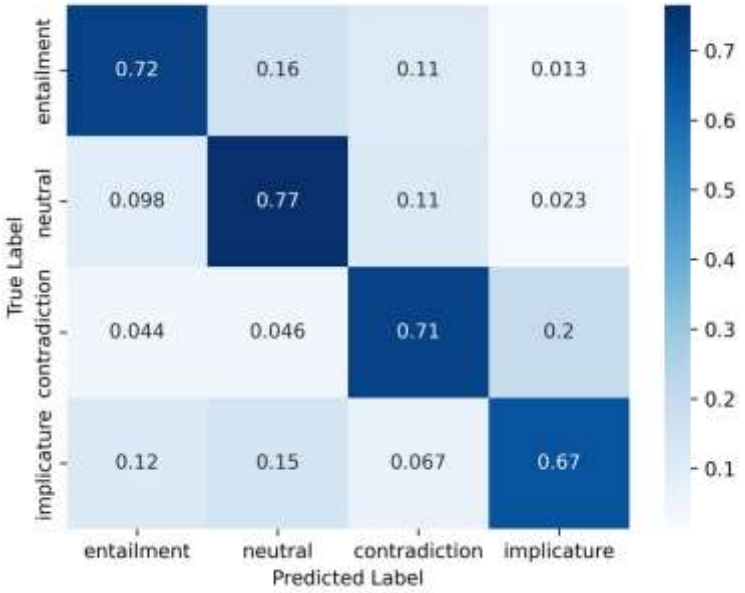
A: Ehliyetini alanların sürüşleri kötü değil.
B: Ehliyetini alanların sürüşleri vasat.
Ground Label: implicature
Predicted Label: implicature
Response: b implies that the runners' performance is good, which is different from a's statement that it is not.

A: Kaynak ustası bu işte on yıldır çalışmıyor.
B: Kaynak ustası bu işte on iki yıldır çalışmıyor.
Ground Label: entailment
Predicted Label: neutral
Response: b is neutral because it does not provide any additional information about the duration of kaynak's work.

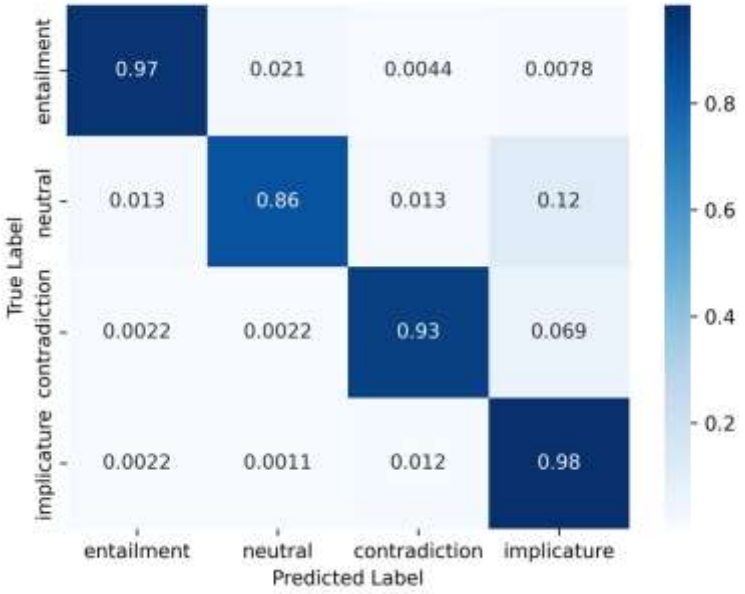
APPENDIX C

CONFUSION MATRICES

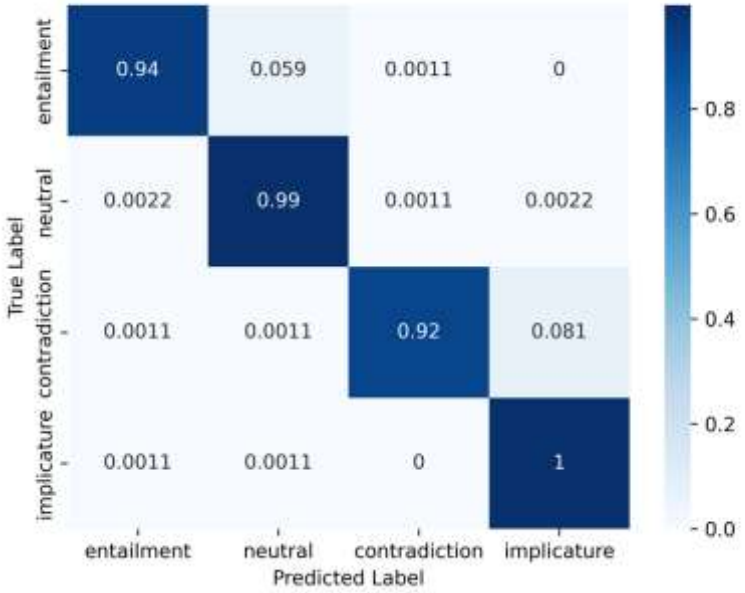
Confusion Matrix of Adjectives



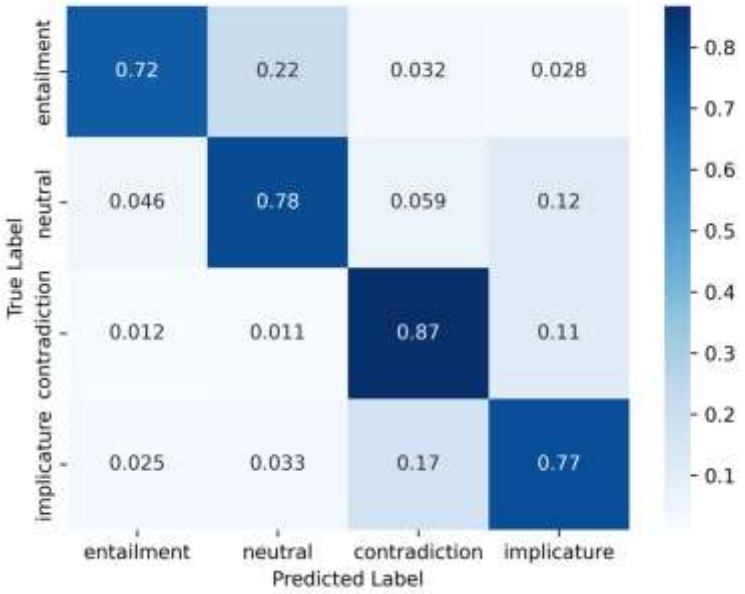

Confusion Matrix of Quantifiers



Confusion Matrix of Modals




Confusion Matrix of the Neutral-2 Sentence



APPENDIX D

FEATURE ANALYSES

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|---|---|---|---|---|
| (1-2) | -0.082 | 0.000 | -0.106 | -0.058 |
| (1-3) | -0.218 | 0.000 | -0.242 | -0.194 |
| (1-4) | -0.247 | 0.000 | -0.271 | -0.223 |
| (1-5) | 0.112 | 0.000 | 0.088 | 0.136 |
| (2-1) | 0.082 | 0.000 | 0.058 | 0.106 |
| (2-3) | -0.137 | 0.000 | -0.161 | -0.113 |
| (2-4) | -0.165 | 0.000 | -0.189 | -0.141 |
| (2-5) | 0.193 | 0.000 | 0.169 | 0.217 |
| (3-1) | 0.218 | 0.000 | 0.194 | 0.242 |
| (3-2) | 0.137 | 0.000 | 0.113 | 0.161 |
| (3-4) | -0.029 | 0.011 | -0.053 | -0.004 |
| (3-5) | 0.330 | 0.000 | 0.306 | 0.354 |
| (4-1) | 0.247 | 0.000 | 0.223 | 0.271 |
| (4-2) | 0.165 | 0.000 | 0.141 | 0.189 |
| (4-3) | 0.029 | 0.011 | 0.004 | 0.053 |
| (4-5) | 0.359 | 0.000 | 0.334 | 0.383 |
| (5-1) | -0.112 | 0.000 | -0.136 | -0.088 |
| (5-2) | -0.193 | 0.000 | -0.217 | -0.169 |
| (5-3) | -0.330 | 0.000 | -0.354 | -0.306 |
| (5-4) | -0.359 | 0.000 | -0.383 | -0.334 |

Logistic Regression Coefficients and Frequencies for NER tags

| Feature | Coefficient | Frequency |
|---|---|---|
| PERCENT | 1.467630 | 0.043722 |
| FAC | 1.044898 | 0.002333 |
| MONEY | 0.933265 | 0.000444 |
| EVENT | 0.794782 | 0.001556 |
| LANGUAGE | 0.685991 | 0.004222 |
| PERSON | 0.602571 | 0.019944 |
| ORG | 0.599811 | 0.001111 |
| NORP | 0.565061 | 0.006222 |
| DATE | 0.433454 | 0.005222 |
| WORK_OF_ART | 0.300296 | 0.004833 |
| GPE | 0.119597 | 0.026556 |
| LOC | 0.099830 | 0.000444 |
| TITLE | -0.146653 | 0.001222 |
| QUANTITY | -0.556223 | 0.003667 |
| TIME | -0.826073 | 0.015000 |
| ORDINAL | -0.922748 | 0.001167 |
| CARDINAL | -1.153119 | 0.148278 |

Features Used in Random Forest Model

| Feature Name |
|---|
| premise_scale_position |
| premise_words_on_right |
| premise_words_on_left |
| hypothesis_words_on_right |
| hypothesis_words_on_left |
| word_similarity_diff |
| hypothesis_token_counts |
| premise_word_similarity |
| hypothesis_word_similarity |
| premise_token_counts |
| hypothesis_scale_tfidf |
| premise_scale_tfidf |
| premise_hypothesis_similarity |

Logistic Regression Model with Coefficients

| Feature | Coefficient |
|---|---|
| premise_scale_position | 1.246053 |
| premise_words_on_right | 1.014938 |
| premise_words_on_left | 0.757901 |
| hypothesis_words_on_right | 0.362788 |
| hypothesis_words_on_left | 0.339886 |
| word_similarity_diff | 0.270488 |
| hypothesis_is_root_verb | 0.223625 |
| avg_premise_token_length | 0.208467 |
| hypothesis_scale_position | 0.190735 |
| premise_is_root_verb | 0.161609 |
| isPol_NegNeg | 0.157276 |
| isPol_PosNeg | 0.147530 |
| isPol_NegPos | 0.086528 |
| avg_hypothesis_token_length | 0.059677 |
| premise_polarity | -0.082447 |
| hypothesis_noun_counts | -0.112200 |
| hypothesis_polarity | -0.143448 |
| premise_noun_counts | -0.174325 |
| isPol_PosPos | -0.229977 |
| hypothesis_token_counts | -0.238279 |
| premise_word_similarity | -0.347963 |
| hypothesis_word_similarity | -0.451775 |
| premise_token_counts | -0.657990 |
| hypothesis_scale_tfidf | -1.450011 |
| premise_scale_tfidf | -2.401104 |
| premise_hypothesis_similarity | -4.236277 |

# REFERENCES


Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*.

Altinok, D. (2023). A Diverse Set of Freely Available Linguistic Resources for Turkish. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13739–13750. https://doi.org/10.18653/v1/2023.acl-long.768

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., … Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 929–947. https://doi.org/10.1145/3620665.3640366

Arslan, S. (2020). When the owner of information is unsure: Epistemic uncertainty influences evidentiality processing in Turkish. *Lingua*, *247*, 102989. https://doi.org/10.1016/j.lingua.2020.102989

Baker, R., Doran, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, *1*(2), 211–248. https://doi.org/10.1163/187730909X12538045489854

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information* (arXiv:1607.04606). arXiv. http://arxiv.org/abs/1607.04606

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. https://doi.org/10.18653/v1/D15-1075

Breheny, R. (2019). Scalar Implicatures. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics* (1st ed., pp. 39–61). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198791768.013.4

Budur, E., Özçelik, R., Gungor, T., & Potts, C. (2020). Data and Representation for Turkish Natural Language Inference. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8253–8267. https://doi.org/10.18653/v1/2020.emnlp-main.662

Capone, A. (2006). On Grice's circle (a theory-internal problem in linguistic theories of the Gricean type). *Journal of Pragmatics*, *38*(5), 645–669. https://doi.org/10.1016/j.pragma.2006.02.005

Carston, R. (2004). Truth-conditional content and conversational implicature. In C. Bianchi (Ed.), *The Semantics/Pragmatics Distinction* (pp. 65--100). CSLI Publications.

Chierchia, G. (2004). Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface. In A. Belletti (Ed.), *Structures and Beyond* (pp. 39–103). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195171976.003.0003

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., Meroni, L., & di Milano-Bicocca, U. (2001). *The Acquisition of Disjunction: Evidence for a Grammatical View of Scalar Implicatures*.

Chierchia, G., Fox, D., & Spector, B. (2012). 87. Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. V. Heusinger, & P. Portner (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science* (pp. 2297–2332). DE GRUYTER. https://doi.org/10.1515/9783110253382.2297

Cho, W. I., Chersoni, E., Hsu, Y.-Y., & Huang, C.-R. (2021). Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2922–2929. https://doi.org/10.18653/v1/2021.findings-acl.258

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). *XNLI: Evaluating Cross-lingual Sentence Representations* (arXiv:1809.05053). arXiv. http://arxiv.org/abs/1809.05053

Cui, R., Hershcovich, D., & Søgaard, A. (2022). *Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks* (arXiv:2204.10615). arXiv. http://arxiv.org/abs/2204.10615

Culpeper, J., & Haugh, M. (2014). *Pragmatics and the English Language*.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché-Buc (Eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (Vol. 3944, pp. 177–190). Springer Berlin Heidelberg. https://doi.org/10.1007/11736790_9

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02151-0

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists* (1st ed). Routledge.

De Haan, F. (2006). Typological approaches to modality. In W. Frawley, E. Eschenroede, S. Mills, & T. Nguyen (Eds.), *The Expression of Modality* (pp. 27–70). Mouton de Gruyter. https://doi.org/10.1515/9783110197570.27

De Melo, G., & Bansal, M. (2013). Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, *1*, 279–290. https://doi.org/10.1162/tacl_a_00227

De Neys, W., & Schaeken, W. (2007). When People Are More Logical Under Cognitive Load: Dual Task Impact on Scalar Implicature. *Experimental Psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

Deng, J., Cheng, J., Sun, H., Zhang, Z., & Huang, M. (2023). *Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements* (arXiv:2302.09270). arXiv. http://arxiv.org/abs/2302.09270

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Enç, M. (1991). The Semantics of Specificity. *Linguistic Inquiry*, *22*(1), 1–25. JSTOR.

Fox, D. (2007). Free Choice and the Theory of Scalar Implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics* (pp. 71–120). Palgrave Macmillan UK. https://doi.org/10.1057/9780230210752_4

Fu, Z., Lam, W., Yu, Q., So, A. M.-C., Hu, S., Liu, Z., & Collier, N. (2023). *Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder* (arXiv:2304.04052). arXiv. http://arxiv.org/abs/2304.04052

Gazdar, G. (1983). *Pragmatics: Implicature, presupposition, and logical form* (2. [print.]). Acad. Pr.

Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., … Kenealy, K. (2024). *Gemma: Open Models Based on Gemini Research and Technology* (arXiv:2403.08295). arXiv. http://arxiv.org/abs/2403.08295

George, E. J., & Mamidi, R. (2020). Conversational implicatures in English dialogue: Annotated dataset. *Procedia Computer Science*, *171*, 2316–2323. https://doi.org/10.1016/j.procs.2020.04.251

Geurts, B. (2009). Scalar Implicature and Local Pragmatics. *Mind & Language*, *24*(1), 51–79. https://doi.org/10.1111/j.1468-0017.2008.01353.x

Glickman, O., & Dagan, I. (2005). A probabilistic setting and lexical cooccurrence model for textual entailment. *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment - EMSEE '05*, 43–48. https://doi.org/10.3115/1631862.1631870

Grice, H. P. (1975). Logic and Conversation. In D. Davidson & G. Harman (Eds.), *The Logic of Grammar* (pp. 64–75).

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Gubelmann, R., Katis, I., Niklaus, C., & Handschuh, S. (2023). Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks. *Journal of Logic, Language and Information*, *33*(1), 21–48. https://doi.org/10.1007/s10849-023-09410-4

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. https://doi.org/10.18653/v1/N18-2017

Hao, Y., Dong, L., Wei, F., & Xu, K. (2019). *Visualizing and Understanding the Effectiveness of BERT* (arXiv:1908.05620). arXiv. http://arxiv.org/abs/1908.05620

He, P., Gao, J., & Chen, W. (2023). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing* (arXiv:2111.09543). arXiv. http://arxiv.org/abs/2111.09543

Hirschberg, J. B. (1985). *A Theory of Scalar Implicature* [University of Pennsylvania]. https://api.semanticscholar.org/CorpusID:122032940

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. University of California.

Horn, L. R. (1985). Metalinguistic Negation and Pragmatic Ambiguity. *Language*, *61*(1), 121–174. JSTOR. https://doi.org/10.2307/413423

Horn, L. R. (2006a). Implicature. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (1st ed., pp. 2–28). Wiley. https://doi.org/10.1002/9780470756959.ch1

Horn, L. R. (2006b). Implicature. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (1st ed.). Wiley. https://doi.org/10.1002/0470018860.s00233

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. http://arxiv.org/abs/2106.09685

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R., & Jin, Z. (2023). State-of-the-art generalisation research in NLP: A taxonomy and review. *Nature Machine Intelligence*, *5*(10), 1161–1174. https://doi.org/10.1038/s42256-023-00729-y

Hurford, J. R. (1974). Exclusive or Inclusive Disjunction. *Foundations of Language*, *11*(3), 409–411. JSTOR.

Jackendoff, R. (1996). The proper treatment of measuring out, telicity, and perhaps even quantification in english. *Natural Language and Linguistic Theory*, *14*(2), 305–354. https://doi.org/10.1007/BF00133686

Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8690–8705. https://doi.org/10.18653/v1/2020.acl-main.768

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv. http://arxiv.org/abs/2310.06825

Jiang, N., & De Marneffe, M.-C. (2019). Evaluating BERT for natural language inference: A case study on the CommitmentBank. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6085–6090. https://doi.org/10.18653/v1/D19-1630

Johanson, L. (2009). 15. Modals in Turkic. In B. Hansen & F. D. Haan (Eds.), *Modals in the Languages of Europe* (pp. 487–510). Mouton de Gruyter. https://doi.org/10.1515/9783110219210.3.487

Jumelet, J., Denić, M., Szymanik, J., Hupkes, D., & Steinert-Threlkeld, S. (2021). *Language Models Use Monotonicity to Assess NPI Licensing* (arXiv:2105.13818). arXiv. http://arxiv.org/abs/2105.13818

Kalouli, A.-L., Sevastjanova, R., Beck, C., & Romero, M. (2022). *Negation, Coordination, and Quantifiers in Contextualized Language Models* (arXiv:2209.07836). arXiv. http://arxiv.org/abs/2209.07836

Katsos, N. (2005). *Interaction of Structural and Contextual Constraints During the On-line Generation of Scalar Inferences*.

Katsos, N., & Cummins, C. (2010). Pragmatics: From Theory to Experiment and Back Again. *Language and Linguistics Compass*, *4*(5), 282–295. https://doi.org/10.1111/j.1749-818X.2010.00203.x

Katsos, N., & Cummins, C. (2012). *Scalar implicature: Theory, processing and acquisition*.

Kaufmann, S., Condoravdi, C., & Harizanov, V. (2006). Formal approaches to modality. In W. Frawley, E. Eschenroede, S. Mills, & T. Nguyen (Eds.), *The Expression of Modality* (pp. 71–106). Mouton de Gruyter. https://doi.org/10.1515/9783110197570.71

Kennedy, C. (1999). *GRADABLE ADJECTIVES DENOTE MEASURE FUNCTIONS, NOT PARTIAL FUNCTIONS*.

Kennedy, C. (2007). Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Kennedy, C., & McNally, L. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, *81*(2), 345–381. https://doi.org/10.1353/lan.2005.0071

Kerimoğlu, C. (2010). On The Epistemic Modality Markers in Turkey Turkish: Uncertainty. *Journal of Turkish Studies*, *Volume 5 Issue 4*(5), 434–478. https://doi.org/10.7827/TurkishStudies.1836

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). *"Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures* (arXiv:2305.13826). arXiv. https://doi.org/10.48550/arXiv.2305.13826

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. http://arxiv.org/abs/1412.6980

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners* (arXiv:2205.11916). arXiv. http://arxiv.org/abs/2205.11916

Korman, D. Z., Mack, E., Jett, J., & Renear, A. H. (2018). Defining textual entailment. *Journal of the Association for Information Science and Technology*, *69*(6), 763–772. https://doi.org/10.1002/asi.24007

Kriz, V., Holub, M., & Pecina, P. (2015). *Feature Extraction for Native Language Identification Using Language Modeling*.

Kumar, S., & Talukdar, P. (2020). *NILE: Natural Language Inference with Faithful Natural Language Explanations* (arXiv:2005.12116). arXiv. http://arxiv.org/abs/2005.12116

Landman, F. (2000). *Events and Plurality* (Vol. 76). Springer Netherlands. https://doi.org/10.1007/978-94-011-4359-2

Lassiter, D. (2010). *Gradable epistemic modals, probability, and scale structure*.

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, *32*(1), 84–100. https://doi.org/10.1017/pan.2023.20

Lee, C. (2008). *Scalar Implicatures: Pragmatic Inferences or Grammar?*

Leech, G. N., Rayson, P., & Wilson, A. (2002). *Word frequencies in written and spoken English: Based on the British National Corpus* (Nachdr.). Longman.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press. https://doi.org/10.7551/mitpress/5526.001.0001

Lim, H. (2021). A Study on Dropout Techniques to Reduce Overfitting in Deep Neural Networks. In J. J. Park, V. Loia, Y. Pan, & Y. Sung (Eds.), *Advanced Multimedia and Ubiquitous Engineering* (Vol. 716, pp. 133–139). Springer Singapore. https://doi.org/10.1007/978-981-15-9309-3_20

Liu, L., Jindal, I., & Li, Y. (2022). *Is Semantic-aware BERT more Linguistically Aware? A Case Study on Natural Language Inference*.

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/BM.2013.018

Miaschi, A., Brunato, D., Dell'Orletta, F., & Venturi, G. (2020). Linguistic Profiling of a Neural Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*, 745–756. https://doi.org/10.18653/v1/2020.coling-main.65

Millar, N. (2009). Modal verbs in TIME: Frequency changes 1923–2006. *International Journal of Corpus Linguistics*, *14*(2), 191–220. https://doi.org/10.1075/ijcl.14.2.03mil

Naik, A., Ravichander, A., Sadeh, N., Rose, C., & Neubig, G. (2018). *Stress Test Evaluation for Natural Language Inference*.

Newstead, S. E., & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 535–546. https://doi.org/10.1016/S0022-5371(83)90328-6

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Ortiz Suárez, P. J., Sagot, B., & Romary, L. (2019). *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures* [Application/pdf]. 337 KB. https://doi.org/10.14618/IDS-PUB-9021

Özyıldız, D. (2017). Quantifiers in Turkish. In D. Paperno & E. L. Keenan (Eds.), *Handbook of Quantifiers in Natural Language: Volume II* (Vol. 97, pp. 857–937). Springer International Publishing. https://doi.org/10.1007/978-3-319-44330-0_17

Panizza, D., Chierchia, G., & Clifton, C. (2009). On the role of entailment patterns and scalar implicatures in the processing of numerals. *Journal of Memory and Language*, *61*(4), 503–518. https://doi.org/10.1016/j.jml.2009.07.005

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282. https://doi.org/10.1016/S0010-0277(02)00179-8

Papafragou, A., & Schwarz, N. (2005). Most Wanted. *Language Acquisition*, *13*(3), 207–251. JSTOR.

Pedersen, W. A. (2014). *Inchoative verbs and adverbial modification: Decompositional and scalar approaches*. McGill University.

Poliak, A. (2020). *A Survey on Recognizing Textual Entailment as an NLP Evaluation* (arXiv:2010.03061). arXiv. http://arxiv.org/abs/2010.03061

Potts, C. (2015). Presupposition and Implicature. In S. Lappin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (1st ed., pp. 168–202). Wiley. https://doi.org/10.1002/9781118882139.ch6

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. http://arxiv.org/abs/1910.10683

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002

Römer, U. (2004). *Textbooks:* A corpus-driven approach to modal auxiliaries and their didactics. In J. McH. Sinclair (Ed.), *Studies in Corpus Linguistics* (Vol. 12, pp. 185–199). John Benjamins Publishing Company. https://doi.org/10.1075/scl.12.14rom

Rotstein, C., & Winter, Y. (2004). Total Adjectives vs. Partial Adjectives: Scale Structure and Higher-Order Modifiers. *Natural Language Semantics*, *12*(3), 259–288. https://doi.org/10.1023/B:NALS.0000034517.56898.9a

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey Of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, *65*, 569–631. https://doi.org/10.1613/jair.1.11640

Russell, B. (2006). Against Grammatical Computation of Scalar Implicatures. *Journal of Semantics*, *23*(4), 361–382. https://doi.org/10.1093/jos/ffl008

Sauerland, U. (2012). The Computation of Scalar Implicatures: Pragmatic, Lexical or Grammatical? *Language and Linguistics Compass*, *6*(1), 36–49. https://doi.org/10.1002/lnc3.321

Schumacker, R., & Tomek, S. (2013). Chi-Square Test. In R. Schumacker & S. Tomek, *Understanding Statistics Using R* (pp. 169–175). Springer New York. https://doi.org/10.1007/978-1-4614-6227-9_8

Schweter, S. (2020). *BERTurk—BERT models for Turkish* (1.0.0) [Computer software]. [object Object]. https://doi.org/10.5281/ZENODO.3770924

Sebüktekin, H. I. (1971). *Turkish-English contrastive analysis: Turkish morphology and corresponding English structures*. De Gruyter. https://doi.org/10.1515/9783111635781

Shetreet, E., Chierchia, G., & Gaab, N. (2014). When *some* is not *every*: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping*, *35*(4), 1503–1514. https://doi.org/10.1002/hbm.22269

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2. ed., Repr). Blackwell.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (arXiv:2206.04615). arXiv. http://arxiv.org/abs/2206.04615

Stacey, J., Belinkov, Y., & Rei, M. (2022). Supervising Model Attention with Human Explanations for Robust Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11349–11357. https://doi.org/10.1609/aaai.v36i10.21386

Talman, A., Apidianaki, M., Chatzikyriakidis, S., & Tiedemann, J. (2021). *NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance* (arXiv:2104.04751). arXiv. http://arxiv.org/abs/2104.04751

Thukral, S., Kukreja, K., & Kavouras, C. (2021). Probing Language Models for Understanding of Temporal Expressions. *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 396–406. https://doi.org/10.18653/v1/2021.blackboxnlp-1.31

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Tomioka, S. (2021). Scalar Implicature, Hurford's Constraint, Contrastiveness and How They All Come Together. *Frontiers in Communication*, *5*, 461553. https://doi.org/10.3389/fcomm.2020.461553

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. http://arxiv.org/abs/2307.09288

Van Rooij, R., & Schulz, K. (2004). Exhaustive Interpretation of Complex Sentences. *Journal of Logic, Language and Information*, *13*(4), 491–519. https://doi.org/10.1007/s10849-004-2118-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., & Lambert, N. (2020). *TRL: Transformer Reinforcement Learning* (0.2.1) [Python]. https://github.com/huggingface/trl (Original work published 2020)

Wendland, A., Zenere, M., & Niemann, J. (2021). Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique. In M. Yilmaz, P. Clarke, R. Messnarz, & M. Reiner (Eds.), *Systems, Software and Services Process Improvement* (Vol. 1442, pp. 289–300). Springer International Publishing. https://doi.org/10.1007/978-3-030-85521-5_19

Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. http://arxiv.org/abs/1910.03771

Wu, S., & Dredze, M. (2019). *Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT* (arXiv:1904.09077). arXiv. http://arxiv.org/abs/1904.09077

Yin, W., Hay, J., & Roth, D. (2019). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach* (arXiv:1909.00161). arXiv. http://arxiv.org/abs/1909.00161

Yu, F., Zhang, H., Tiwari, P., & Wang, B. (2023). *Natural Language Reasoning, A Survey* (arXiv:2303.14725). arXiv. http://arxiv.org/abs/2303.14725

Zheng, Z., Qiu, S., Fan, L., Zhu, Y., & Zhu, S.-C. (2021). GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2074–2085. https://doi.org/10.18653/v1/2021.findings-acl.182