Arabic Natural Language Inference with Pre-trained Transformers: Exploring the Power of Semantic Understanding

Fouzi Takelait
Kennedy College of Sciences
University of Massachusetts Lowell, USA
fouzi_takelait@student.uml.edu

Abstract

Natural Language Inference (NLI) is a crucial aspect of Natural Language Processing (NLP) that involves classifying the relationship between two sentences into categories such as entailment or contradiction. In this paper, we propose a novel approach for Arabic NLI utilizing pre-trained transformer models - Arabert-v2, Marbert , and Qarib. The existing methods mostly focus on syntactic, lexical, and semantic strategies, but these approaches lack the sophistication of understanding intricate language semantics. The novelty of our work lies in the ability of transformer models to capture both lexical and semantic features, thus potentially outperforming traditional NLI methods. We present a comprehensive study and comparison of different pre-trained models using the Arabic datasets ArNLi and ArbTEDS. Our results are compared with state-of-the-art methods to provide a new benchmark in the literature. Our work aims to contribute to the evolving landscape of Arabic NLI, providing advanced methods for further research and applications in the field. By improving the capability to infer relationships between sentences, we hope to refine machine understanding of Arabic language, paving the way for more sophisticated applications in areas like machine translation, question answering, and text summarization.

Keywords: Natural Language inferences , NLP, Transformers

1. Introduction

Natural Language Inference (NLI), also known as recognizing textual entailment (RTE), is a vital task in natural language processing (NLP) that involves determining the logical relationship between a pair of texts, referred to as the premise (P) and hypothesis (H). The task of NLI has gained significant attention among researchers, as it plays a crucial role in various NLP and artificial intelligence (AI) systems. NLI encompasses identifying the relationship, such as entailment, contradiction, or neutral, between the premise and hypothesis sentences. The NLI task has extensive applications in numerous domains, including dialogue improvement, question answering, machine translation, text classification, summarization, and information retrieval. By comprehending the meaning and context of user input or source language text, NLI facilitates the generation of more accurate and relevant responses, translations, categorizations, summaries, and information retrieval. Various approaches have been explored for the NLI task, such as rule-based methods, machine learning, and deep learning approaches. Neural networks (NN) have proven to be particularly effective in NLI tasks due to their ability to understand the complexities of natural language, which is essential for accurately determining the relationships between sentences. However, there is limited research on applying NN techniques and RNN architectures specifically to address the Arabic NLI problem, with most existing works relying on feature engineering and statistical methods. In this paper, we propose a novel approach that leverages pre-trained transformer models to fine-tune the NLI task on Arabic datasets. Transformers models, known for their ability to capture both lexical and semantic features, offer the potential to outperform traditional NLI methods. We aim to bridge the gap in Arabic NLI research by adopting a neural network approach and exploring the effectiveness of pre-trained transformer models for the Arabic NLI task. The main contributions of our work include:

- Introducing an NLI system for Arabic based on pre-trained transformer models, treating the inference task as a classification problem without the need for manual feature engineering.
- Evaluating different types of pre-trained transformer models and loss functions to enhance the robustness of our model for the Arabic NLI task.
- Evaluating These models on different Arabic NLI datasets, namely ArbTEDS, ArNLI, to comprehensively evaluate the performance of our proposed system.

To evaluate our approach, we compare the results with existing Arabic NLI benchmarks and translation of renowned English benchmarks due to the lack of benchmarks in Arabic. The remainder of this paper is organized as follows: Section 2 provides an overview of Related Work. In Section 3, we describe the Existing Arabic NLI datasets. Section 4 presents the methodology in detail, including the fine-tuning of transformer models . Section 5 discusses the experimental results, and finally, Section 6 concludes the paper and outlines future research directions.

dataset	Total of samples	Num of classes	Training Percentage	Validation Percentage	test Percentage
ArNli	5945	3	80%	10%	10%
ArbTEDS	600	2	70%	15%	15%

Table 1: Datasets Distribution

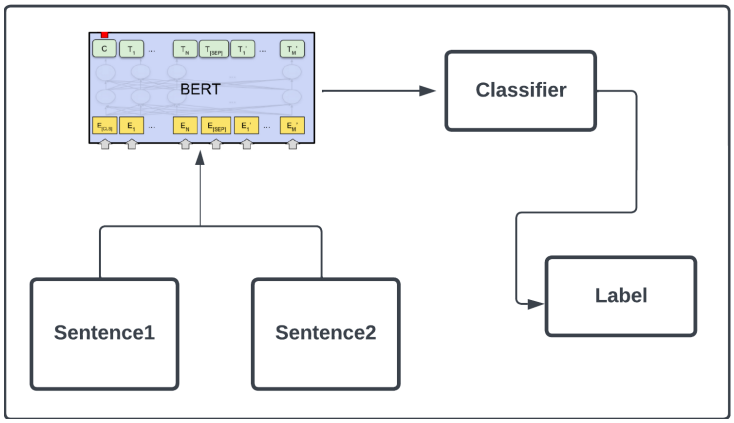

Figure 1: Proposed Architecture: Fine-Tuning on Pre-trained transformer models

2. Related Work

In recent years, there has been a growing interest in Natural Language Inference (NLI) research, particularly in the context of downstream NLP tasks (Mishra et al., 2020). Most of the existing studies in NLI have focused on two-way RTE, which involves binary classification of Entailment vs. Non-entailment scenarios. However, there has been relatively less research on three-way RTE, specifically emphasizing contradiction. The majority of advancements in NLI have been achieved for English, primarily due to the availability of large-scale datasets and the utilization of deep learning models such as BERTNLI, RoBERTa, XLNET, and DeBERTa, which are based on transformer architectures. However, when it comes to languages other than English, the progress in NLI has been limited due to the scarcity of reliable datasets. Various research efforts have been made to create NLI datasets for languages like Japanese (Mishra et al., 2020), Chinese (Hu et al., 2020), Portuguese (Rocha and Lopes Cardoso, 2018), Italian (Bos et al., 2009), German (Eichler et al., 2014), Brazilian (Fonseca et al., 2016), Persian (Amirkhani et al., 2020), and Turkish (Budur et al., 2020). In the case of the Arabic language, there is an existing Arabic dataset for two-way RTE, but it consists of only 600 pairs and does not provide enough data for deep learning approaches (Alabbas, 2011). Arabic NLI research faces several challenges stemming from the unique features of the Arabic language [(Almarwani and Diab, 2017) (Ben-Sghaier et al., 2020) (Khader et al., 2016). Lexical ambiguity is a significant challenge due to the absence of diacritics in Arabic texts, making it difficult to process them accurately. Additionally, Arabic exhibits rich synonymy, where multiple words may have the same surface meaning. Furthermore, Arabic lacks comprehensive computational resources like a large-scale WordNet, which are widely available and used for English. To address the scarcity of large-scale entailment datasets in Arabic, researchers have proposed various methods and systems. Alabbas developed ArbTE, a system that evaluates existing textual entailment techniques applied to the Arabic language (Khader et al., 2016). They extended the basic version of the Tree Edit Distance (TED) algorithm in (Alabbas, 2011) to enhance the matching algorithm for Arabic textual entailment. Alabbas also created the publicly available ArbTEDS14 dataset, consisting of 618 text-hypothesis pairs collected from Arabic news websites and hand-annotated pairs. AlKhawaldeh et al. explored the impact of resolving negation and analyzing the polarity of text-hypothesis pairs on Arabic entailment accuracy (Alabbas and Ramsay, 2013). They achieved an accuracy of 69% on the ArbTEDS dataset by considering polarity information. Almarwani and Diab applied Support Vector Machine (SVM) and Random Forest classifiers with word embeddings to detect textual entailment in Arabic (Al-Khawaldeh, 2019). They achieved an accuracy of 76.2% on the ArbTEDS dataset by incorporating various features such as similarity scores, named entities, and word overlap. Boudaa et al. used the Support Vector Machine (SVM) algorithm and alignment-based features to detect textual entailment in Arabic (Habash et al., 2017). Their system achieved an accuracy of 75.84% on the ArbTEDS dataset by leveraging features like named entities, tem-

poral expressions, and word sequences. Khader et al. proposed a lexical analysis technique for Arabic textual entailment, combining word overlap and semantic matching to enhance the precision of their system (Boudaa et al., 2019). They achieved a precision of 68% for entailment and 58% for non-entailment with an overall recall of 61% on the ArbTEDS dataset. in (Jallad and Ghneim, 2022) introduce an Arabic NLI system that utilizes a novel dataset called ArNLI, created specifically for this purpose. The authors employ a diverse set of features including named entity recognition, WordNet similarity, customized stopwords, as well as features related to numerical values, dates, and times. Furthermore, they experiment with various language models, such as TF-IDF, N-Grams, and word embeddings. Building upon the introduction of the ArNLI dataset, (Bensghaier et al., 2023) presents deep learning approaches that utilize various Recurrent Neural Networks (RNNs) to overcome the limitations of rule-based methods and traditional machine learning algorithms. These studies highlight the efforts made in addressing Arabic NLI challenges and developing systems and approaches specifically tailored for the Arabic language. However, there is still a need for further research and exploration to improve Arabic textual entailment accuracy and to bridge the gap between English-centric advancements and other languages.

3. Dataset

3.1. ArNLI

ArNLI is a newly created dataset specifically designed for Arabic NLI research. It is developed to provide a valuable resource for training and evaluating NLI systems in Arabic (Jallad and Ghneim, 2022).

The authors of the paper present ArNLI as a novel dataset that aims to address the lack of suitable data for Arabic NLI tasks. The dataset is constructed with careful consideration of various factors relevant to the Arabic language, ensuring its relevance and effectiveness in capturing the specific challenges and characteristics of Arabic text.

The ArNLI dataset is likely designed to cover a wide range of linguistic phenomena, including syntactic structures, semantic relationships, word usage, and other linguistic features. It may include a diverse set of sentence pairs, each consisting of a premise and a hypothesis, with various inference relationships such as entailment, contradiction, and neutrality. These pairs are carefully annotated by human annotators, providing gold-standard labels indicating the correct inference relationship between the premise and hypothesis.

To enhance the richness and coverage of the dataset, the authors may have incorporated various linguistic resources and features. These could include named entity recognition, WordNet similarity, special stopwords, and features related to numbers, dates, and times. Additionally, the authors may have utilized different language models such as TF-IDF, N-Grams, and word embeddings to augment the dataset with valuable linguistic information.

The ArNLI dataset is likely comprehensive in its coverage and diverse in its content, aiming to provide a challenging and realistic benchmark for evaluating Arabic NLI models. By introducing this dataset, the authors contribute to the research community by enabling further advancements in Arabic NLI and facilitating the development of more accurate and robust NLI systems for the Arabic language. The statistics of the dataset distributed as follows : **Training pairs 5092, Testing pairs 1274**.

3.2. ArbTEDS

The NLI task in Arabic saw extensive utilization of the Arabic Textual Entailment Dataset (ArbTEDS). (Maytham, 2013) Alabbas created ArbTEDS, the first dataset of its kind for Arabic NLI, employing a semi-automatic approach. It consists of approximately **600 pairs of text/hypothesis (T/H) statements**, evenly distributed between the "entails" and "not entails" classes. However, despite its widespread use, ArbTEDS is constrained by its relatively limited size. Furthermore, its annotation is limited to two classes, restricting its applicability to binary classification scenarios.

4. Methodology

In this section, we discuss the essential components of the proposed method, starting with an overview of the pre-trained models used and going over the fine-tuning process on the various datasets to overcome the dataset's small sample size problem.

4.1. Representation Learning

Representation learning plays a fundamental role in machine learning tasks, as it involves capturing and encoding the underlying features and patterns of input data. Traditionally, methods like TF-IDF and variant LSTM models have been used for representation learning, each with its limitations. TF-IDF relies on a simple bag-of-words approach, lacking the ability to capture complex structures and dependencies in the data. LSTM models, while effective in capturing sequential information, face challenges in modeling long-range dependencies and suffer from vanishing or exploding gradient problems. Transformers, on the other hand, have emerged as a powerful paradigm for representation learning, revolutionizing natural language processing tasks. The key strength of transformers lies in their attention mechanism, which enables them to capture dependencies between all input tokens simultaneously. This attention mechanism allows transformers to attend to relevant information across the entire input sequence, regardless of the distance between tokens. As a result, transformers excel at capturing complex structures and long-range dependencies, making them highly effective for a wide range of tasks, including language translation, sentiment analysis, and natural language inference. Furthermore,

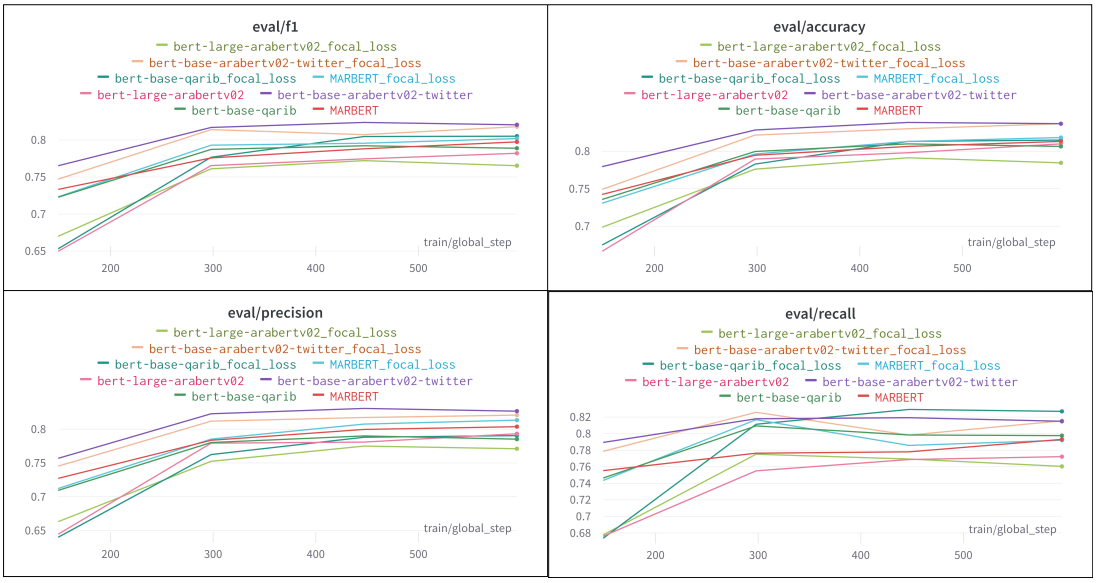

Figure 2: charts of Evaluation Performance on ArNLI Dataset.

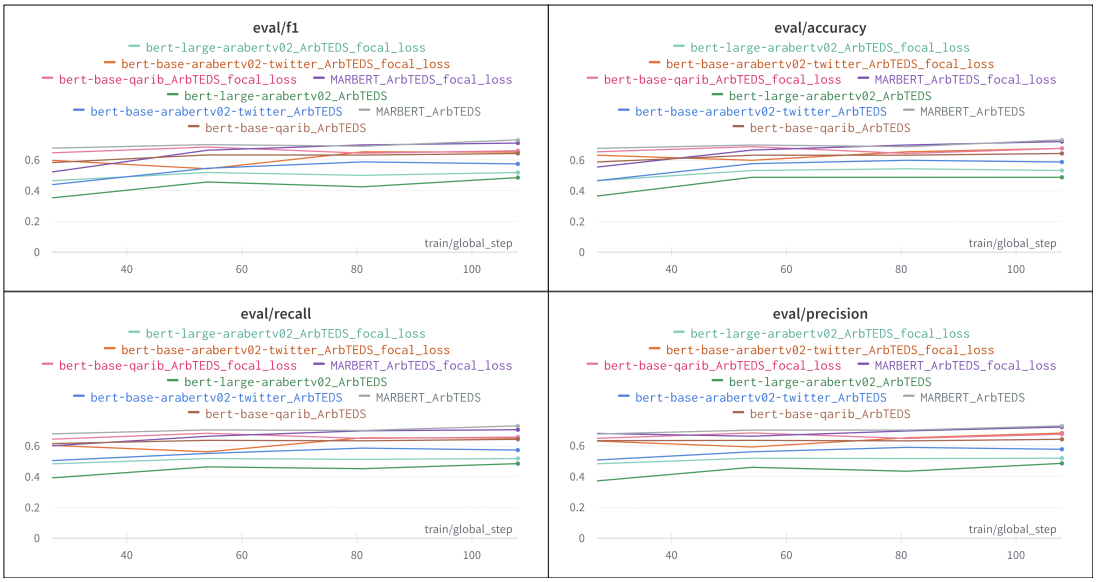

Figure 3: charts of Evaluation Performance on ArbTEDS Dataset.

transformers have gained attention for their ability to handle large-scale datasets and parallelize computations efficiently. This scalability makes them well-suited for training on massive amounts of data, which is crucial for capturing the diverse and nuanced nature of language.

4.2. Models

In our methodology, we build upon the insights gained from previous work on natural language inference (NLI) and address the inherent challenges associated with capturing both semantic and grammatical information. Prior approaches have often focused on either semantics or grammar, leading to limitations in accurately classifying the relationship between two sentences. To overcome these drawbacks, we leverage the power of pre-trained transformer models, specifically (Abdul-Mageed et al., 2020) MARBERT,(Abdelali et al., 2021) qarib, and two version from (Antoun et al., 2020) aubmindlab, By utilizing these pre-trained models, we benefit from their ability to effectively encode and represent complex linguistic patterns and relationships. These models have been trained on extensive amounts of text data, enabling them to learn rich contextual embeddings that capture the nuanced meaning and structure of Arabic sentences. Leveraging the strengths of these models, we aim to accurately classify the relationship between pairs of sentences in our NLI task. To evaluate the performance of our methodology, we conduct comprehensive

ArNLI Dataset								
Models	**Cross Entropy**				**Focal Loss**			
	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
MARBERT	79.1	77.2	77.7	76.8	81.0	79	79.1	79.0
bert-base-qarib	78.6	77.2	76.9	77.8	78.1	77.1	76.0	78.6
bert-base-arabertv02-twitter	80.8	79.5	80.5	**78.7**	81.0	79.4	80.4	78.5
bert-large-arabertv02	**82.5**	**80.4**	**82.9**	78.4	80.5	78.7	79.3	78.2

Table 2: The Results Of Applying Different Models on ArNLI Data-Set

ArbTEDS Dataset								
Models	**Cross Entropy**				**Focal Loss**			
	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
MARBERT	**76.6**	**76.1**	**76.5**	**76.0**	63.3	54.8	71.2	59.2
bert-base-qarib	62.2	62.2	63.0	63.0	57.7	57.6	59.2	59.0
bert-base-arabertv02-twitter	62.2	62.1	63.8	63.5	67.7	67.5	70.7	69.5
bert-large-arabertv02	46.6	43.2	50.0	50.0	51.1	48.5	56.0	54.3

Table 3: The Results Of Applying Different Models on ArbTEDS DataSet

experiments using different pre-trained models. This allows us to assess the effectiveness of each model in capturing the nuances of the Arabic language and accurately predicting the relationship between sentences. By comparing the results obtained from different models, we gain valuable insights into their performance characteristics and identify the most suitable model for our specific NLI task. Additionally, One of the addressed challenges in this study pertains to imbalanced datasets. In line with previous research (Jamal et al., 2022), we have undertaken a comparative analysis to evaluate the performance disparity of models utilizing two distinct loss functions namely cross entropy and focal loss.

5. Results and Discussion

5.1. Performance Metrics

When assessing the performance of the Natural Language Inference (NLI) task, one approach is to treat it as a classification problem, where the goal is to categorize the relationship between two sentences into a set number of classes, such as two, three, or even multiple classes. To measure the effectiveness of NLI models, we rely on widely adopted evaluation metrics, namely accuracy, precision, recall, and F1-score.

By considering NLI as a classification task, we aim to accurately predict the relationship between the premise and hypothesis sentences. Accuracy serves as a fundamental metric, representing the overall correctness of the model's predictions. It measures the proportion of correct classifications out of the total number of instances in the evaluation dataset.

In addition to accuracy, precision, recall, and F1-score provide further insights into the performance of NLI models. Precision focuses on the correctness of positive predictions made by the model, determining the proportion of true positive predictions out of all positive predictions. Recall, on the other hand, measures the model's ability to identify all positive instances correctly, calculating the proportion of true positives out of the total actual positive instances in the dataset. F1-score combines both precision and recall into a single metric, harmonizing their contributions to evaluate the model's overall effectiveness.

By employing these evaluation metrics, we can assess the model's capability to classify the relationships between sentences accurately. The accuracy metric ensures the overall correctness of the predictions, while precision, recall, and F1-score provide a deeper understanding of the model's performance in correctly identifying positive instances and avoiding false positives and false negatives.

5.2. Experimental Results

In this section, we present the outcomes of our extensive experimentation involving a diverse range of models and architectures trained on various datasets, highlighting the impact of these factors on performance on the test set only. Our analysis focused on several pre-trained models specifically designed for Arabic NLI, leading us to identify Arabertv2 large model and MARBERT as the top-performing models across the ArNLI and ArbTEDS datasets See Figure 2 and Figure 3 . These models were trained using two distinct loss functions: cross-entropy and focal loss.

Interestingly, our experimental findings demonstrate that the Cross Entropy function consistently outperformed the focal loss across all models. Contrary to the assumption presented in (Lin et al., 2017), the Cross Entropy function exhibited superior performance in handling imbalanced datasets.

When comparing our results to the benchmark performance, our models showcased remarkable improvements. Specifically, our model achieved an impressive f1-score of **80.4%** on the ArNLI dataset and **76.1%** on the ArbTEDS dataset. Furthermore, our model demonstrated a notable accuracy of **82.5%** on ArNLI and **76.6%** on ArbTEDS See Table 2 and Table 3. These results demonstrate the effectiveness of our approach in surpassing the benchmark performance on both datasets.

5.3. Results Analysis

In this section, we thoroughly examine the results obtained from our proposed methods. Specifically, on the ArNLI dataset, we observed that the large-sized Arabertv2 model consistently outperformed all other models, including the state-of-the-art MARBERT model, regardless of the chosen loss function. This notable performance can be attributed to the fact that the ArNLI dataset bears more similarities to the dataset on which Arabertv2 was trained.

Conversely, when evaluating the ArbTEDS dataset, we found that the MARBERT model achieved a significantly higher accuracy compared to all other models. Impressively, the MARBERT model attained an accuracy of 76.6% on the test set, surpassing the performance of alternative models.

Models	Accuracy(%)
Syntactic approaches	66,3%
Lexical approaches (Enriched representation of P and H)	75.84%
Traditional Word Embedding	76.20%
Earth Mover's Distance + word embedding	76.50%
Use different language models (TFIDF, N-Grams, and Word Embeddings)	75%
Training different types of recurrent neural network models with words embeddings	73,60%
Our Proposed Model	Accuracy - F1-score
MARBERT + Cross-Enrtopy Loss	76.6% - 76.1%

Table 4: Different Approaches With an Accuracy On ArbTEDS dataset.

Models	Accuracy(%)
Use different language models (TFIDF,N-Grams, and Word Embedding)	75%
Training different types of recurrent neural network modelswith words embeddings	63,65%
Our Proposed Model	Accuracy - F1-score
bert-large-arabertv02 + Cross-Enrtopy Loss	82.5% - 80.4%

Table 5: Different Approaches With an Accuracy On ArNLI Dataset.

6. Conclusion and Future Works

In conclusion, our proposed method has achieved state-of-the-art results on the evaluated datasets, surpassing the performance of all other existing methods. For detailed performance comparisons, please refer to the accompanying Table 2 and Table 3. Our method was developed specifically for dealing with NLI in Arabic language, leveraging the efficiency and performance of pre-trained transformer models to handle the semantic and syntactic complexities of Arabic text. Additionally, considering the imbalanced nature of the dataset, we conducted experiments using different loss functions to examine their impact on model performance. In future work, we intend to expand our experimentation to include more diverse and representative datasets. Moreover, increasing the size of the dataset holds the potential to enhance the model's robustness.

Source Code: To replicate our results, we have made our code open-source, allowing anyone to access and conduct the experiments. Additionally, researchers can utilize this code to explore further with additional pre-trained models.

GitHub: `https://github.com/ftakelait/ArabicNLI.git`

7. Bibliographical References

Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., and Samih, Y. (2021). Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2020). Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Al-Khawaldeh, F. T. (2019). A study of the effect of resolving negation and sentiment analysis in recognizing text entailment for arabic. *arXiv preprint arXiv:1907.03871*.

Alabbas, M. and Ramsay, A. (2013). Natural language inference for arabic using extended tree edit distance with subtrees. *Journal of Artificial Intelligence Research*, 48:1–22.

Alabbas, M. (2011). Arbte: Arabic textual entailment. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 48–53.

Almarwani, N. and Diab, M. (2017). Arabic textual entailment with word embeddings. pages 185–190.

Amirkhani, H., AzariJafari, M., Pourjafari, Z., Faridan-Jahromi, S., Kouhkan, Z., and Amirak, A. (2020). Farstail: A persian natural language inference dataset. *arXiv preprint arXiv:2009.08820*.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Ben-Sghaier, M., Bakari, W., and Neji, M. (2020). Classification and analysis of arabic natural language inference systems. *Procedia Computer Science*, 176:551–560.

Bensghaier, M., Bakari, W., and Neji, M. (2023). Investigating the use of different recurrent neural networks for natural language inference in arabic.

Bos, J., Zanzotto, F. M., and Pennacchiotti, M. (2009). Textual entailment at evalita 2009. *Proceedings of EVALITA*, 2009(6.4):2.

Boudaa, T., El Marouani, M., and Enneya, N. (2019). Alignment based approach for arabic textual entailment. *Procedia computer science*, 148:246–255.

Budur, E., Özçelik, R., Güngör, T., and Potts, C. (2020).

Data and representation for turkish natural language inference. *arXiv preprint arXiv:2004.14963*.

Eichler, K., Gabryszak, A., and Neumann, G. (2014). An analysis of textual inference in german customer emails. pages 69–74.

Fonseca, E., Borges, L., Santos, D., Criscuolo, M., and Aluısio, S. (2016). Assin: Evaluation of semantic similarity and textual inference.

Habash, N., Diab, M., Darwish, K., El-Hajj, W., Al-Khalifa, H., Bouamor, H., Tomeh, N., El-Haj, M., and Zaghouani, W. (2017). Proceedings of the third arabic natural language processing workshop.

Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., and Moss, L. S. (2020). Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Jallad, K. A. and Ghneim, N. (2022). Arnli: Arabic natural language inference for entailment and contradiction detection. *arXiv preprint arXiv:2209.13953*.

Jamal, S., Kassem, A. M., Mohamed, O., and Ashraf, A. (2022). On the arabic dialects' identification: Overcoming challenges of geographical similarities between arabic dialects and imbalanced datasets. pages 458–463.

Khader, M., Awajan, A., and Alkouz, A. (2016). Textual entailment for arabic language based on lexical and semantic matching. *International Journal of Computing & Information Sciences*, 12(1):67–74.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Maytham, A. (2013). A dataset for arabic textual entailment. In *Proceedings of the Student Research Workshop associated with RANLP 2013*.

Mishra, A., Patel, D., Vijayakumar, A., Li, X., Kapanipathi, P., and Talamadupula, K. (2020). Reading comprehension as natural language inference: a semantic analysis. *arXiv preprint arXiv:2010.01713*.

Rocha, G. and Lopes Cardoso, H. (2018). Recognizing textual entailment: challenges in the portuguese language. *Information*, 9(4):76.