Quantum Natural Language Processing for High-Level Sentence Classification

Dhruv Sachdeva
BTech Engineering Physics
Delhi Technological University
Delhi, India

Hussein Shiri
Physics Graduate - Faculty of Science
Lebanese University
Beirut, Lebanon

Kiran Kaur
BS Data Science and its Applications
Indian Institute of Technology Madras
Chennai, India

Maksym Husarov
Software Engineer
EPAM
Malaga, Spain

Rishi Koushik Reddy Thippireddy
BTech Data Science and AI
Indian Institute of Information Technology Dharwad
Dharwad, India

***Abstract*—Quantum Natural Language Processing takes Classical Natural Language Processing to a whole new level by harnessing the power and properties of quantum physics and quantum computing. This paper delves into the domain of Quantum Natural Language Processing (QNLP) and its application in high-level sentence classification tasks. We aims to explore its potential by employing both Classical and Quantum Pipelines. Leveraging the capabilities of quantum systems to achieve improved accuracy and efficiency in classifying sentences based on their semantic meaning.**

I. Introduction

In the realm of natural language processing (NLP), the boundaries of what was once thought possible are rapidly expanding as quantum physics and quantum computing converge to create a paradigm shift. Quantum Natural Language Processing (QNLP) represents a quantum leap beyond classical NLP, unveiling new horizons for understanding and processing language. This paper embarks on a journey into this domain, investigating the fusion of quantum principles with linguistic data to propel high-level sentence classification tasks to new heights.

Quantum Natural Language Processing takes Classical Natural Language Processing to a whole new level by harnessing the power and properties of quantum physics and quantum computing. It's a pioneering endeavor that strives to unravel the full potential of Quantum Natural Language Processing by blending both Classical and Quantum Pipelines. This integration leverages the intrinsic capabilities of quantum systems to enhance the accuracy and efficiency of sentence classification based on their semantic meaning. As the volume and complexity of textual data continue to escalate, the need for innovative and efficient Natural Language Processing solutions has never been more pronounced.

In this paper, we embark on a journey to explore the transformative power of Quantum Natural Language Processing. We will delve into the fundamental principles of quantum mechanics and how they can be harnessed to revolutionize the way we understand and classify sentences. Quantum computing, with its ability to operate in superposition and entanglement, holds the promise of addressing the computational challenges posed by large language datasets. The implications of this convergence extend far beyond the confines of academia, with potential applications in diverse fields such as information retrieval, sentiment analysis, and machine translation.

Our aim is to provide a comprehensive overview of the current state of Quantum Natural Language Processing research, showcasing recent breakthroughs, limitations, and opportunities for future exploration. By the end of this paper, readers will have a deeper understanding of the promise and challenges of Quantum Natural Language Processing, equipped with the knowledge to navigate this evolving landscape of advanced sentence classification. As the fusion of quantum physics and language processing continues to unfold, it is evident that Quantum Natural Language Processing is poised to unlock new levels in language understanding and classification, and this paper aims to guide the way.

II. Methodology

A. Data Collection and Preprocessing:

A heterogeneous selection of sentences originating from diverse domains was meticulously curated and subjected to rigorous preprocessing procedures to establish a standardized

and contextually relevant dataset. To ensure the utmost diversity and relevance, our data collection process involved the extraction of text from a variety of sources. Specifically, we employed web scraping techniques to acquire a comprehensive array of textual data that contained various reviews of various domains.

The objective was to compile a dataset that not only represented the broad expanse of linguistic diversity but also encapsulated real-world contexts and user-generated content, thus fostering the robustness of our research.In parallel with data acquisition, a rigorous preprocessing regimen was applied to ensure uniformity and reliability within the dataset. The purpose of these preprocessing steps was to enhance the quality of the dataset by reducing noise and harmonizing the textual data for subsequent analysis.

B. Classical Pipeline:

We established a classical Natural Language Processing (NLP) pipeline as the foundational benchmark for our study. This classical pipeline encompassed conventional NLP techniques, including tokenization, feature extraction, and the employment of traditional machine learning models for sentence classification.

The implementation was orchestrated within the Apache Spark framework, leveraging the powerful capabilities of Spark-NLP to ensure efficient data processing. This classical NLP pipeline served as a reference point to assess the performance of subsequent quantum-enhanced and hybrid pipelines.

C. Quantum Pipeline:

Leveraging quantum computing frameworks lambeq and qiskit, a Quantum NLP pipeline was developed. Quantum-enhanced techniques like Quantum Embedding and Quantum Amplitude Encoding were employed to encode sentence semantics into quantum states.

D. Hybrid Approach:

The hybrid approach combines classical and quantum elements to leverage the respective strengths of both paradigms in sentence classification. This hybrid methodology involves classical pre-processing and post-processing steps, such as data cleaning, tokenization, and result interpretation, seamlessly integrated with quantum computations to enhance the overall classification framework.

E. Performance Comparison:

The performance of the Classical and Quantum Pipelines will be thoroughly evaluated using metrics like accuracy and precision. A detailed analysis will be conducted to understand the scenarios where Quantum NLP excels.

III. Data Collection and Preprocessing

The initial phase of our data collection process involved the systematic extraction of sentences from various review sections on a wide array of websites, forming the foundation of our comprehensive dataset.

To accomplish this, we employed the BeautifulSoup library, a Python tool for web scraping that allowed us to gather diverse textual content efficiently and effectively. Subsequently, meticulous data cleaning procedures were carried out to enhance the quality and uniformity of our dataset.

Furthermore, we leveraged the TextBlob library to perform sentiment analysis, affording us insights into the emotional tone and polarity of the collected sentences. The results of this sentiment analysis were then meticulously extracted and recorded in a structured text file, setting the stage for the subsequent stages of our research analysis."

IV. Classical Pipeline

The classical pipeline employed a suite of well-established Natural Language Processing (NLP) techniques. This pipeline was structured into three integral phases, each contributing to the robustness of our research methodology.

The initial phase encompassed data extraction, a pivotal step involving the systematic amalgamation of data from multiple source files into a consolidated dataset. This consolidated dataset served as the foundational repository for our subsequent analysis and experimentation.

Following data extraction, the transformation phase came into play, where extensive data preprocessing techniques were applied. This involved the removal of duplicate sentences and the execution of sentiment analysis to discern the polarity of each sentence, thereby adding a critical dimension to our data. This transformation phase was imperative for ensuring the quality and relevance of the dataset used in our research.

In the final phase of our classical pipeline, the results were loaded into a Jupyter notebook, affording us a convenient platform for further analysis, visualization, and experimentation. Subsequently, to ensure the accessibility and transparency of our research process, the results were systematically archived in a GitHub repository, thus providing a publicly accessible record of our work.

To do all of this we used the spark-nlp, pyspark, glob and pandas libraries. The following is a figure of sample data after the classical pipeline.

label	text	classified_label
1	I am glad I took ...	1.0
1	I ordered May 31 .	0.0
1	Thank you .	1.0
1	Was unused at fir...	0.0
0	Very disappointed...	0.0
0	A bit ugly .	0.0
0	Manufacturing def...	0.0
0	Very disappointed...	1.0
0	The case is great...	1.0
0	Bad quality .	0.0
0	Overpriced .	0.0
1	I ordered the pho...	1.0
1	A very reasonable...	1.0
1	I ordered an phon...	1.0
1	I ordered an phon...	1.0
1	Good quality .	1.0
1	Fast, friendly, r...	1.0
1	Got this phone fo...	1.0
1	I ordered an iPho...	1.0
1	Very happy with m...	1.0

only showing top 20 rows

Accuracy: 0.7697594501718213

Fig. 1. Example of a figure of data after the classical pipeline

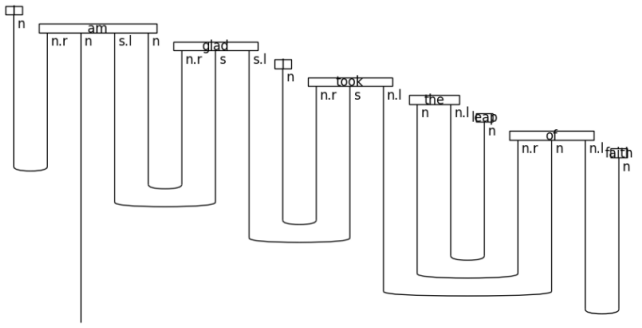

Fig. 2. Transforming sentences to diagrams in the quantum pipeline

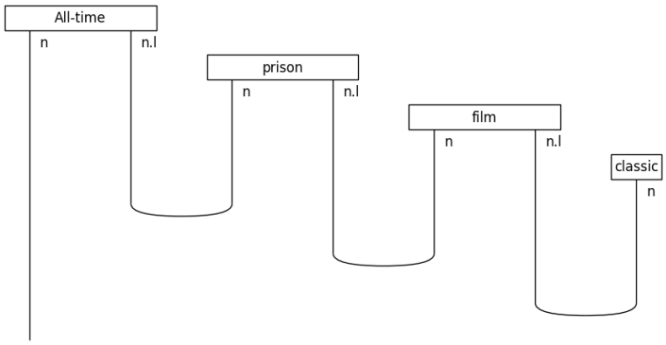

Fig. 3. Transforming sentences to diagrams in the quantum pipeline

V. Quantum Pipeline

Utilizing quantum computing frameworks such as Lambeq and pytket-qiskit, we developed a Quantum Natural Language Processing (NLP) pipeline. This quantum pipeline closely resembles the hybrid pipeline, with a key distinction being the adoption of a quantum trainer using lambeq in place of the classical PyTorch trainer.

The consolidation of sentences from three distinct files into a single cohesive file marked the initial phase of data preparation. Subsequently, several transformations were applied, including the conversion of sentences into diagrammatic representations.

Additionally, an accuracy assessment was conducted to evaluate the pipeline's performance. The outcomes of these operations were recorded in log files and stored within the Jupyter notebook for future reference.

For a deeper exploration of the shared aspects of our methodology, please refer to the hybrid pipeline section, where you will find a comprehensive overview.

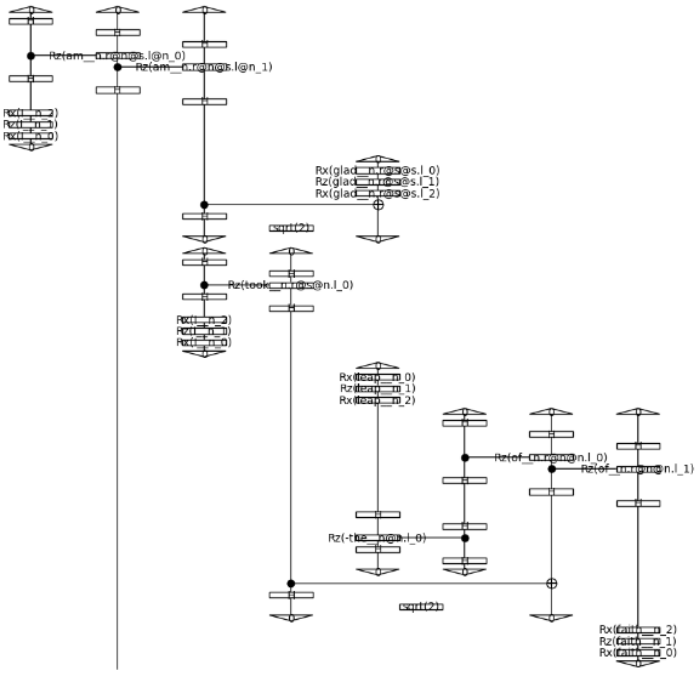

Fig. 4. Quantum circuit diagram for quantum pipeline

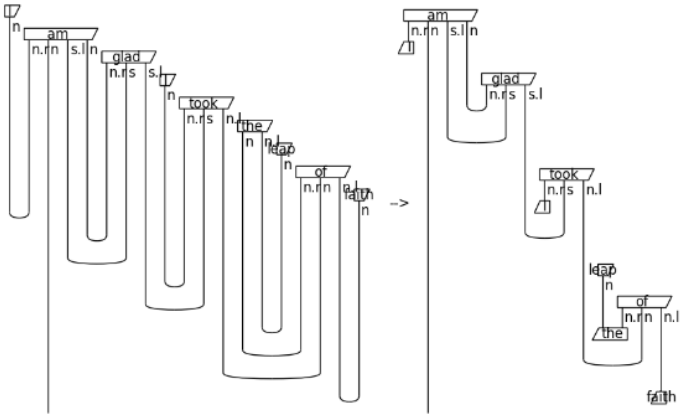

Fig. 5. Transforming sentences to diagrams in the quantum pipeline

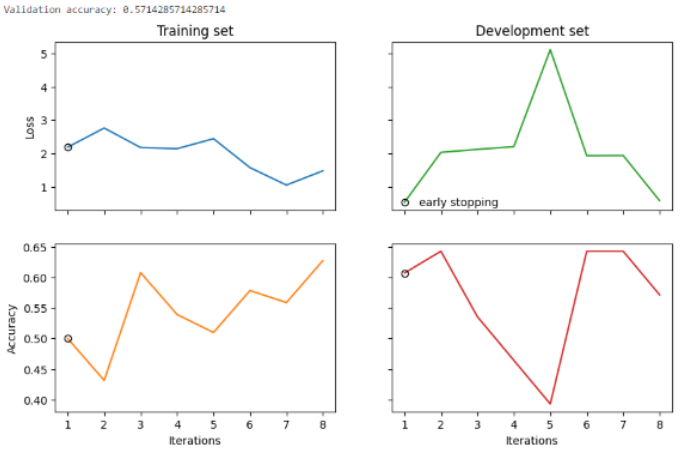

Fig. 6. Results of the quantum pipeline

VI. Hybrid Approach

The hybrid approach combines classical and quantum elements to leverage the respective strengths of both paradigms in sentence classification. Here we used the Lambeq library as well as qiskit. Lambeq is a python library made by Quantinuum to give researchers the possibility to work on not only classical but also hybrid and quantum pipelines for NLP(natural language processing).

Notably, both hybrid and quantum pipelines incorporate quantum components. Their distinction lies primarily in the choice of trainers. The hybrid approach employs a classical trainer, whereas the quantum counterpart employs a quantum trainer.

In this notebook, our primary focus centers on the implementation of the hybrid pipeline, with the subsequent assessment of its accuracy. This accuracy assessment allows for meaningful comparisons with alternative approaches, such as quantum and classical methods. For compatibility with IBM backends and simulators, we leverage pytket-qiskit, an extension of Pytket designed to execute Pytket circuits on IBM's quantum computing infrastructure.

Within the context of our experiment, epochs denotes the maximum number of iterations the trainer can perform, and batch size represents the quantity of samples processed before model updates occur. Data preparation involves the consolidation of all sentences into a single file, followed by the partitioning of sentences into training and testing sets, while labels and sentences are stored in separate lists.

One of the distinguishing features of Lambeq is its utilization of categorical quantum mechanics, a framework where states, effects, scalars, and other metrics are transformed into diagrammatic representations. In our study, sentences undergo a transformation into diagrams composed of boxes and wires, which can subsequently be translated into a quantum circuit using the zx calculus.

The quantum aspect of our methodology involves the application of a quantum ansatz known as IQPAnsatz. An ansatz serves as an initial circuit or starting point for execution on a quantum computer. The sentences provided in our dataset encompass a variety of linguistic components, including nouns, verbs, and other linguistic elements, each assigned a specific value or score.

To execute our quantum circuit, we have the option to utilize real quantum computers or simulators provided by platforms like Pennylane and Qiskit. In our study, we opt for the use of a local simulator from Qiskit known as AerBackend. The dataset is curated to align with the requirements of our trainer, ensuring its compatibility and suitability for the hybrid approach

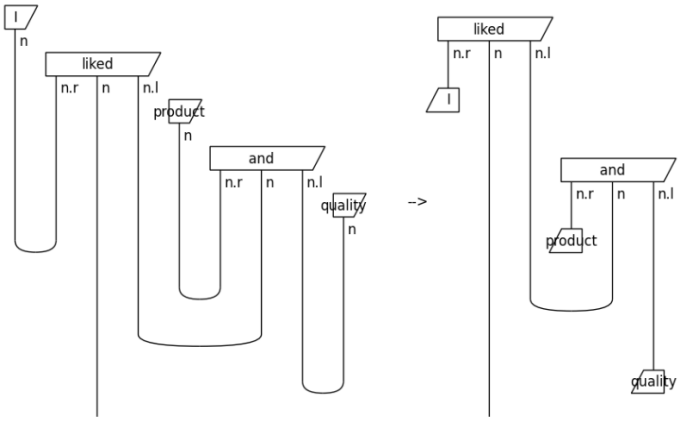

Fig. 7. Example of a figure of data after the hybrid pipeline

Fig. 8. Result of the hybrid pipeline

VII. Performance Comparison and Conclusion

In this study, we conducted a comprehensive analysis comparing the time to completion, memory consumption, and accuracy of various Natural Language Processing (NLP) pipelines.

	classical	hybrid	quantum
time	shortest	medium	longest
memory	least	medium	biggest
accuracy	best	worst	medium

Our evaluation aimed to provide insights into the choice between classical and quantum NLP techniques. It is evident that the classical NLP pipeline currently stands as the most viable option. Classical NLP has a well-established history with numerous advancements and refinements.

These accumulated developments have been seamlessly integrated into classical NLP programs, rendering them robust and highly effective. Notably, Spark-NLP, a leading player in the classical NLP domain, is recognized for its superior speed, advanced capabilities, and exceptional accuracy.

In contrast, quantum NLP is still an emerging field. Although it shows promise, it remains in its infancy compared to classical NLP. The quantum approach's advantage in terms of accuracy, particularly with a reduced training sentence requirement, is noteworthy.

However, the trade-off lies in its higher memory consumption. Notably, our analysis considers quantum NLP with noise-free quantum simulators. The introduction of noise could significantly affect quantum NLP results, potentially tilting the balance further in favor of the classical pipeline.

In summary, the classical NLP pipeline currently outperforms its quantum counterpart in terms of practicality and efficiency. While the quantum approach excels in accuracy under certain conditions, its higher memory requirements and the potential impact of noise make the classical pipeline the safer and more dependable choice for most applications.

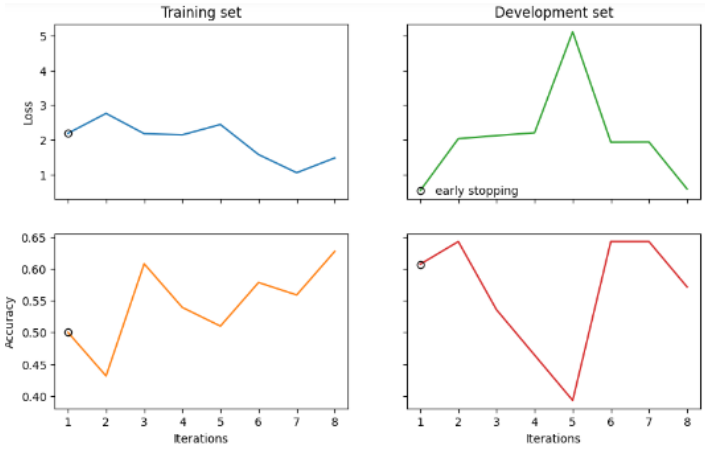

Fig. 9. Example of a figure of data after the quantum pipeline

Acknowledgment

I would like to express my heartfelt gratitude to QuantumAI for providing the invaluable platform and support that facilitated the realization of this research project. Their commitment to advancing quantum computing technology has been instrumental in our exploration of Quantum Natural Language Processing.

I also extend my sincere thanks to HyPy for their cloud services, which played a pivotal role in the computational aspects of our research, ensuring the seamless execution of quantum algorithms and data processing.

I am profoundly thankful to my dedicated team members, whose unwavering collaboration, expertise, and dedication significantly contributed to the success of this project. Their collective efforts and shared vision have been a source of inspiration and motivation throughout this research journey.

The realization of this research paper would not have been possible without the support, expertise, and camaraderie of these organizations and individuals. Their contributions are deeply appreciated, and I look forward to future collaborations in our continued pursuit of scientific exploration and innovation.

References

[1] Lambeq tutorial for experimental Quantum Natural Language Processing (QNLP), created by Quantinuum's QNLP team.

[2] Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, Bob Coecke, "QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer."

[3] Dimitri Kartsaklis, Ian Fan, Richie Yeung, Anna Pearson, Robin Lorenz, Alexis Toumi, Giovanni de Felice, Konstantinos Meichanetzidis, Stephen Clark, Bob Coecke, ”lambeq: An Efficient High-Level Python Library for Quantum NLP.”
[4] An Introduction to Quantum Natural Language Processing by Srinjoy Gangulyon Udemy.
[5] Sentiment analysis of twitter reviews github repositary by Hussein Shiri.
[6] Shervin Le Du, UPM, ”A gentle introduction to Quantum Natural Language Processing.”