ATHENS UNIVERSITY OF BUSINESS & ECONOMICS

BACHELOR'S THESIS

Harnessing BERT for Multilingual Sentiment Analysis: A Comparative Study

---

Author:
Nikolaos Mermigkas

Supervisor:
Professor Emmanouil Zachariadis

A thesis submitted in fulfillment of the requirements for my Bachelor's Degree

Department of Management Science and Technology

Abstract

The thesis provides a literature review on sentiment analysis, focusing on the use of BERT models in multilingual sentiment analysis. The review explores the challenges associated with multilingual sentiment analysis, such as handling language-specific nuances and resource distribution among languages. It also covers recent advances and emerging trends in multilingual sentiment analysis using BERT models, including enhancements in pretraining techniques and cross-lingual transfer learning. The paper highlights novel techniques like domain adaptation, hierarchical BERT, and lightweight BERT variants, as well as the latest benchmark datasets and evaluation metrics for multilingual sentiment analysis. Furthermore, the IBM use case discussed in the thesis focuses on sentiment analysis using the watBERT model, a BERT version of IBM. The methodology used in the study includes data selection, evaluation metrics, experimental setup, and statistical analysis. The study utilizes TSA (Targeted Sentiment Analysis) labeled datasets, such as TSA-MD, which consists of manually labeled reviews from multiple domains, and MAMS, a dataset over restaurant reviews. The watBERT model is tested on various multilingual datasets, including YASO, MAMS, and SE16, and the performance is evaluated using precision, recall, and F1 score. The study also explores compression techniques, distillation, and CPU-optimized models to improve the model's throughput and performance. A part of the study comes from internal testing of the watBERT model, developed by the Watson NLP team at IBM, which allowed the use of the tests for the purpose of this comparative study. The results obtained from the experiments are presented through tables and plots, providing insights into the model's strengths and limitations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Sentiment analysis, or opinion mining, has become increasingly important in various applications and industries, such as social media monitoring, customer feedback analysis, and market research [15]. With the rapid growth of digital content in multiple languages, multilingual sentiment analysis has emerged as a vital subfield within natural language processing (NLP). Analyzing sentiment across different languages poses significant challenges due to linguistic variations, including differences in syntax, semantics, and cultural nuances.

Bidirectional Encoder Representations from Transformers (BERT) [4] is a pre-trained deep learning model that has achieved state-of-the-art results in a wide range of NLP tasks, including sentiment analysis. BERT's architecture and pretraining process make it particularly suitable for multilingual tasks, providing a promising avenue for addressing the challenges in multilingual sentiment analysis [1].

In this thesis, we focus on a bibliographic study of BERT models and their application in multilingual sentiment analysis, as well as an IBM use case where we, as an intern in the NLP team, have had the opportunity to apply these models in a real-world context.

1.2 Objectives and research questions

The main objectives of this thesis are to:

Conduct a comprehensive literature review on multilingual sentiment analysis techniques, with a focus on BERT models.

Investigate the challenges and advancements in multilingual sentiment analysis and how BERT models can address these issues.

Examine the IBM use case, outlining the methodology used, the results obtained, and the lessons learned from applying BERT models to a real-world multilingual sentiment analysis problem.

The primary research questions we aim to answer are:

How have BERT models contributed to the field of multilingual sentiment analysis?

What are the main challenges in multilingual sentiment analysis, and how can BERT models help overcome them? [3]

How have BERT models been applied in the IBM use case, and what insights can be drawn from this application?

1.3 Scope and Limitations

The scope of this thesis is focused on the bibliographic study of multilingual sentiment analysis techniques, particularly BERT models, and the IBM use case. While we provide an extensive review of the literature and a detailed examination of the IBM use case, the results and findings may be limited by the available research and the specific context of the IBM project. Additionally, our analysis may not cover all possible techniques or approaches for multilingual sentiment analysis but will serve as a solid foundation for understanding the topic and its current state.

Chapter 2

Literature Review

2.1 Sentiment Analysis: Overview and Techniques

In this section, we provide an overview of sentiment analysis, its applications, and the key techniques that have been employed in the field. We discuss traditional methods, such as rule-based and lexicon-based approaches, as well as machine learning techniques, including supervised, unsupervised, and deep learning methods. We also briefly touch upon the transition from these earlier techniques to advanced models like BERT.

Sentiment Analysis, also known as opinion mining, is an essential domain within Natural Language Processing (NLP) that focuses on identifying and categorizing opinions expressed in a piece of text, particularly to determine the writer's attitude towards a particular topic, event, or product [15]. The task is commonly defined as classifying the polarity of a given text at the document, sentence or feature/aspect level -whether the expressed opinion is positive, negative or neutral [19].

In recent years, with the rise of social media platforms, blogs and discussion forums -among others, sentiment analysis has become an extremely popular topic of research. This is attributed to the numerous applications of sentiment analysis in fields such as customer service, marketing, public opinion monitoring and even political profiling.

From increasing customer satisfaction at an eshop based on product reviews to crafting political profiles based on tweets, sentiment analysis can be utilized in numerous ways. Traditional Sentiment Analysis methodologies can broadly be divided into three categories: rule-based approaches, machine learning-based approaches, and hybrid approaches.

Rule-based approaches employ a set of manually curated rules and often rely on lexicon-based techniques where words are assigned sentiment scores and the total sentiment of a sentence or document is computed accordingly. Such systems may also consider syntactic, morphological, and discourse level cues [13]. Although rule-based systems can be highly interpretable, they are often brittle and require extensive manual tuning.

Machine Learning-based approaches take advantage of both supervised and unsupervised learning algorithms. Early works in this domain utilized traditional machine learning algorithms such as Naïve Bayes, Decision Trees, and Support Vector Machines with handcrafted features such as bag of words, n-grams, or sentiment lexicons [15]. More recent approaches use deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which learn to extract features directly from the data [20].

Hybrid approaches combine rule-based and machine learning methods, often to leverage the interpretability of rule-based methods and the generalization capability of machine learning models [**cambria2013**].

The advent of Transformer-based models, most notably BERT [4], has shifted the paradigm of sentiment analysis. With their ability to model complex patterns and dependencies in text, these models have achieved state-of-the-art performance on a range of sentiment analysis tasks, transforming the landscape of sentiment analysis techniques.

In the next sections, we'll examine the challenges and approaches related to conducting sentiment analysis across multiple languages, as well as the specific role that BERT models play in this context.

2.2 Multilingual Sentiment Analysis: Challenges and Approaches

Here, we delve into the specific challenges associated with multilingual sentiment analysis, such as language-specific nuances, cultural differences, and limited resources for low-resource languages. We review various approaches to address these challenges, including cross-lingual transfer learning, unsupervised representation learning, and multilingual pretraining.

Multilingual sentiment analysis has attracted substantial attention as the digital universe expands into a myriad of languages, driven by the global nature of the internet and social media platforms. This section will explore the specific challenges associated with multilingual sentiment analysis and discuss various methods proposed in recent literature to tackle these challenges.

One of the primary challenges in multilingual sentiment analysis is handling language-specific nuances and idiomatic expressions that hold different sentiment values in different cultures and languages [17]. For example, an English language phrase directly translated to another language might not carry the same sentiment due to cultural and contextual differences. This calls for sophisticated models that can understand the subtleties of different languages and cultures.

Another significant challenge lies in resource distribution among languages. Many languages, particularly those spoken in less economically developed regions, are often low-resource in terms of available annotated sentiment analysis datasets [3]. This imbalance poses a significant challenge, as most sentiment analysis models require large amounts of annotated data to train effectively.

Cross-lingual transfer learning is one of the prominent methods used to address these challenges. Models are trained on a resource-rich language and then transferred to other languages, benefiting from the similarities between languages [3, 14]. BERT models pre-trained on multilingual corpora, for example, are often employed in this regard, leveraging the shared structures among languages in the Transformer's hidden states.

Unsupervised representation learning is another approach that has gained traction. This involves learning language representations from large amounts of unlabeled data, bypassing the need for extensive annotation [3]. Techniques such as language model pretraining and autoencoding are often employed in this area.

To conclude, the field of multilingual sentiment analysis presents unique challenges that are being addressed by increasingly sophisticated approaches. These range from leveraging shared structures among languages, learning representations in an unsupervised manner, to creating sophisticated models that understand the cultural and contextual nuances of different languages.

In the following sections, we are going to explore the BERT model and its architecture (transformer based , as well as its variants that tackle the aforementioned challenges.

2.3 Transformers

At the heart of many modern NLP models lies a Transformer architecture which leverages the power of attention mechanisms to model dependencies between words in a sentence, regardless of their position, bettering previous sequential approaches [21].

The Transformer, based solely on attention mechanisms, has brought about a significant shift in the field of Natural Language Processing (NLP). Discarding the traditional reliance on recurrence and convolutions, the Transformer model adopts a novel approach that allows parallel processing of input and output sequences.

Central to the Transformer model is the "Scaled Dot-Product Attention" mechanism. This attention mechanism computes the relevance of different parts of the input sequence for each element of the output sequence, essentially allowing the model to focus on various parts of the input when generating each component of the output. This attribute has proven highly effective, especially for translation tasks, where individual words in the input sentence may hold different levels of relevance for each word in the output sentence.

Another significant contribution of Vaswani et al. (2017) [21] is the concept of "Positional Encoding". Given the absence of recurrence in the Transformer architecture, this technique enables the model to consider the order of words in the input sequence, an otherwise inaccessible feature.

The design principles and mechanisms introduced by the Transformer model have had a profound influence on later models in NLP, most notably BERT, GPT-2, GPT-3, and others.

As such, the Transformer architecture forms the fundamental backbone of BERT, facilitating its exceptional performance in various NLP tasks, including sentiment analysis.

2.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model that learns from unlabelled text data. Devlin et al. (2018) [4] demonstrated that pre-training language representations prior to fine-tuning on specific tasks significantly improved upon the then state-of-the-art across multiple NLP benchmarks. With a multi-layered stack of encoders from the Transformer architecture, one of its strengths lies in its ability to take entire text sequences into account at once, as opposed to analyzing individual words or phrases in isolation. One key innovation

of BERT was its utilization of bi-directional (left to right and right to left) context in contrast to previous unidirectional or shallow bidirectional architectures.

What characterizes BERT is its design around the concept of pre-training and fine-tuning. During the pre-training phase, BERT is trained on a large text corpus (such as Wikipedia) with unsupervised learning. Two tasks are performed in this phase: the masked language model (MLM) task, and the next sentence prediction (NSP) task. The MLM task randomly masks words in a sentence and then prompts the model to predict the masked word based on the context provided by the non-masked words. The NSP task involves determining whether a sentence logically follows a given preceding sentence.

Following pre-training, BERT undergoes a fine-tuning phase for specific NLP tasks, such as sentiment analysis, named entity recognition (NER), and question-answering tasks. In this phase, BERT is fine-tuned on task-specific labeled data to perform specific tasks with high precision.

BERT's architecture and learning mechanisms have contributed significantly to its exceptional performance, often outperforming other models in a wide range of NLP tasks. It has set new state-of-the-art performances on eleven NLP tasks and has become a benchmark in the field.

2.5 Variants Of BERT

Following the success of BERT, several variants and improvements on the model have been proposed. These variants, designed with various enhancements and specific purposes, have achieved notable success in different NLP tasks, including sentiment analysis.

RoBERTa [12], or Robustly Optimized BERT Pretraining Approach, sought to optimize the BERT pre-training process by making it longer and more dynamic. This variant extended BERT's approach by dynamically changing the masking pattern during training, leading to better performance on downstream tasks. RoBERTa uses larger batch sizes, removes the next sentence prediction task from pretraining, and trains on larger amounts of data for longer periods. RoBERTa has been shown to outperform BERT on several NLP benchmarks

Meanwhile, ALBERT [10], officially published as A Lite BERT, aimed to solve the problem of the increasing model size by sharing parameters across layers, reducing redundancy and maintaining high performance. It employs cross-layer parameter sharing to reduce the parameters and computational cost, and a factorized embedding parameterization to decompose large vocabulary embedding into smaller matrices. Despite its smaller size, ALBERT performs comparably or even better than its larger counterparts on several benchmarks.

DistilBERT [18] is a distilled version of BERT, making it smaller, faster, and lighter, while retaining most of BERT's performance. This model was trained using knowledge distillation, a technique where a smaller model (the student) is trained to reproduce the behavior of a larger model (the teacher).

Another noteworthy model is SpanBERT [8] which improved upon BERT by representing and predicting spans of text instead of individual tokens.

Recent works have started to explore the applications of BERT's attention mechanism in different scenarios. One such instance is Sun et al.'s (2019) [19] investigation

on the use of intermediate layers of BERT for aspect-based sentiment analysis. Their study found that lower-level layers are beneficial for the extraction of syntax-related features, while higher layers are adept at extracting semantic features.

These advancements illustrate the continued evolution and importance of BERT and its variants in the landscape of NLP. Yet, while these models achieve high performance in various tasks, it remains vital to explore how their capabilities extend to more complex, multilingual scenarios, especially in the context of sentiment analysis.

2.6 BERT Models for Multilingual Sentiment Analysis

This section focuses on BERT models and their application in multilingual sentiment analysis. We discuss the architecture, pre-training process, and fine-tuning strategies for BERT, as well as its multilingual variants, such as XLM-RoBERTa [3]. We review key papers that demonstrate the effectiveness of BERT models in multilingual sentiment analysis tasks and discuss their implications for the field.

BERT and its multilingual extension, mBERT, revolutionized the NLP field with their ability to understand the context of language bidirectionally [16]. Traditional language models analyze text either from left to right or right to left, whereas BERT can assess both directions simultaneously. This results in an understanding of the context of words within their full sentence structure. Hence, it's much better equipped to understand the subtleties of sentiment analysis tasks, even in a multilingual setting.

mBERT [16], specifically, was trained on a corpus of 104 languages, leveraging shared subword tokenization to bridge languages. This means that it uses a shared vocabulary to represent subword units in all languages. This approach, while economically effective, also helps recognize the similarities in structure and semantics across languages, leading to more generalized representations.

The studies by Wang (2020) [22] and Conneau (2020) [3] offer compelling examples of mBERT's potential in multilingual sentiment analysis tasks. They demonstrate how mBERT's shared structure aids in transferring sentiment knowledge from one language to another, even in low-resource language scenarios.

However, research did not stop with mBERT. XLM-RoBERTa (XLM-R) [3] pushed the boundaries further by enlarging the training corpus to include even more languages (up to 100) and introducing refined training techniques. XLM-R, compared to mBERT, showed significant improvements in various cross-lingual benchmarks, including sentiment analysis tasks.

Furthermore, other variants of BERT have been introduced with more specific focus. For instance, ALBERT [10] modifies the architecture of BERT to be more parameter-efficient, which can be valuable in deploying models for real-world applications where computational resources might be a concern. Another notable variant is ELECTRA [2] that innovates on the pretraining method for more sample-efficient learning.

In conclusion, the introduction of BERT and its variants has been a transformative force in the field of multilingual sentiment analysis. These models have offered innovative solutions to the challenges of understanding and translating sentiments across diverse languages and cultures, presenting an exciting avenue for future research.

2.7 Extensions and Improvements to BERT

In this part, we highlight recent advances and emerging trends in multilingual sentiment analysis using BERT models. We cover novel techniques, such as domain adaptation, hierarchical BERT, and lightweight BERT variants, as well as the latest benchmark datasets and evaluation metrics for multilingual sentiment analysis.

In the short span of time since its introduction, the BERT model has seen several extensions and improvements. A major thrust of this work has been towards resource optimization and expanded language support.

Resource Optimization

With BERT's heavy computational requirements, one line of research has focused on making the model more efficient, aiming for comparable performance with fewer resources. Several lighter versions of BERT have been developed to meet this need:

DistilBERT [19] is a smaller, faster, and cheaper version of BERT. It retains 95% of BERT's performance while being 40% smaller and 60% faster, achieved by a process called knowledge distillation where the distilled model learns from the teacher model's output distributions.

ALBERT [10] achieves further parameter reduction by factorizing the embedding layer and sharing parameters across the hidden layers. It retains the high performance of BERT while significantly reducing the number of parameters.

TinyBERT [7] also utilizes knowledge distillation, but distills both the hidden states and the attention distribution, resulting in a model that is only 28% the size of BERT, but retains a significant portion of its performance.

Language Support

Expanding BERT's applicability beyond English has been another focus of research:

mBERT [16] is a multilingual version of BERT that has been pre-trained on the top 104 languages with the largest Wikipedias. It has proven useful for many cross-lingual transfer tasks, however, it has been noted that it struggles with low-resource languages.

XLM-R [3] extends mBERT by incorporating RoBERTa's training improvements and increasing the amount of training data. The model supports 100 languages and has achieved state-of-the-art results on many cross-lingual benchmarks.

UDify, a multilingual BERT based model, is examined in Dan Kondratyuk's and Milan Strakathe's et al. (2019) [9] study regarding its language support capabilities. The model demonstrates remarkable proficiency in predicting part-of-speech, morphological features, lemmas, and dependency trees across multiple languages. Pretrained on 104 languages and fine-tuned on Universal Dependencies treebanks, UDify achieves state-of-the-art performance in various syntactic tasks without requiring language-specific components or recurrent models. The study highlights the advantages of multilingual modeling, emphasizing the benefits derived from cross-linguistic annotations, particularly for low-resource languages. Furthermore, the exceptional capabilities of pretrained self-attention networks in multilingual dependency

parsing are underscored. Overall, UDify's extensive language support showcases its potential in advancing multilingual natural language processing research.

2.8 Recent Advances and Emerging Trends

With the advent of Transformer-based models like BERT, the field of NLP has seen an explosion of research activity. The landscape of sentiment analysis is continuously changing and evolving with the development of more sophisticated techniques and models. This section will highlight some of the recent advances and emerging trends in sentiment analysis using BERT and its variants.

Enhancements in Pretraining Techniques

Research is ongoing to improve pretraining techniques, furthering our understanding of the capabilities and limitations of these methods. For example, ELECTRA [2] introduces a new pretraining approach, the replaced token detection task, that trains models to distinguish "real" tokens in the input from "fake" ones output by a generator network. ELECTRA models have shown competitive results with significantly less computational resources compared to BERT.

Advances in Cross-lingual Transfer Learning

Cross-lingual transfer learning has shown great promise in overcoming the data scarcity problem in low-resource languages. XLM [**conneau2029**] and XLM-R [3] represent significant advances in this direction, achieving state-of-the-art results on several cross-lingual benchmarks.

While multilingual models like mBERT and XLM-R have made strides in cross-lingual transfer learning, they still struggle with languages that are under-represented in the training data. Recent research has proposed methods to better handle these languages, such as leveraging monolingual data for unsupervised pretraining or designing better cross-lingual objective functions (Hu et al., 2020).

Leveraging Unsupervised and Semi-supervised Learning

Another trend is the increased use of unsupervised and semi-supervised learning to make the most of the vast amount of unlabeled data available. ULMFiT [5] employs a semi-supervised approach, using unsupervised pretraining followed by supervised fine-tuning, and has demonstrated success in several sentiment analysis tasks.

Focus on Interpretability and Explainability

As deep learning models become more complex, there is a growing emphasis on interpretability and explainability. This is particularly relevant for models like BERT, which are often described as "black boxes". An approach such as attention-based explanations [6] has been proposed to shed light on what these models learn. This topic, although immensely exciting and emergingly important in the future, will

not consitute a topic of analysis. It is highly suggested to refer to this study about explainability and interpretability[6].

Incorporation of Domain Knowledge

Finally, there is a growing interest in incorporating domain knowledge into BERT models to improve their performance on specific tasks. This includes methods like fine-tuning with in-domain data, injecting expert knowledge into models, and adapting models to specific domains (Lee et al., 2020).

As we move forward, these trends will likely continue to shape the future of sentiment analysis, opening up new possibilities and challenges.

Robustness and Adversarial Training

As Transformer-based models are increasingly used in sensitive and high-stakes applications, it's important for them to be robust against adversarial attacks. Adversarial training, a form of robust optimization, has been proposed as a solution to improve the robustness of these models (Jin et al., 2019). This involves training the model on perturbed versions of the input data, forcing it to learn more general and robust representations.

Combining BERT with Other Modalities

Recent studies have also started to explore the combination of BERT with other modalities, such as visual or acoustic signals, for tasks like multimodal sentiment analysis [23]. This allows models to leverage additional information from other sources, potentially leading to better performance.

The field of sentiment analysis using BERT and its variants is rapidly evolving, with many exciting developments on the horizon. It's an exciting time to be conducting research in this area, and it will be interesting to see where these trends lead in the coming years.

Chapter 3

Targeted Multilingual Sentiment Analysis: An IBM Use Case

3.1 Project Context and Objectives

This section provides an overview of the IBM project, Watson NLP, some concrete background to it, and describes the goals it seeks to achieve. This project revolves, among others, around the field of multilingual sentiment analysis, which is a challenging task due to language-specific nuances, cultural differences, and limited resources in low-resource languages [11, 24]. The goal of the project is to harness the power of the BERT model to develop an effective and efficient solution for sentiment analysis in multiple languages.

IBM has been experimenting with variations of BERT models for multilingual sentiment analysis. However, through a lot of trial and error and experiments, the R&D department of IBM reached the decision to utilize watBERT: an internally developed fine-tuned version of the base BERT, tailored to the needs of the company. IBM currently supports multilingual sentiment analysis in 24 languages and is working on supporting new languages such as Greek.

While the primary objective of multilingual sentiment analysis is to analyze sentiment across multiple languages, it naturally extends to targeted sentiment analysis, which aims to identify sentiment-bearing aspects or entities within the text [19].

Targeted sentiment analysis plays a crucial role in comprehending the sentiment linked to particular facets, entities, or subjects of interest in a given text. By extracting sentiment from detailed aspects or entities, businesses can obtain deeper insights into customer preferences, product attributes, and brand perception. Hence, conducting a thorough comparison of targeted sentiment analysis involving watBERT and alternative algorithms becomes highly pertinent, introducing a valuable layer of analysis in the domain of multilingual sentiment analysis.

The comparative analysis aims to evaluate the performance, efficacy, and adaptability of watBERT in contrast to cutting-edge algorithms. Through the examination of metrics like accuracy, precision, recall, and F1 score across diverse datasets and languages, the study provides valuable insights into the strengths and limitations of watBERT in targeted sentiment analysis scenarios. This comparative assessment not only substantiates the capabilities of watBERT but also illuminates its distinct contributions to sentiment analysis tasks within multilingual contexts.

Having established this connection, the following sections will present comprehensive findings and metrics derived from the comparative analysis of targeted

sentiment analysis utilizing watBERT and other algorithms. By showcasing these outcomes, the performance of watBERT in relation to alternative approaches and underscore its proficiency in extracting sentiment from specific aspects or entities can be effectively demonstrated. As a result, this research contributes significantly to the broader landscape of multilingual sentiment analysis.

The following sections describe the methodology used, the results obtained, and the insights gained through the application of the watBERT model on IBM use cases.

3.2 Methodology

To ensure a rigorous and reliable analysis, a systematic approach was followed. The methodology encompasses several key steps, including data selection, evaluation metrics, experimental setup, and statistical analysis. By adhering to this methodology, the study ensures a fair and objective comparison that yields meaningful insights into the strengths and limitations of watBERT.

3.2.1 Data

Training

TSA labeled datasets:

TSA-MD: A diverse TSA English training set obtained by manual labeling. IBM's internal team collected 952 sentences of reviews from multiple domains: First, reviews were written by crowd annotators in a given domain, on a topic of their choice. Then, the reviews were annotated for TSA by asking annotators to mark all sentiment-bearing targets in each sentence.

MAMS: A TSA English dataset over restaurant reviews. In MAMS, each sentence has at least two targets annotated with different sentiments. The sentiments are either positive, negative or neutral. To match our setup, the neutral labels were removed.

Weekly labeled data:

Unlabeled texts were used in a self-supervised manner to produce weak labels. The texts were automatically annotated for TSA by a "seed" TSA model. The prediction were filtered, and reused as weak labels for another training phase. A similar process using the YELP reviews dataset is described in the paper Multi-Domain Targeted Sentiment Analysis.

The unlabeled texts that were used are:

Unlabeled sentiment data (SD), which contains sentences conveying sentiments, from multiple sources:

ALMR, ALMS, ALROT, EMARC, EMAVA, EMFCASE, EMICONV, EMMOVDLG, EMPAY, EMPUBCROWD, EMRED, EMSPEECH, EMTED, EMVOG, EMWT-BOT, EMWVA, FPB_75, WDSHLTH, Medallia - IBM NPS, comments from IBM customers, Austin Survey dataset.

LexisNexis: English reviews mined from the LexisNexis corpus. The LexisNexis dataset was used in first releases. It was replaced by the SD dataset.

Evaluation

YASO: A dataset in English we collected of 2215 sentences annotated for TSA. The sentences are taken from user reviews from multiple sources. This dataset covers reviews from many domains, and is thus a good choice for multi-domain evaluation. The construction of the YASO dataset is detailed in the paper YASO: A Targeted Sentiment Analysis Evaluation Dataset for Open-Domain Reviews.

MAMS (test set): 500 English sentences from of the MAMS test set were used for evaluation.

SemEval'2016 (SE'16): A multilingual TSA dataset.

Example Review with annotation

Apart from the conceptual difference of the sentiment analysis and the targeted sentiment analysis, it is important to have a clear and concise example about the difference in the data. As the name suggests, *targeted sentiment analysis (TSA)* can output the sentiment for a number of targets/entities of a document/sentence.

The following example demonstrates a review of one of the datasets that are used for the training of the model:

Listing 3.1: JSON example of Targeted Sentiment Analysis

```
{
  "text": "Great content, categories, and variety of
     material o choose from and nice service to the
     customer also.",
  "targets": [
    {
      "text": "content",
      "location": {
        "begin": 6,
        "end": 13
      },
      "sentiment": "positive",
      "detected_by": {
        "detected_positive": 4.0,
        "detected_negative": 0.0,
        "detected_mixed": 0.0
      }
    },
    {
      "text": "service to the customer",
      "location": {
        "begin": 74,
```

```
        "end": 97
      },
      "sentiment": "positive",
      "detected_by": {
        "detected_positive": 4.0,
        "detected_negative": 0.0,
        "detected_mixed": 0.0
      }
    },
    {
      "text": "categories",
      "location": {
        "begin": 15,
        "end": 25
      },
      "sentiment": "positive",
      "detected_by": {
        "detected_positive": 4.0,
        "detected_negative": 0.0,
        "detected_mixed": 0.0
      }
    },
    {
      "text": "and variety of material",
      "location": {
        "begin": 27,
        "end": 50
      },
      "sentiment": "positive",
      "detected_by": {
        "detected_positive": 4.0,
        "detected_negative": 0.0,
        "detected_mixed": 0.0
      }
    }
  ]
}
```


That was a random example and a total positive one, since all the entities are characterized as positive by all the annotators. This is to be seen in the *"detected_by"* field of the example. It is called polarity and indicates the unanimity (or not) of the annotators.

The annotations refer to four targets of the selected document. The *content,service to the customer, categories, and variety of material* are all positively annotated, as the review suggest.

Of course, there might be cases where a review is totally negative or mixed, containing neutral, positive and negative sentiments.

This is exactly what IBM through watBERT seeks to achieve: To capture with optimal performance the sentiments of different targets in a document.

3.2.2 Algorithms

As part of the methodology, several compression techniques were incorporated, which aimed at enhancing the model's throughput. These techniques include Layer Removal, which reduces the number of encoder layers, specifically targeting the last layers of the model. Additionally, we have also implemented techniques to reduce the Feed Forward Network (FFN) dimension, essentially pruning the intermediate size of all encoders. Another notable strategy is Distillation, where a distilled version of the TSA models by fine-tuning the IBM Foundation Distilled Watson Multilingual XLM-RoBERTa Model was created.

To provide a more concise context we list the algorithms presented and compared as follows:

- BERT - Bidirectional Encoder Representations from Transformers. The Transformer Encoder stack was released by Google in 2018. [4]
- Multilingual BERT (mBERT). [16]
- WatBERT.

Compression techniques are used for improving the model's throughput. These techniques were used:

- Layer Removal : reduces the number of encoder layers. The removed layers were the last layers of the model.
- Reducing Feed Forward Network (FFN) Dimension / FFN Pruning : Reduces the intermediate size of all encoders.

Distillation: a distilled version of the TSA models was created by fine-tuning the IBM Foundation Distilled Watson Multilingual XLM-RoBERTa Model.

3.2.3 Evaluation Metrics

The graphs below show the Precision and Recall obtained while varying the prediction threshold (the threshold itself is not plotted). The Y-Axis shows the Precision at a specific threshold (P@TH), while the X-Axis shows the Recall at a specific threshold (R@TH).

The formulas for the above metrics (Precision and Recall) are the following:

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures the proportion of correctly predicted positive observations to the total predicted positives. High precision relates to a low false-positive rate

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall expresses the ability to find all relevant instances in a dataset. High recall means that the class is correctly recognized, thus having a small number of False Negatives (FN).

F1 Score:

$$F1 \text{ Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

The F1 score is the harmonic mean of precision and recall. While the regular mean treats all values equally, the harmonic mean gives much more weight to low values. As a result, the classifier will only get a high F1 score if both recall and precision are high.

Relevant Terminology:

- **TP**: True Positives, the number of instances that were correctly identified as positive.
- **FP**: False Positives, the number of instances that were incorrectly identified as positive.
- **FN**: False Negatives, the number of instances that were incorrectly identified as negative.

Note that average recall (R@TH) and average precision (P@TH) are calculated using the formulas above by dividing with the number of training runs (in most cases 5).

The graphs are either:

In the experimentation phase: averages over 5 training runs with different random seeds.

In the section describing the delivered models: results of a single model.

3.2.4 Experimentation Phase

For a more holistic approach, the experiments are split into two sections; the ones without compression/distillation and the ones with compression and distillation.

To gain a comprehensive understanding of the role compression and distillation techniques play on model performance, the study is designed to encompass two distinct experimental segments.

The initial segment involves experimentation without the application of compression and distillation strategies. This enables to establish a baseline performance metric that encapsulates the unaltered efficiency and capability of the models. It offers an opportunity to understand the inherent strengths and potential limitations of the models, which serves as a fundamental building block for the subsequent stages of our study.

The latter segment integrates the application of compression and distillation techniques. The performance metrics derived from this phase, juxtaposed against the baseline established in the initial segment, allows us to ascertain the degree of improvement achieved through these modifications. It underscores the specific contributions of compression and distillation techniques and enables a critical examination of any potential trade-offs.

By dividing the study in this manner, we provide an in-depth and encompassing investigation into our modifications, thereby ensuring a robust evaluation of the techniques employed to enhance the model's throughput and overall efficacy.

Basic Setup

Despite the fact that there are two distinct types of experiments, for consistency purposes, there is a basic setup. The experiments in this phase included training with 5 different random seeds, and reporting averaged results. All the graphs in this section (of Experimentation Phase) represent such averages, unless stated otherwise. Of course, each experiment, irrespective of the section (distilled/compressed or not) has its own setup, which is clearly stated.

Following, there are the hyperparameters listed that accompany each experiment (unless stated otherwise).

Hyperparameters

- Train batch size: 64
- Number of epochs: depends on the experiment. This parameter is important for obtaining good quality performance.
- Early stopping: not used, since it is not yet implemented for Transformers blocks.
- Learning rate: 3e-5
- Lower casing: not used
- Train/Predict stride: 64
- Max sequence length (both for train and predict): 128

Main experiments without compression/distillation

WatBERT Vs. Google-based BERT (G-BERT)

Setup

- Same data: training on MAMS, TSA-MD and weak labels from LexisNexis.
- Number of epochs: 18 for the WatBERT model.
- BERT training with the same parameters.
- WatBERT was trained for a fixed number of 18 epochs.
- No compression, for both models.
- No threshold applied to the confidence scores for WatBERT.
- The default confidence threshold of 0.82 was applied for the G-BERT model.
- Results are averages over 5 runs.

Results

Performance was comparable on both YASO (top) and MAMS (bottom), as can be viewed in the plots below for YASO 3.1 and for MAMS 3.2 respectively.

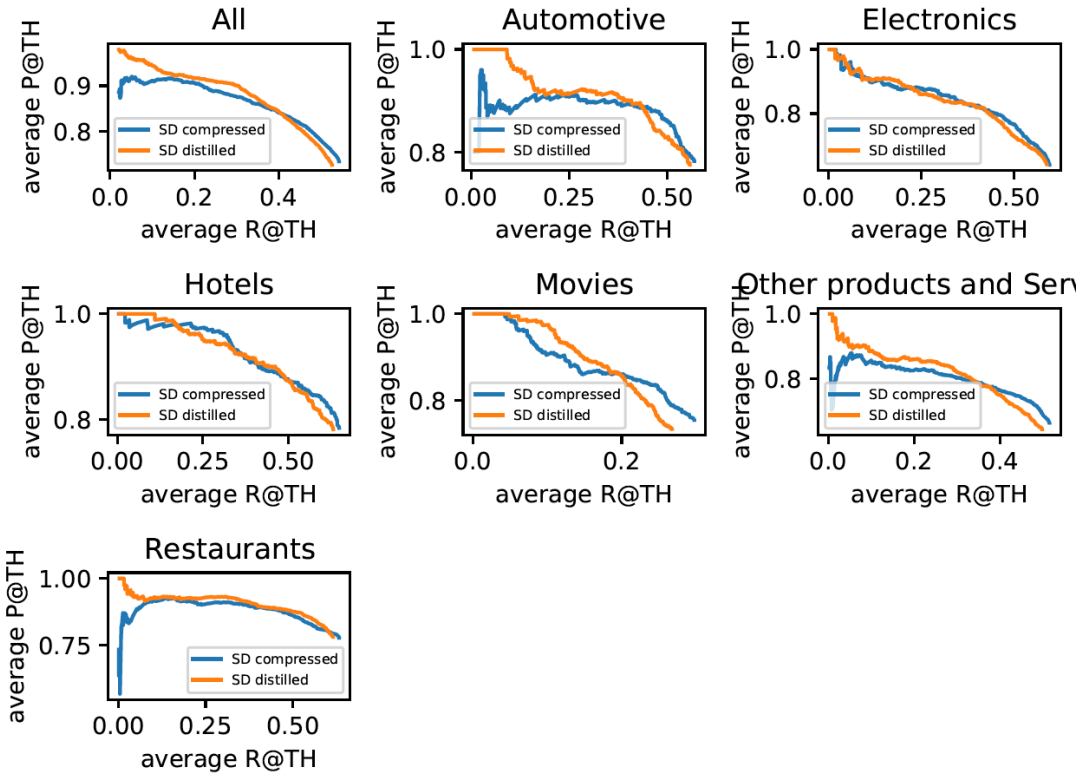

FIGURE 3.1: Compressed vs distilled SD on YASO dataset

Note: the fact that G-BERT [4] shows lower recall should be ignored, it stems from the confidence threshold that was applied for that model.

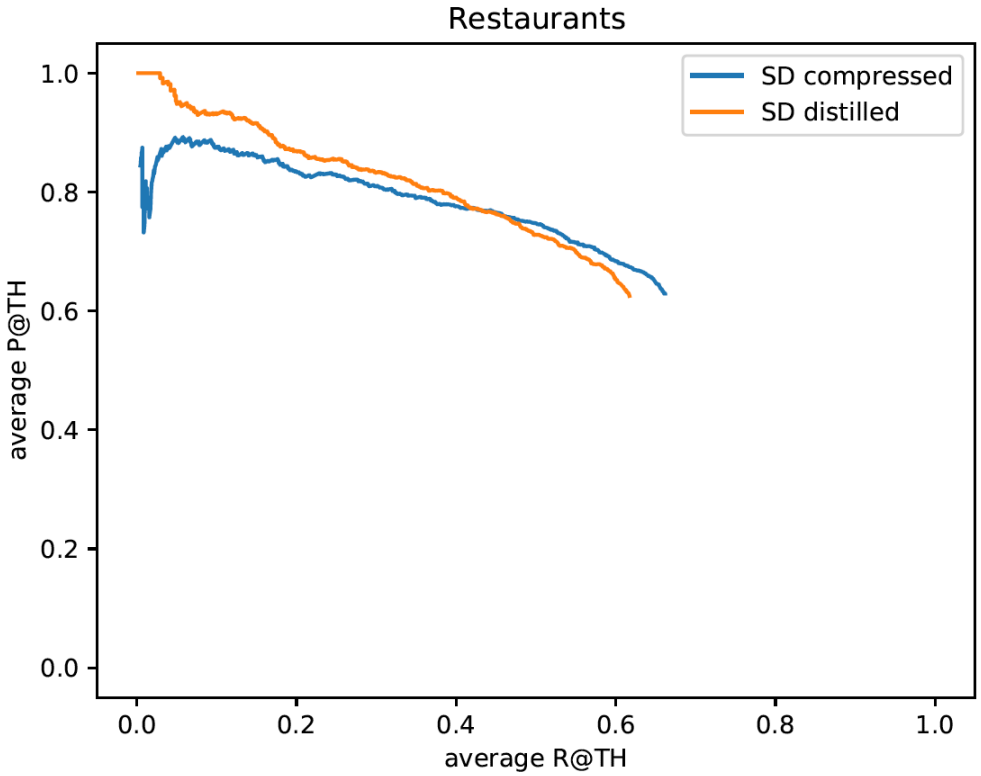

FIGURE 3.2: Compressed vs distilled SD on MAMS dataset

Compression and distillation experiments

The following section presents experiments aimed at improving runtime and memory consumption.

Compression vs. distillation with SD data

This experiment compares WatBERT models compressed to 8 layers, with 30% pruning, to the distilled WatBERT model released in April'23 (V2). The training data is the same: the SD weak labels dataset, with strong labels from MAMS and TSA-MD.

Setup

- Number of epochs: 30 for distilled WatBERT, 25 for the compressed WatBERT.
- No threshold applied to the confidence scores.
- Results are averages over 5 runs.

Results

Results are comparable, with a very small advantage to the compressed model (on YASO).

Compression with SD data vs. LexisNexis data

Setup

- Number of epochs: 30 for LexisNexis, 25 for SD.
- No threshold applied to the confidence scores.

Results

- To achieve a quality that close to the quality of the uncompressed models, the used number of epochs should be higher than the number of epochs required for the uncompressed models.
- When comparing the two weak labels datasets, the results are comparable.
- Overall at the confidence threshold that will be defined for a precision of 0.8, the same recall is obtained.

The above results can be seen in the plots 3.3 for YASO, in 3.4 for MAMS and in 3.5 for SE16.

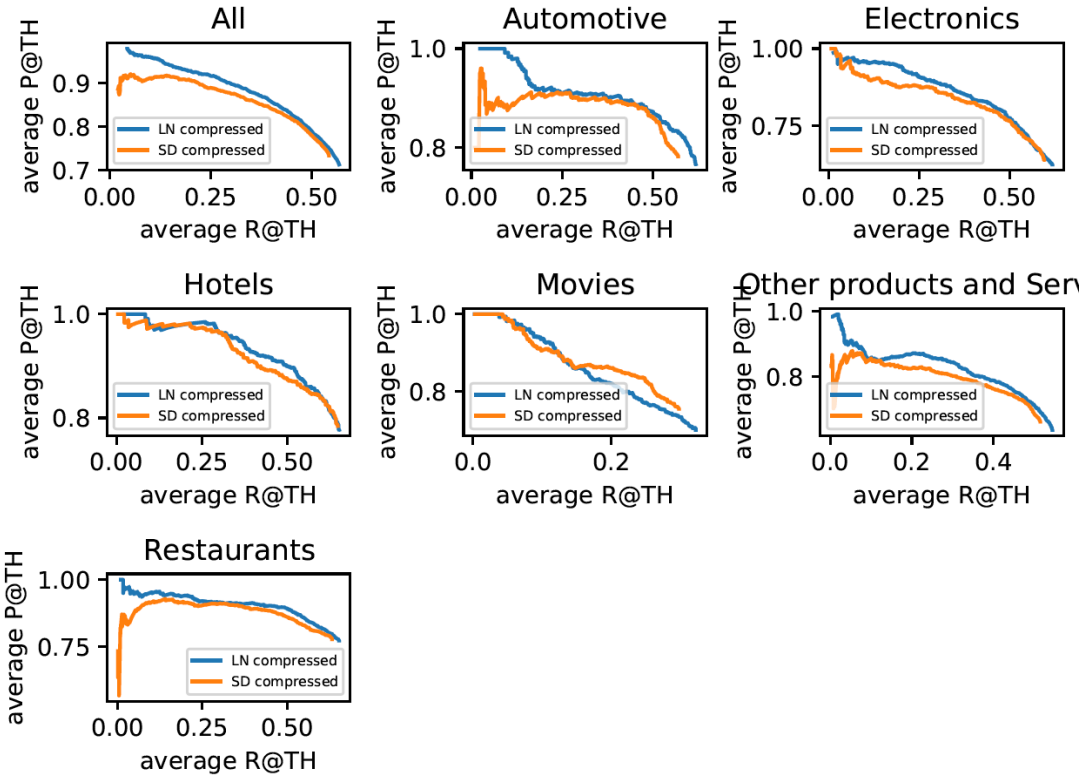

FIGURE 3.3: Compression with SD data vs. LexisNexis data on YASO dataset

Compression with LexisNexis data

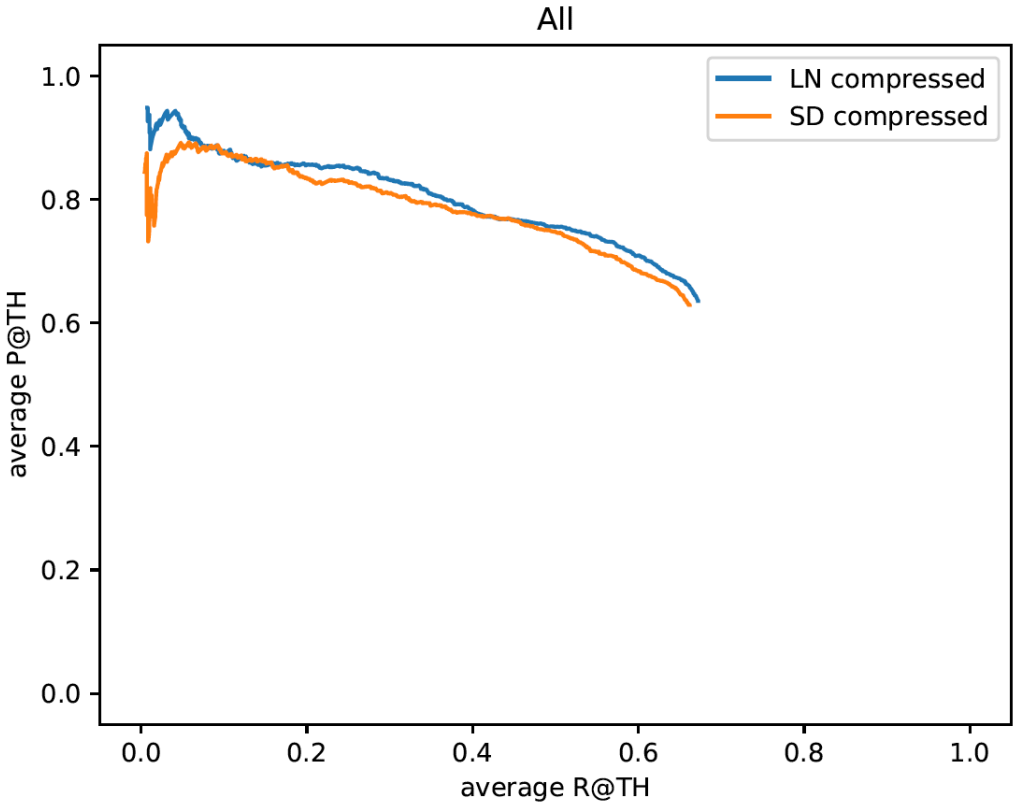

FIGURE 3.4: Compression with SD data vs. LexisNexis data on MAMS dataset

In the presented experiment, two compression techniques were applied.

Setup

- Layer removal: reduces the number of encoder layers.
- FFN pruning: Reduces the size of the intermediate layer within all encoders.

The best selected configuration included 8 layers with 30% pruning of the neurons in each intermediate layer.

Results

Applying compression (WatBERT comp.) degrades performance, as expected, but not by much, as can bee seen in the plots **??** and 3.7.

3.2.5 Delivered Models

This section presents the results for the delivered models, after these experiments. As far as the results are concerned, the results are all:

- From a single model (i.e not averages over multiple runs like the experimental phase)
- Thresholded using the confidence threshold selected per model, aiming at a precision of 80%.

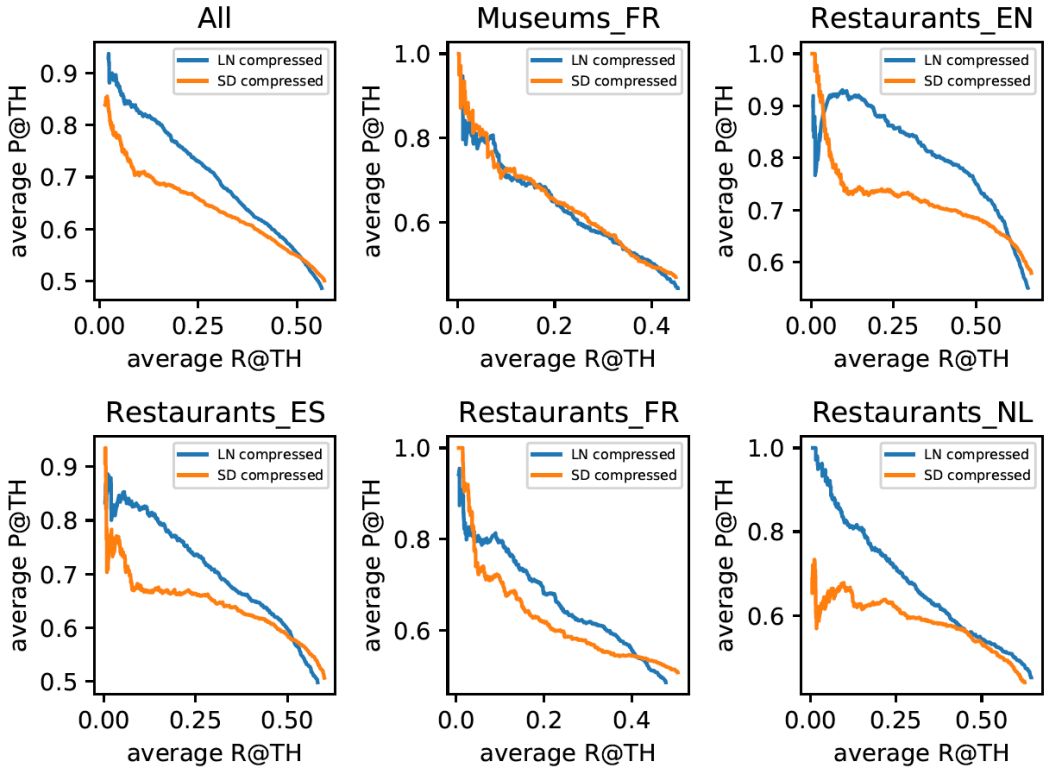

FIGURE 3.5: Compression with SD data vs. LexisNexis data on SE16 dataset

Quality

In the ensuing segment, we undertake a thorough evaluation of the diverse array of delivered BERT models used for targeted sentiment analysis. This section's primary goal is to rigorously examine these models' performance and functionality across various sectors and languages.

The models' resilience, the accuracy of their outputs, and their overall effectiveness in different application scenarios are carefully condsidered and probed into. The industries examined include automotive, electronics, hospitality, cinema, miscellaneous products, services, dining, and museums. These models' performance is evaluated through comprehensive metrics, including precision, recall, and F1 score, across multiple languages.

By dissecting and assessing the quality of each model, this section seeks to provide a broad understanding of their potential strengths and possible shortcomings. This intensive investigation aims to furnish insights that could guide the future refinement and creation of BERT models that are more robust, accurate, and efficient in the field of targeted sentiment analysis [19].

Distilled and WatBERT models

The following comparison includes:

- The current WatBERT stock model trained with LexisNexis data.

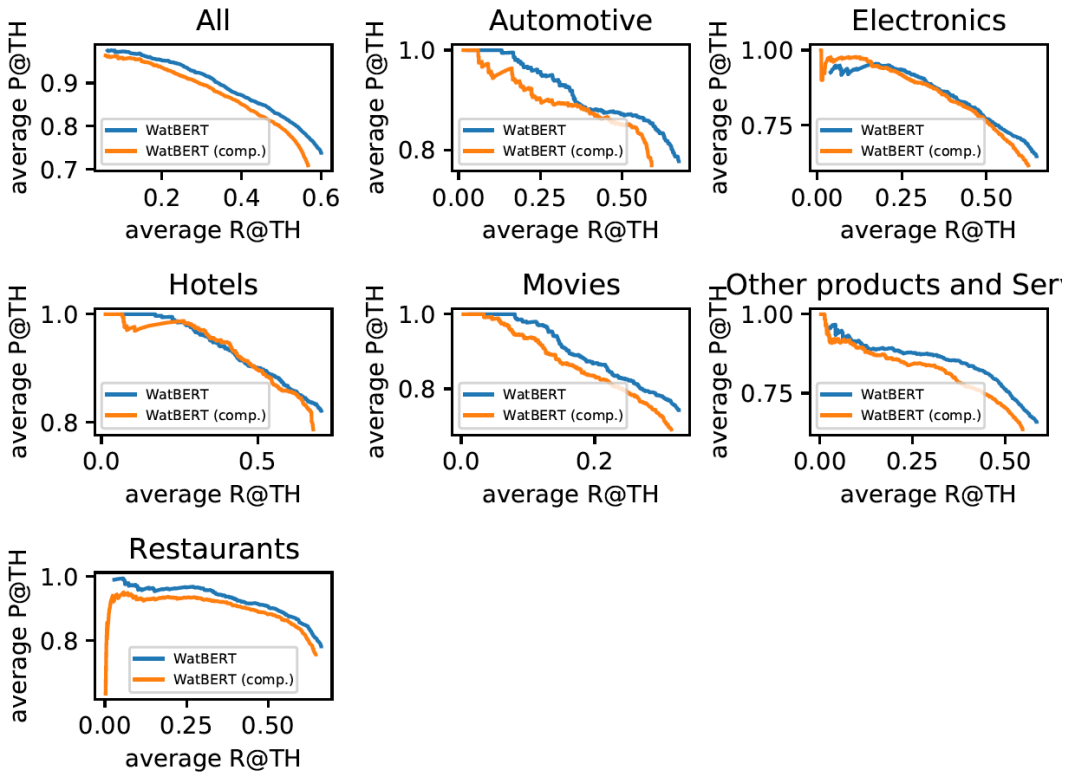

FIGURE 3.6: watBERT compression: YASO dataset

- A new Watbert models trained with SD data
- A new WatBERT distilled model trained with SD data.

Results

The results on YASO, MAMS and SE16 can be thoroughly analyzed through the corresponding tables 3.1, 3.2, 3.3 and the plots:

CPU-optimized models

1st Experiment

This experiment compares the quality of:

1. A WatBERT model trained on SD weak labels (named watbert).
2. A CPU-optimized variant of (1), named watbert-cpu.
3. A distilled WatBERT model trained on SD weak labels data, named distilled.
4. A CPU-optimized variant of (3), named distilled-cpu.

The results on YASO, MAMS and SE16 can be thoroughly analyzed through the corresponding tables 3.4, 3.5, 3.6 and the plots:

2nd Experiment

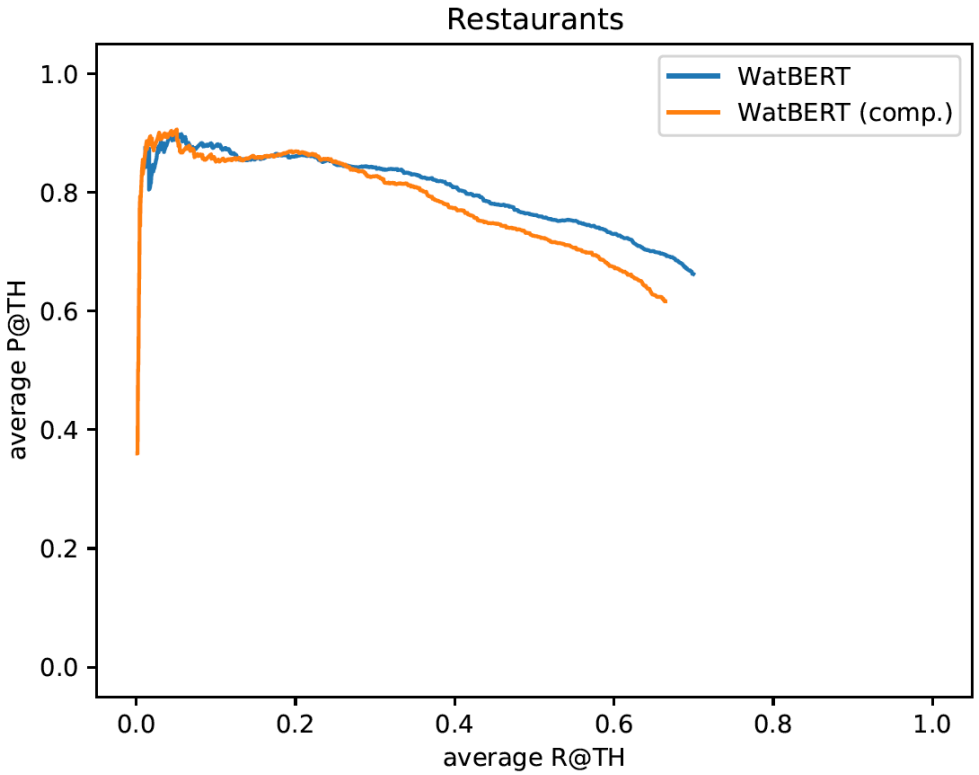

FIGURE 3.7: WatBERT compression: MAMS dataset

This experiment compares the quality of:

1. A WatBERT model trained on LexisNexisData (named stock).
2. A CPU-optimized variant of the (1) model, named stock-cpu.
3. A compressed WatBERT model trained on LexisNexis data, named compressed (in the graphs with an old name fast).
4. A CPU-optimized variant of (3), named compressed-cpu (in the graphs with an old name fast-cpu).

Results

The results on YASO, MAMS and SE16 can be thoroughly analyzed through the corresponding tables 3.7, 3.8, 3.9 and the corresponding plots 3.14, 3.15, 3.16:

Note: Results vary between languages. For example, in Spanish (ES) and Dutch (NL), the ‘stock-cpu‘ and ‘compressed-cpu‘ models yield about the same quality. For French (FR), the ‘stock-cpu‘ model yields better results in SE16.

3rd Experiment: A comparison to the Google-BERT based model

This comparison includes:

1. The delivered WatBERT stock model.

Domain	Model	Precision	Recall	F1
Automotive	watbert	0.81	0.54	0.65
Automotive	stock	0.86	0.57	0.68
Automotive	distilled	0.84	0.50	0.63
Electronics	watbert	0.71	0.59	0.65
Electronics	stock	0.71	0.57	0.64
Electronics	distilled	0.72	0.55	0.62
Hotels	watbert	0.84	0.63	0.72
Hotels	stock	0.86	0.63	0.72
Hotels	distilled	0.84	0.57	0.68
Movies	watbert	0.83	0.25	0.38
Movies	stock	0.84	0.26	0.40
Movies	distilled	0.85	0.22	0.35
Other products and Services	watbert	0.77	0.50	0.61
Other products and Services	stock	0.75	0.50	0.60
Other products and Services	distilled	0.72	0.43	0.54
Restaurants	watbert	0.84	0.61	0.71
Restaurants	stock	0.86	0.61	0.71
Restaurants	distilled	0.85	0.53	0.65

TABLE 3.1: watBERT vs stock vs distilled: Recall, Precision, F1 for each model on YASO dataset

Domain	Model	Precision	Recall	F1
Restaurants	watbert	0.70	0.55	0.62
Restaurants	stock	0.74	0.59	0.66
Restaurants	distilled	0.74	0.49	0.59

TABLE 3.2: watBERT vs stock vs distilled: Recall, Precision, F1 for each model on MAMS dataset

2. A WatBERT compressed model.
3. The existing TSA stock model: a compressed Google-BERT (G-BERT) based model.
4. An uncompressed TSA model (retrained from scratch).

Results

The graphs below (3.17, 3.18, 3.19) show the Precision and Recall obtained while varying the prediction threshold (the threshold itself is not plotted). The Y-Axis shows the Precision at a specific threshold (P@TH), while the X-Axis shows the Recall at a specific threshold (R@TH).

They are accompanied by the corresponding tables: 3.10, 3.11, 3.12

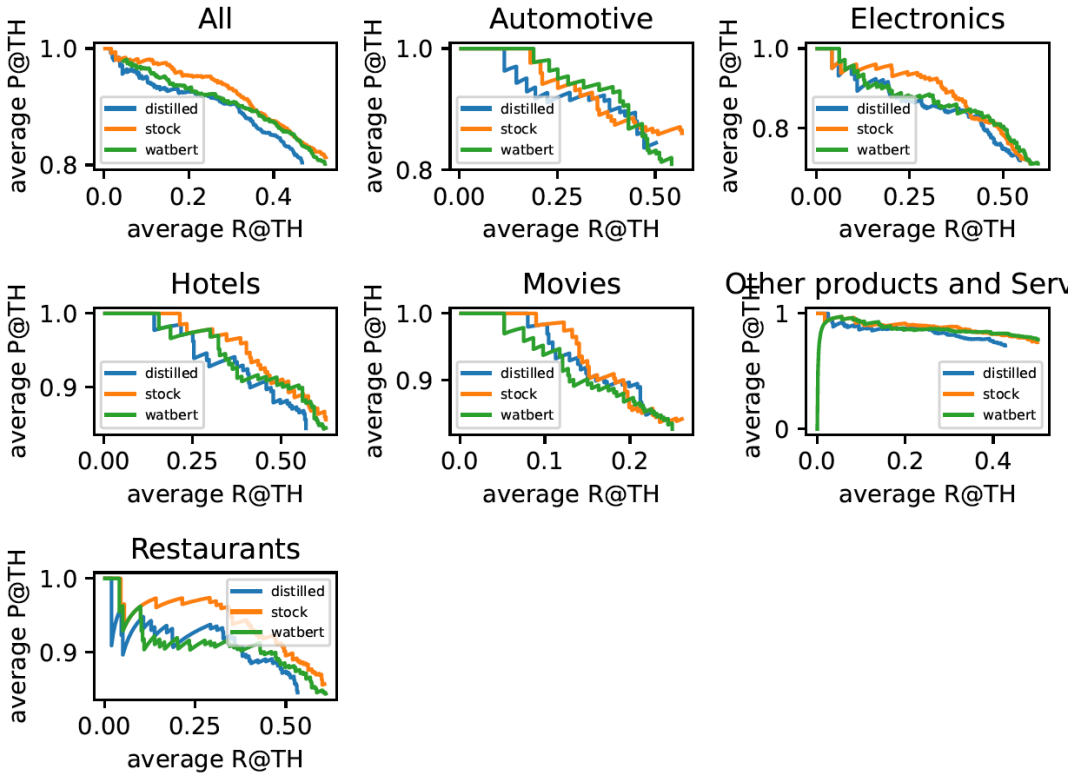

FIGURE 3.8: Delivered results for YASO

Even though the delivered model was not-trained on any non-English data, it is capable of accepting input in any non-English languages supported by multilingual BERT.

The model is tested on non-English TSA labeled data from the SemEval'16 (SE16) dataset. This is a dataset that is less suited to our needs, as it lacks the annotation of all sentiment-bearing terms as in YASO. For example, in English, SE16 contains the sentence "Don't leave the restaurant without it.", yet it lacks an annotation of the positive sentiment towards it. As a result, the performance (for English) on this dataset is lower than the reported results on YASO above. Still, SE16 provides some insight on the performance for non-English languages.

The graph 3.19 shows results using 5 SE16 datasets:

- Restaurant reviews in English (EN).
- Restaurant reviews in Spanish (ES)
- Restaurant reviews in Dutch (NL).
- Restaurant reviews in French (FR).
- Museum reviews in French (FR).

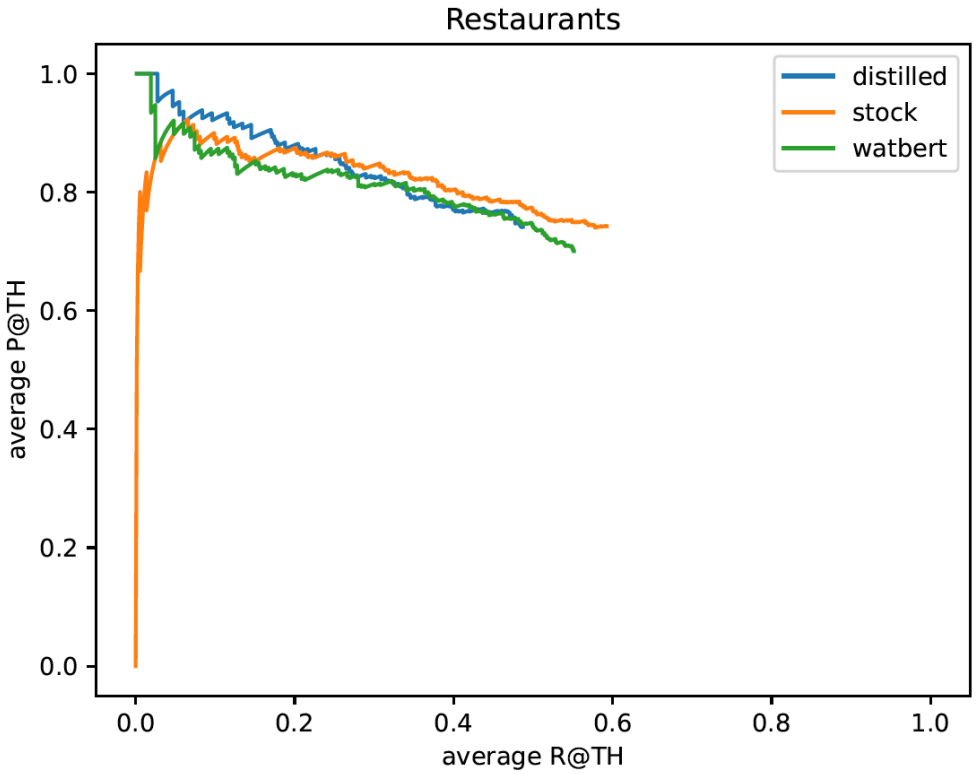

FIGURE 3.9: Delivered results for MAMS

3.2.6 Runtime

Method:

The tests were on English, using the Watson NLP runtime testing framework. Each experiment included running the runtime test for five times, and averaging the runtime over the last 4 runs. When utilizing a GPU, the first run was often slower, thus excluded from the report.

GPU

The figure 3.20, 3.21 show runtime measurements performed on a GPU.

Comparison to distilled WatBERT

This comparison pertains to the models trained on the new weakly labeled data SD.

Comparison to G-BERT based models

Compared models:

Language	Domain	Model	Precision	Recall	F1
French	Museums	watbert	0.55	0.48	0.51
French	Museums	stock	0.53	0.47	0.50
French	Museums	distilled	0.51	0.36	0.42
English	Restaurants	watbert	0.62	0.63	0.62
English	Restaurants	stock	0.65	0.65	0.65
English	Restaurants	distilled	0.64	0.56	0.60
Spanish	Restaurants	watbert	0.59	0.63	0.61
Spanish	Restaurants	stock	0.59	0.60	0.60
Spanish	Restaurants	distilled	0.59	0.52	0.55
French	Restaurants	watbert	0.59	0.52	0.55
French	Restaurants	stock	0.61	0.50	0.55
French	Restaurants	distilled	0.59	0.36	0.45
Dutch	Restaurants	watbert	0.54	0.65	0.59
Dutch	Restaurants	stock	0.52	0.64	0.58
Dutch	Restaurants	distilled	0.53	0.49	0.51

TABLE 3.3: watBERT vs stock vs distilled: Recall, Precision, F1 for each model on SE16

1. Green: The WatBERT model, trained on weak labels from LexisNexis (named stock in the table).
2. Red: A compressed WatBERT model (named fast).
3. Blue: A compressed G-BERT based model (an existing model).
4. Orange: An uncompressed G-BERT based model (retrained from scratch).

Results

- Without compression, WatBERT (38.35 throughput) is faster than G-BERT (29.05 throughput) by about 31%.
- Compression improves the runtime of WatBERT by 22% (46.51 vs. 38.35).
- The obtained runtime of compressed WatBERT (46.5) is comparable to the runtime of the compressed G-BERT model (50.3).
- The difference presumably stems from the compressed G-BERT employing 6 layers, while the compressed WatBERT has 8 layers. No additional layers were removed from WatBERT, since it had the impact of considerably degrading performance.
- G-BERT memory footprint is about 67% of the memory footprint of WatBERT (2.68 vs. 4.01). Without compression, that difference is 72% (3.20 vs. 4.44).

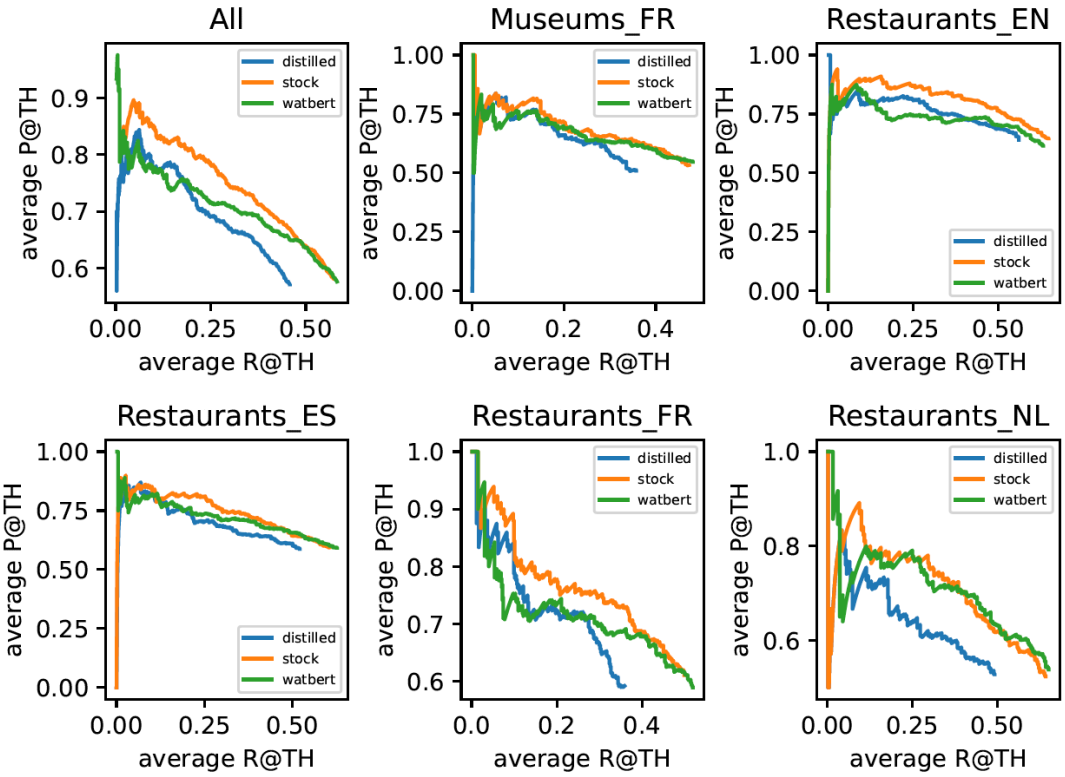

FIGURE 3.10: Delivered results for SE16

CPU-optimized models

The tables 3.22 and 3.23 show the runtime gains obtained by these models.

Comparison to distilled WatBERT

- This comparison is on models trained on weak labels from the SD dataset (in addition to the standard strong labels from MAMS and TSA-MD). This should not however be a factor that changes runtime. Comparing the watbert and watbert-cpu runtime results here (3.53 and 9.34 respectively) to the results of the stock and stock-cpu models, hey are indeed about the same.
- A speedup of more than 13x is obtained by the CPU-optimized distilled model (distil-cpu in the table).
- Similarly, distil-cpu consumes 14% of the memory of the standard watbert model.

Comparison to distilled models

The table 3.23 compares the runtime and memory consumption of the regular WatBERT-based model, and a compressed variant of it.

Domain	Model	Precision	Recall	F1
Automotive	watbert	0.81	0.54	0.65
Automotive	distilled	0.84	0.50	0.63
Automotive	watbert-cpu	0.82	0.49	0.61
Automotive	distilled-cpu	0.84	0.45	0.58
Electronics	watbert	0.71	0.59	0.65
Electronics	distilled	0.72	0.55	0.62
Electronics	watbert-cpu	0.73	0.52	0.61
Electronics	distilled-cpu	0.67	0.51	0.58
Hotels	watbert	0.84	0.63	0.72
Hotels	distilled	0.84	0.57	0.68
Hotels	watbert-cpu	0.86	0.54	0.67
Hotels	distilled-cpu	0.85	0.50	0.63
Movies	watbert	0.83	0.25	0.38
Movies	distilled	0.85	0.22	0.35
Movies	watbert-cpu	0.83	0.18	0.30
Movies	distilled-cpu	0.86	0.21	0.34
Other products and Services	watbert	0.77	0.50	0.61
Other products and Services	distilled	0.72	0.43	0.54
Other products and Services	watbert-cpu	0.77	0.44	0.56
Other products and Services	distilled-cpu	0.74	0.43	0.54
Restaurants	watbert	0.84	0.61	0.71
Restaurants	distilled	0.85	0.53	0.65
Restaurants	watbert-cpu	0.79	0.53	0.63
Restaurants	distilled-cpu	0.85	0.46	0.60

TABLE 3.4: CPU optimized model on YASO dataset

The models in this comparison were trained on weak labels from the LexisNexis dataset (in addition to the standard strong labels from MAMS and TSA-MD). On a CPU, the fastest model is a compressed model optimized for CPU use (fast-cpu in the table). It improves runtime by almost 7x, compared to the stock model. Still, utilizing a GPU is faster, up-to a speedup of 20x. Memory consumption is also lowest for the compressed CPU-optimized model (fast-cpu): only 37% of the memory required for the stock model on a CPU.

3.3 Key Takeaways

Here, we present the results obtained from applying BERT models to the IBM use case, including quantitative performance metrics and qualitative insights. We also discuss the model's strengths and weaknesses, as well as potential improvements and future work.

- WatBERT results on English are an improvement over the G-BERT results.
- For non-English, WatBERT achieves a significant boost over G-BERT.

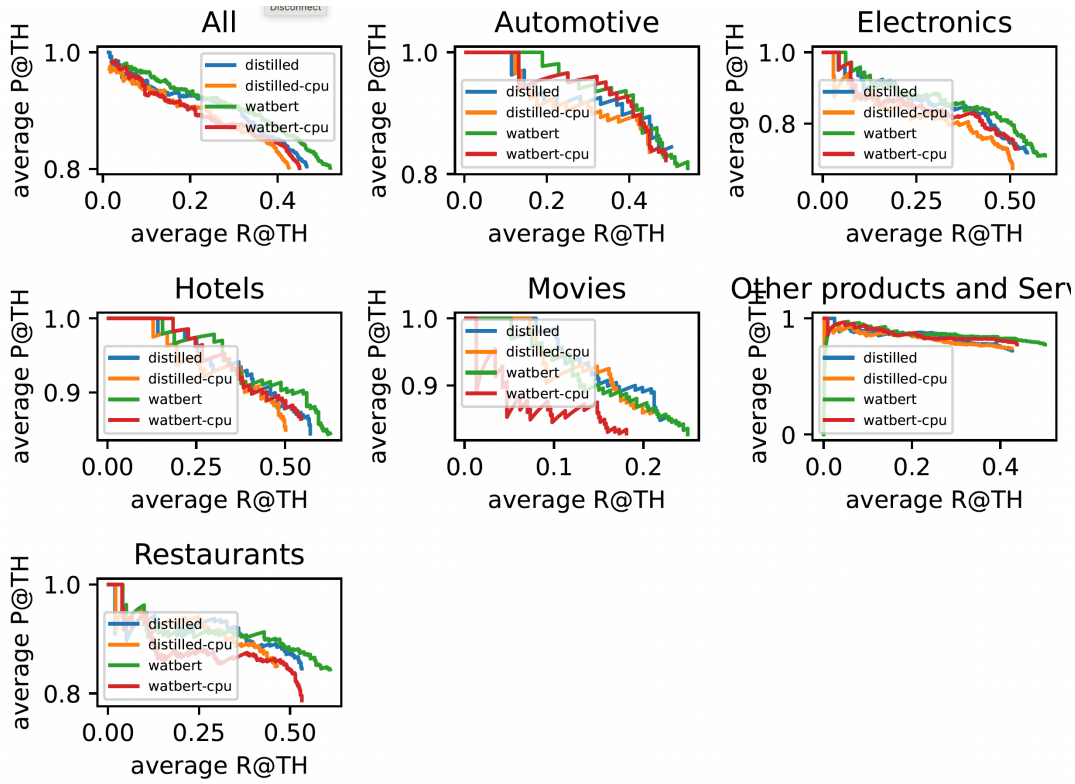

FIGURE 3.11: CPU watbert and distilled: 1st experiment on YASO dataset

Model	Precision	Recall	F1
watbert	0.70	0.55	0.62
distilled	0.74	0.49	0.59
watbert-cpu	0.61	0.49	0.55
distilled-cpu	0.75	0.39	0.52

TABLE 3.5: CPU optimized model on MAMS dataset

- The distilled WatBERT-based model performance is slightly lower than the performance of the WatBERT-based model, with an improvement in runtime of more than 50% (on a GPU).
- CPU-optimized models may be used when runtime is important, at some cost to quality. The watbert-cpu and distill-cpu models each provide a different tradeoff between quality and runtime.

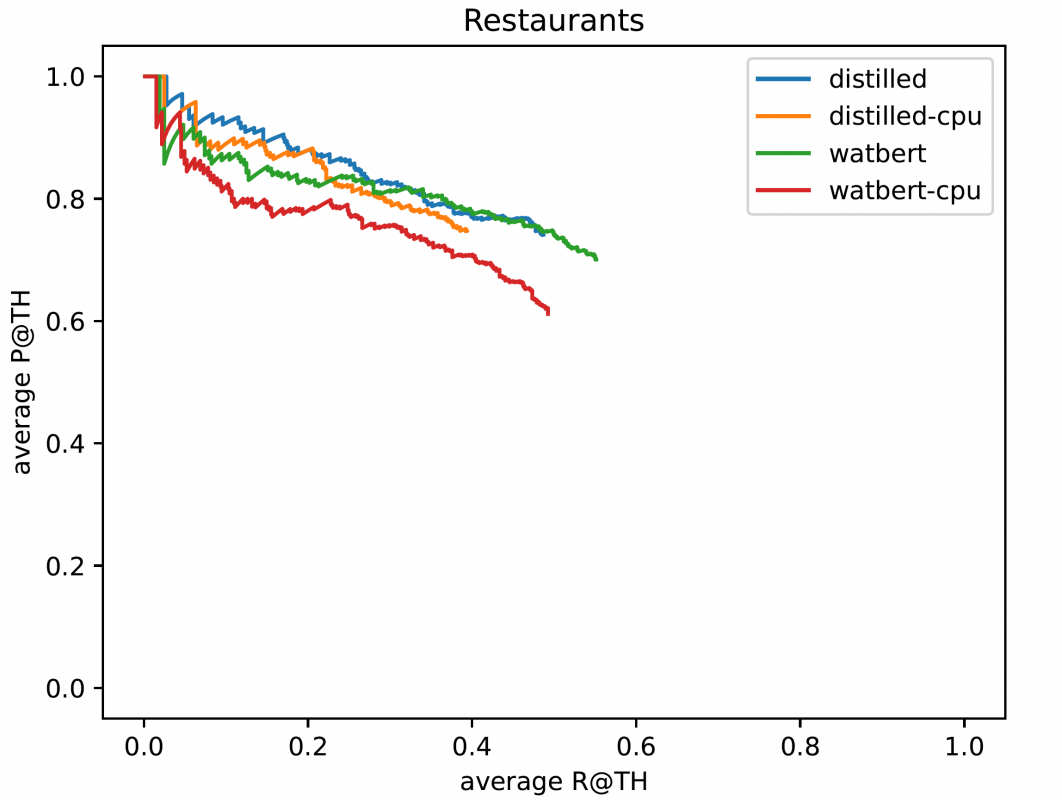

FIGURE 3.12: CPU watbert and distilled: 1st experiment on MAMS dataset

Domain	Language	Model	Precision	Recall	F1
Museums	French	watbert	0.55	0.48	0.51
Museums	French	distilled	0.51	0.36	0.42
Museums	French	watbert-cpu	0.53	0.41	0.46
Museums	French	distilled-cpu	0.52	0.35	0.42
Restaurants	English	watbert	0.62	0.63	0.62
Restaurants	English	distilled	0.64	0.56	0.60
Restaurants	English	watbert-cpu	0.61	0.61	0.61
Restaurants	English	distilled-cpu	0.64	0.52	0.58
Restaurants	Spanish	watbert	0.59	0.63	0.61
Restaurants	Spanish	distilled	0.59	0.52	0.55
Restaurants	Spanish	watbert-cpu	0.56	0.57	0.56
Restaurants	Spanish	distilled-cpu	0.61	0.49	0.54
Restaurants	French	watbert	0.59	0.52	0.55
Restaurants	French	distilled	0.59	0.36	0.45
Restaurants	French	watbert-cpu	0.53	0.44	0.48
Restaurants	French	distilled-cpu	0.60	0.31	0.41
Restaurants	Dutch	watbert	0.54	0.65	0.59
Restaurants	Dutch	distilled	0.53	0.49	0.51
Restaurants	Dutch	watbert-cpu	0.50	0.61	0.55
Restaurants	Dutch	distilled-cpu	0.55	0.46	0.50

TABLE 3.6: CPU optimized model on SE16 dataset

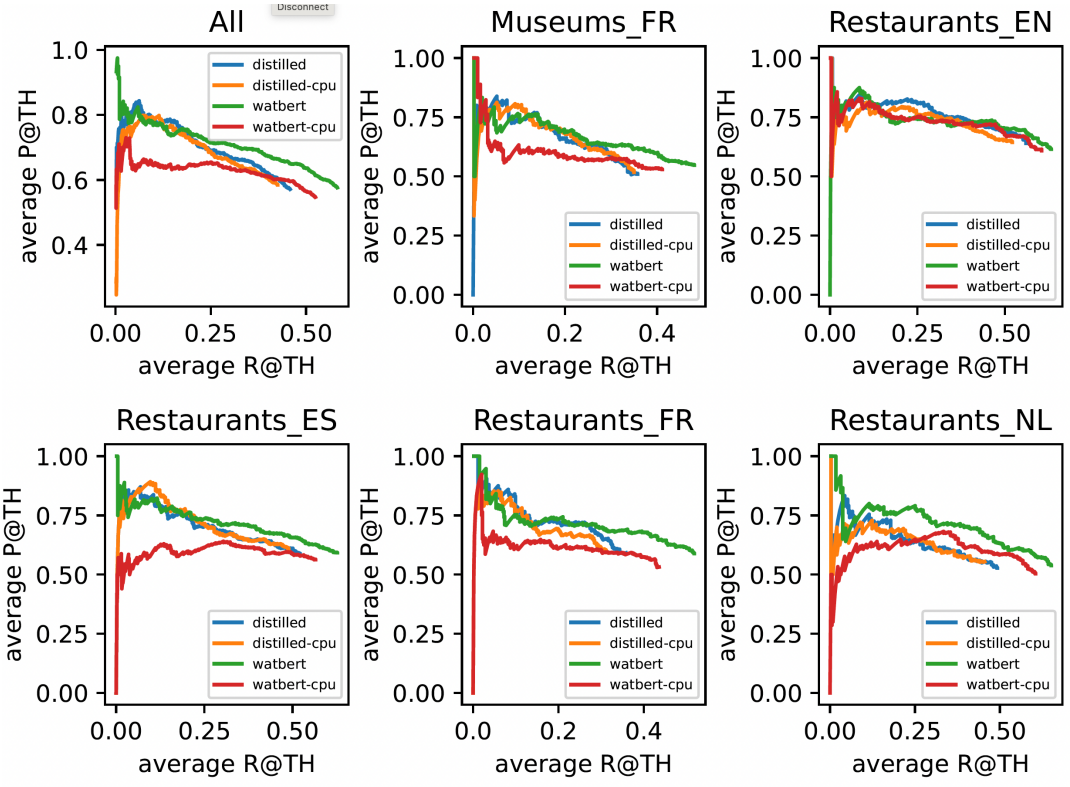

FIGURE 3.13: CPU watbert and distilled: 1st experiment on SE16 dataset

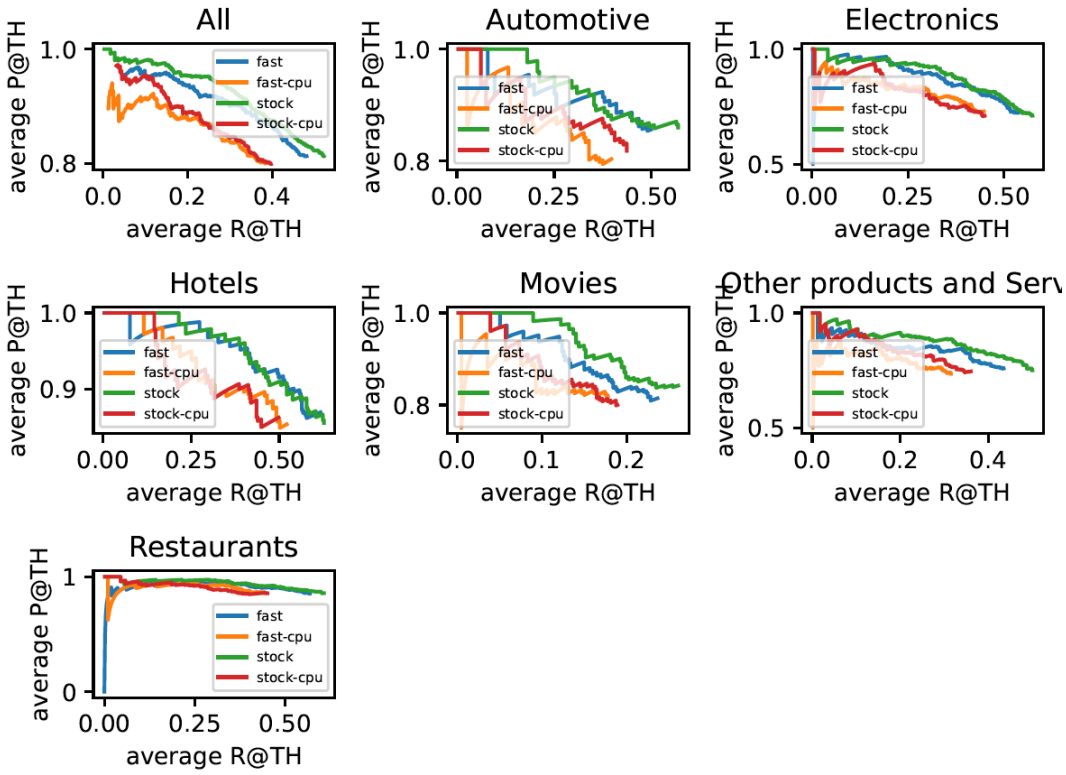

FIGURE 3.14: CPU watbert and distilled: 2nd experiment on YASO dataset

Domain	**Model**	**Precision**	**Recall**	**F1**
Automotive	stock	0.86	0.57	0.68
Automotive	compressed	0.86	0.50	0.63
Automotive	stock-cpu	0.82	0.44	0.57
Automotive	compressed-cpu	0.80	0.40	0.53
Electronics	stock	0.71	0.57	0.64
Electronics	compressed	0.72	0.54	0.62
Electronics	stock-cpu	0.71	0.45	0.55
Electronics	compressed-cpu	0.73	0.45	0.55
Hotels	stock	0.86	0.63	0.72
Hotels	compressed	0.87	0.62	0.72
Hotels	stock-cpu	0.86	0.50	0.63
Hotels	compressed-cpu	0.85	0.52	0.65
Movies	stock	0.84	0.26	0.40
Movies	compressed	0.81	0.24	0.37
Movies	stock-cpu	0.80	0.19	0.31
Movies	compressed-cpu	0.82	0.18	0.29
Other products and Services	stock	0.75	0.50	0.60
Other products and Services	compressed	0.76	0.43	0.55
Other products and Services	stock-cpu	0.75	0.36	0.49
Other products and Services	compressed-cpu	0.74	0.32	0.44
Restaurants	stock	0.86	0.61	0.71
Restaurants	compressed	0.85	0.57	0.68
Restaurants	stock-cpu	0.86	0.45	0.59
Restaurants	compressed-cpu	0.86	0.44	0.59

TABLE 3.7: Performance metrics of stock vs compressed vs stock-cpu vs compressed-cpu on YASO dataset

Model	**Precision**	**Recall**	**F1**
stock	0.74	0.59	0.66
compressed	0.71	0.53	0.61
stock-cpu	0.78	0.43	0.56
compressed-cpu	0.70	0.38	0.49

TABLE 3.8: Performance metrics of stock vs compressed vs stock-cpu vs compressed cpu on MAMS dataset

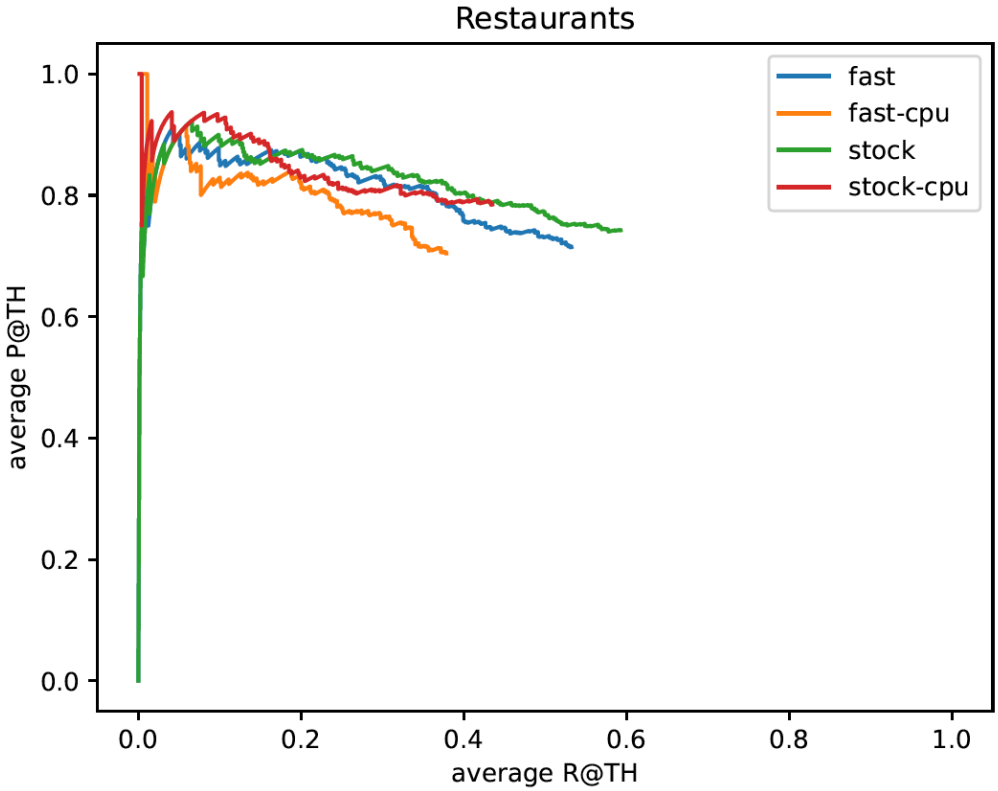

FIGURE 3.15: CPU watbert and distilled: 2nd experiment on MAMS dataset

Domain	Language	Model	Precision	Recall	F1
Museums	French	stock	0.53	0.47	0.50
Museums	French	compressed	0.53	0.33	0.40
Museums	French	stock-cpu	0.47	0.27	0.34
Museums	French	compressed-cpu	0.45	0.17	0.24
Restaurants	English	stock	0.65	0.65	0.65
Restaurants	English	compressed	0.66	0.61	0.64
Restaurants	English	stock-cpu	0.65	0.47	0.54
Restaurants	English	compressed-cpu	0.67	0.52	0.59
Restaurants	Spanish	stock	0.59	0.60	0.60
Restaurants	Spanish	compressed	0.58	0.45	0.51
Restaurants	Spanish	stock-cpu	0.52	0.39	0.45
Restaurants	Spanish	compressed-cpu	0.56	0.29	0.38
Restaurants	French	stock	0.61	0.50	0.55
Restaurants	French	compressed	0.57	0.36	0.44
Restaurants	French	stock-cpu	0.58	0.29	0.39
Restaurants	French	compressed-cpu	0.50	0.20	0.28
Restaurants	Dutch	stock	0.52	0.64	0.58
Restaurants	Dutch	compressed	0.54	0.50	0.51
Restaurants	Dutch	stock-cpu	0.50	0.40	0.45
Restaurants	Dutch	compressed-cpu	0.52	0.35	0.42

TABLE 3.9: Performance metrics of stock vs compressed vs stock-cpu vs compressed cpu on SE16 dataset

Domain	Model	Precision	Recall	F1
Automotive	WatBERT	0.86	0.57	0.68
Electronics	WatBERT	0.71	0.57	0.64
Hotels	WatBERT	0.86	0.63	0.72
Movies	WatBERT	0.84	0.26	0.40
Other products and Services	WatBERT	0.75	0.50	0.60
Restaurants	WatBERT	0.86	0.61	0.71
Automotive	WatBERT comp.	0.86	0.50	0.63
Electronics	WatBERT comp.	0.72	0.54	0.62
Hotels	WatBERT comp.	0.87	0.62	0.72
Movies	WatBERT comp.	0.81	0.24	0.37
Other products and Services	WatBERT comp.	0.76	0.43	0.55
Restaurants	WatBERT comp.	0.85	0.57	0.68

TABLE 3.10: Precision, Recall, F1 on domains and models for YASO dataset

Model	Precision	Recall	F1
WatBERT	0.74	0.59	0.66
WatBERT comp.	0.71	0.53	0.61

TABLE 3.11: Precision, Recall, F1 on domain and models for MAMS dataset

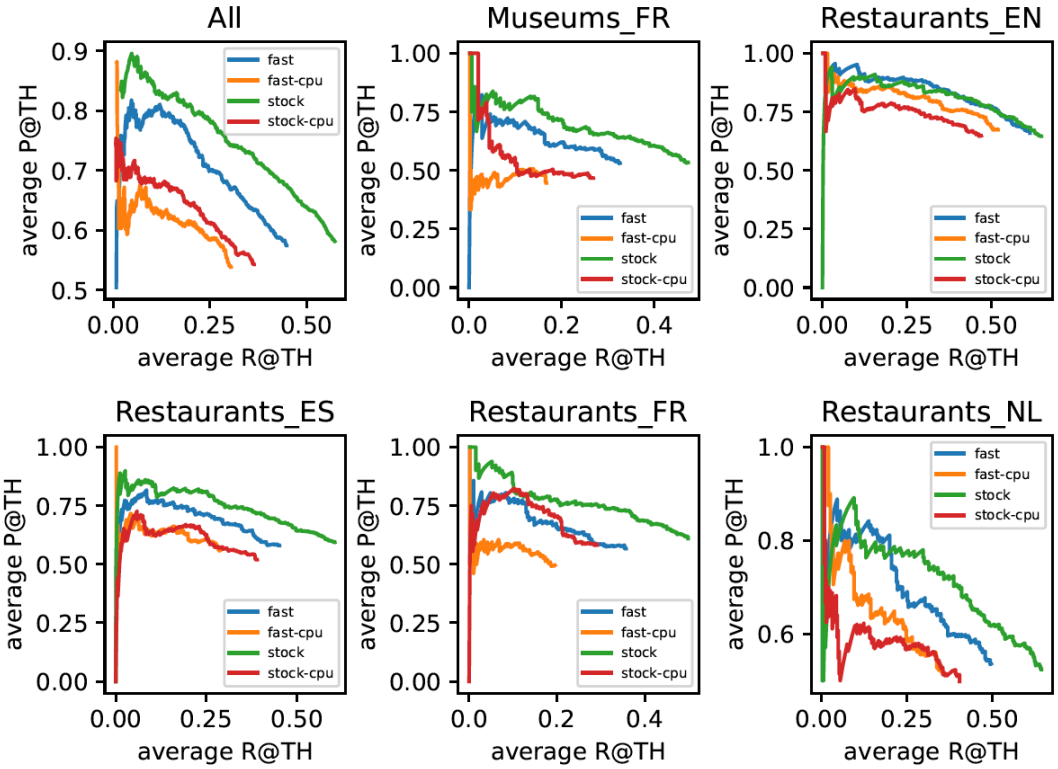

FIGURE 3.16: CPU watbert and distilled: 2nd experiment on SE16 dataset

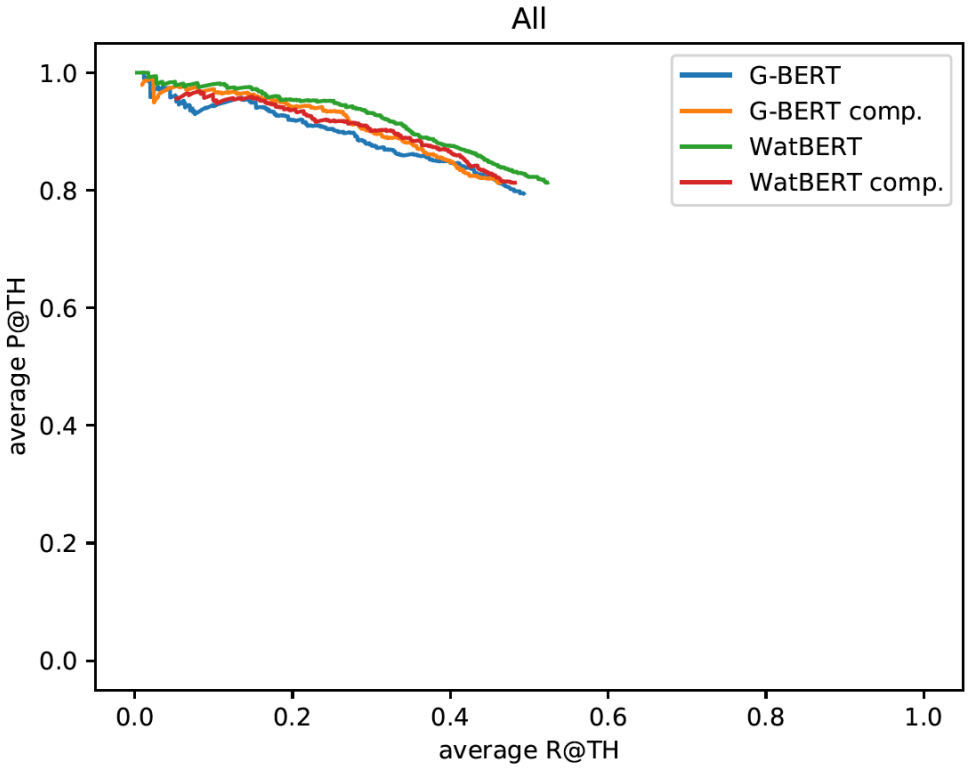

FIGURE 3.17: Delivered Model YASO

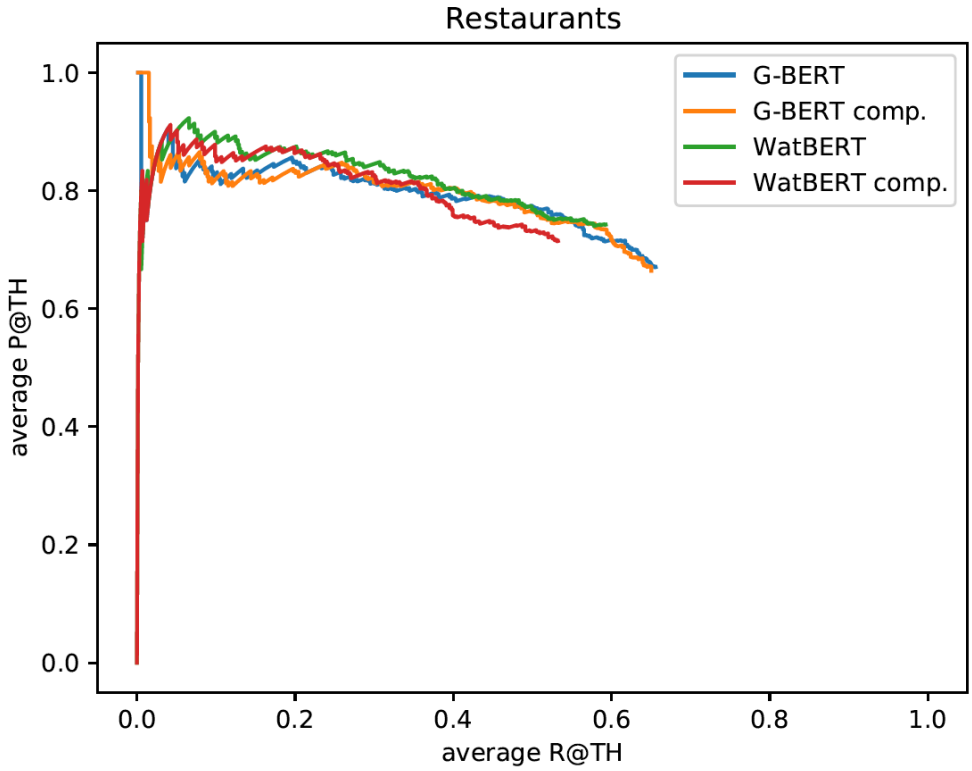

FIGURE 3.18: Delivered Model MAMS

Dataset	Model	Precision	Recall	F1
Restaurants_FR	WatBERT	0.61	0.50	0.55
Restaurants_ES	WatBERT	0.59	0.60	0.60
Restaurants_NL	WatBERT	0.52	0.64	0.58
Museums_FR	WatBERT	0.53	0.47	0.50
Restaurants_EN	WatBERT	0.65	0.65	0.65
Restaurants_FR	WatBERT comp.	0.57	0.36	0.44
Restaurants_ES	WatBERT comp.	0.58	0.45	0.51
Restaurants_NL	WatBERT comp.	0.54	0.50	0.51
Museums_FR	WatBERT comp.	0.53	0.33	0.40
Restaurants_EN	WatBERT comp.	0.66	0.61	0.64

TABLE 3.12: Precision, Recall, F1 on domain and models for SE16 dataset

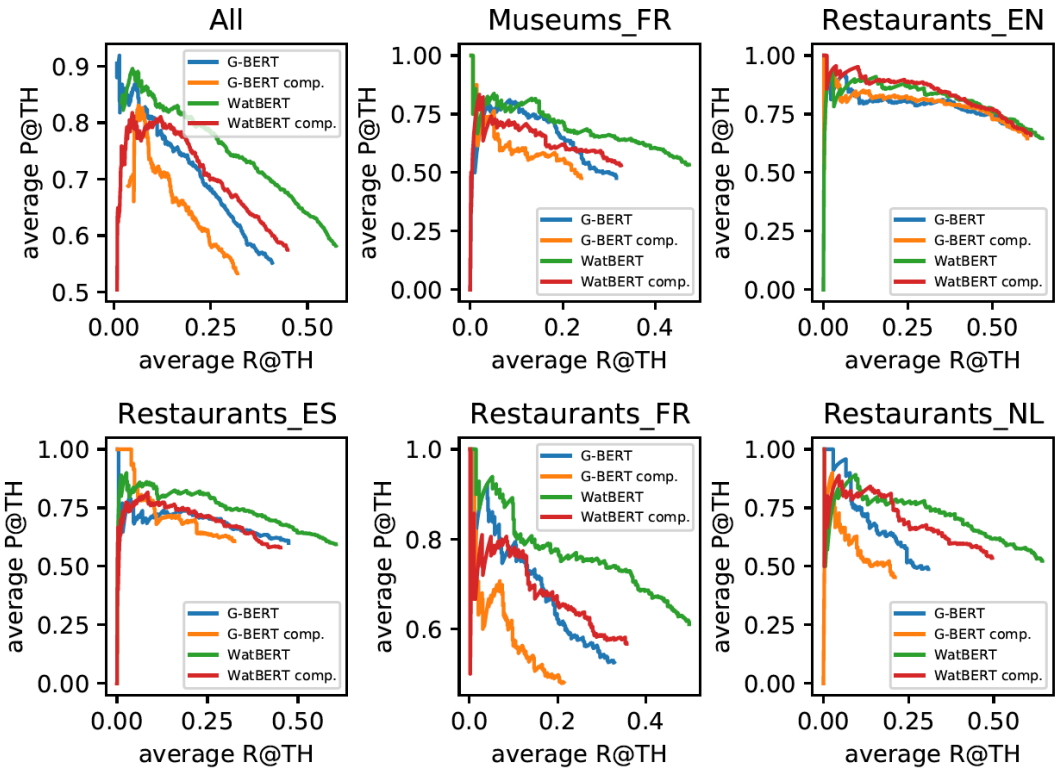

FIGURE 3.19: Delivered Model SE16

Hardware	Model description	Function Throughput (Kcodepoints/sec)	Speedup	Memory Footprint (GB)	Memory consumption
GPU	watbert	40.10	1.00	3.04	100%
GPU	distil	62.97	1.57	1.66	55%

FIGURE 3.20: GPU runtime: distilled vs watbert

GPU: V100	Compressed	Platform	PLM	Function Throughput (Kcodepoints/sec)	Memory Footprint (GB)	GPU Peak Memory Footprint (GB)
	Yes	TensorFlow	Goggle BERT	50.35	2.68	0.47
	Yes	PyTorch	WatBERT	46.51	4.01	0.96
	No	TensorFlow	Goggle BERT	29.05	3.20	0.67
	No	PyTorch	WatBERT	38.35	4.44	1.15

FIGURE 3.21: GPU runtime: delivered models

Hardware	Model description	Distilled?	CPU optimized	Function Throughput (Kcodepoints/sec)	Speedup	Memory Footprint (GB)	Memory consumption
CPU	watbert	No	No	3.53	1.00	2.45	100%
CPU	distil	Yes	No	20.08	5.68	1.46	59%
CPU	watbert-cpu	No	Yes	9.34	2.64	1.33	54%
CPU	distil-cpu	Yes	Yes	47.57	13.46	0.35	14%

FIGURE 3.22: Runtime and Memory Consumption

Hardware	Model description	Compressed	CPU optimized	Function Throughput (Kcodepoints/sec)	Speedup	Memory Footprint (GB)	Memory consumption
CPU	stock	No	No	2.28	1.00	2.46	100%
CPU	fast	Yes	No	3.88	1.70	2.14	87%
CPU	stock-cpu	No	Yes	9.55	4.20	1.30	53%
CPU	fast-cpu	Yes	Yes	15.58	6.85	0.91	37%
GPU	stock	No	No	38.35	16.86	4.44	181%
GPU	fast	Yes	No	46.51	20.44	4.01	163%

FIGURE 3.23: Runtime and Memory Consumption

Chapter 4

Conclusions

The conclusion of this thesis provides a reflection on the broad topic of sentiment analysis, specifically focusing on multilingual sentiment analysis using BERT and its variants. It recapitulates the significant challenges in this field, the evolution of transformer models, and the potential of BERT models in addressing multilingual sentiment analysis.

Through a comprehensive literature review, the thesis shed light on a variety of techniques used in sentiment analysis, highlighting the paradigm shift towards transformer models, such as BERT and its variants. A deep dive into these models revealed their advantages, especially their pre-training capabilities, and their potential for multilingual sentiment analysis. It also underscored the active research and emerging trends in this domain, suggesting an optimistic future with room for improvement and innovation.

The application of these theories and models was exemplified in the IBM use case, a focused project on targeted sentiment analysis. The project's methodology, from data collection, training, evaluation to experimentation, demonstrated a practical application of BERT models in a real-world context. The use case reinforced the viability of these models, their quality, and also addressed computational considerations, such as runtime optimization for GPU and CPU-optimized models. The success of this use case and the delivered models underscores the transformative potential of BERT for multilingual sentiment analysis.

However, as every solution brings its own set of challenges, there are aspects of this work that call for future research. The limitations of BERT models, especially their resource-intensive nature, present an area of potential improvement. Moreover, while the IBM use case provided valuable insights, further studies and use cases are necessary to broaden the understanding and applicability of these models across different languages and contexts.

To sum up, this thesis highlighted the transformative potential of BERT models in the domain of multilingual sentiment analysis, as well as the areas of improvement. It contributes to the ongoing discourse on leveraging AI for understanding and interpreting human sentiment across languages, paving the way for further research.

The journey to fully understanding and implementing multilingual sentiment analysis may still be a long one, but this work has added a crucial piece to the puzzle. It is hoped that the knowledge gained and shared through this thesis will serve as a foundation for further studies and improvements in the field.

Bibliography

[1] Y. Cao, R. Xu, and Y. Lin. "An Overview of Deep Learning-based Sentiment Analysis: Techniques, Challenges, and Interpretability". In: *Information Fusion* (2021).

[2] Kevin Clark et al. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. 2020. arXiv: 2003.10555 [cs.CL]. URL: https://arxiv.org/abs/2003.10555.

[3] A. Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: (2020). URL: https://arxiv.org/abs/1911.02116.

[4] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (2018). URL: https://arxiv.org/abs/1810.04805.

[5] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *arXiv preprint arXiv:1801.06146* (2018). URL: https://arxiv.org/abs/1801.06146.

[6] Sarthak Jain et al. "Learning to Faithfully Rationalize by Construction". In: *arXiv preprint arXiv:2005.00115* (2020). URL: https://arxiv.org/abs/2005.00115.

[7] Xiaoqi Jiao et al. *TinyBERT: Distilling BERT for Natural Language Understanding*. 2019. arXiv: 1909.10351 [cs.CL]. URL: https://arxiv.org/abs/1909.10351.

[8] Mandar Joshi et al. *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. 2019. arXiv: 1907.10529 [cs.CL]. URL: https://arxiv.org/abs/1907.10529.

[9] Dan Kondratyuk and Milan Straka. "75 Languages, 1 Model: Parsing Universal Dependencies Universally". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019. URL: https://aclanthology.org/D19-1279.pdf.

[10] Zhenzhong Lan et al. *Albert: A Lite Bert for Self-supervised Learning of Language Representations*. 2019. arXiv: 1909.11942 [cs.CL]. URL: https://arxiv.org/abs/1909.11942.

[11] X. Li et al. "A Survey on Sentiment Analysis: Techniques, Challenges, and Applications". In: *Knowledge-Based Systems* (2020). URL: https://link.springer.com/article/10.1007/s10462-022-10144-1.

[12] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[13] R. Martınez-Tomas M. Taboada and J.M. Ferrandez. "New perspectives on the application of expert systems". In: *The Journal of Knowledge Engineering* (2011). DOI: 10.1111/j.1468-0394.2011.00599.x.

[14] T. Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems* (2013). URL: https://arxiv.org/abs/1310.4546.

[15] B. Pang and L. Lee. "Opinion Mining and Sentiment Analysis". In: *Foundations and Trends in Information Retrieval* (2008).

[16] Telmo Pires, Eva Schlinger, and Dan Garrette. *How multilingual is Multilingual BERT?* 2019. arXiv: 1906.01502 [cs.CL]. URL: https://arxiv.org/abs/1906.01502.

[17] S. Ruder, I. Vulic, and A. Søgaard. "A Survey of Cross-lingual Word Embedding Models". In: *Journal of Artificial Intelligence Research* 65 (2019). URL: https://arxiv.org/abs/1706.04902.

[18] V. Sanh et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper, and Lighter". In: (2019). URL: https://arxiv.org/abs/1910.01108.

[19] C. Sun, X. Qiu, and X. Huang. "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence". In: (2019). arXiv: 1903.09588 [cs.CL].

[20] D. Tang, B. Qin, and T. Liu. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification". In: 2015. URL: https://aclanthology.org/D15-1167.

[21] A. Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems* (2017).

[22] A. Wang et al. "Cross-lingual Language Model Pretraining". In: (2020). URL: https://arxiv.org/abs/2001.08210.

[23] Junjie Ye et al. "Sentiment-aware multimodal pre-training for multimodal sentiment analysis". In: *Knowledge-Based Systems* (2022). DOI: 10.1016/j.knosys.2022.110021. URL: https://doi.org/10.1016/j.knosys.2022.110021.

[24] H. Zhang et al. "Multilingual Sentiment Analysis: A Survey". In: *Information Processing Management* (2021).