UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS

Application of Pre-training and Fine-tuning AI Models to Machine Translation:
A Case Study of Multilingual Text Classification in Baidu

Guangxin Wei

Relatório de Estágio orientado pela Professora Doutora Helena Gorete Silva Moniz e coorientado pelo Dr. Bo Yin, Gestor de Produto na Baidu, especialmente elaborado para a obtenção do grau de Mestre em Tradução.

2022

Acknowledgements

I would like to take this opportunity to thank my supervisors, Professor Helena Moniz and Dr. Bo Yin, whose knowledge, expertise and support were invaluable and decisive for the present work, and whose insightful feedback has allowed me to develop and sharpen my perspective as a young Researcher, pushing me to a higher level. I am also deeply grateful to the full Baidu´s NLP team who have always been prompt and ready for valuable insights and new ideas.

Moreover, I am very thankful for all support and encouragement from my peers, family and friends, who have stood by me all along the way. Thank You.

Abstract

With the development of international information technology, we are producing a huge amount of information all the time. The processing ability of information in various languages is gradually replacing information and becoming a rarer resource. How to obtain the most effective information in such a large and complex amount of multilingual textual information is a major goal of multilingual information processing.

Multilingual text classification helps users to break the language barrier and accurately locate the required information and triage information. At the same time, the rapid development of the Internet has accelerated the communication among users of various languages, giving rise to a large number of multilingual texts, such as book and movie reviews, online chats, product introductions and other forms, which contain a large amount of valuable implicit information and urgently need automated tools to categorize and process those multilingual texts.

This work describes the Natural Language Process (NLP) sub-task known as Multilingual Text Classification (MTC) performed within the context of Baidu, a Chinese leading AI company with a strong Internet base, whose NLP division led the industry in deep learning technology to go online in Machine Translation (MT) and search. Multilingual text classification is an important module in NLP machine translation and a basic module in NLP tasks. It can be applied to many fields, such as Fake Reviews Detection, News Headlines Categories Classification, Analysis of positive and negative reviews and so on.

In the following work, we will first define the AI model paradigm of 'pre-training and fine-tuning' in deep learning in the Baidu NLP department. Then investigated the application scenarios of multilingual text classification. Most of the text classification systems currently available in the Chinese market are designed for a single language, such as Alibaba's text classification system. If users need to classify texts of the same category in multiple languages, they need to train multiple single text classification systems and then classify them one by one.

However, many internationalized products do not have a single text language, such as AliExpress cross-border e-commerce business, Airbnb B&B business, etc. Industry needs to understand and classify users' reviews in various languages, and have conducted in-depth statistics and marketing strategy development, and multilingual text classification is particularly important in this scenario.

Therefore, we focus on interpreting the methodology of multilingual text classification model of machine translation in Baidu NLP department, and capture sets of multilingual data of reviews, news headlines and other data for manual classification and labeling, use the labeling results for fine-tuning of multilingual text classification model, and output the quality evaluation data of Baidu multilingual text classification model after fine-tuning. We will discuss if the pre-training and fine-tuning of the large model can substantially improve the quality and performance of multilingual text classification.

Finally, based on the machine translation-multilingual text classification model, we derive the application method of pre-training and fine-tuning paradigm in the current cutting-edge deep learning AI model under the NLP system and verify the generality and cutting-edge of the pre-training and fine-tuning paradigm in the deep learning-intelligent search field.

Keywords: Machine Translation; Multilingual Text Classification; AI Model; Quality Assessment; Deep Learning; Pre-training and Fine-tuning

Resumo

Com o desenvolvimento da tecnologia de informação internacional, estamos sempre a produzir uma enorme quantidade de informação e o recurso mais escasso já não é a informação, mas a capacidade de processar informação em cada língua. A maior parte da informação multilingue é expressa sob a forma de texto. Como obter a informação mais eficaz numa quantidade tão considerável e complexa de informação textual multilingue é um dos principais objetivos do processamento de informação multilingue.

A classificação de texto multilingue ajuda os utilizadores a quebrar a barreira linguística e a localizar com precisão a informação necessária e a classificá-la. Ao mesmo tempo, o rápido desenvolvimento da Internet acelerou a comunicação entre utilizadores de várias línguas, dando origem a um grande número de textos multilingues, tais como críticas de livros e filmes, chats, introduções de produtos e outros distintos textos, que contêm uma grande quantidade de informação implícita valiosa e necessitam urgentemente de ferramentas automatizadas para categorizar e processar esses textos multilingues.

Este trabalho descreve a subtarefa do Processamento de Linguagem Natural (PNL) conhecida como Classificação de Texto Multilingue (MTC), realizada no contexto da Baidu, uma empresa chinesa líder em IA, cuja equipa de PNL levou a indústria em tecnologia baseada em aprendizagem neuronal a destacar-se em Tradução Automática (MT) e pesquisa científica. A classificação multilingue de textos é um módulo importante na tradução automática de PNL e um módulo básico em tarefas de PNL. A MTC pode ser aplicada a muitos campos, tais como análise de sentimentos multilingues, categorização de notícias, filtragem de conteúdos indesejados (do inglês *spam*), entre outros.

Neste trabalho, iremos primeiro definir o paradigma do modelo AI de 'pré-treino e afinação' em aprendizagem profunda no departamento de PNL da Baidu. Em seguida, realizaremos a pesquisa sobre outros produtos no mercado com capacidade de classificação de texto — a classificação de texto levada a cabo pela Alibaba. Após a pesquisa, verificamos que a maioria dos sistemas de classificação de texto atualmente disponíveis no mercado chinês são concebidos para uma única língua, tal

como o sistema de classificação de texto Alibaba. Se os utilizadores precisarem de classificar textos da mesma categoria em várias línguas, precisam de aplicar vários sistemas de classificação de texto para cada língua e depois classificá-los um a um.

No entanto, muitos produtos internacionalizados não têm uma única língua de texto, tais como AliExpress comércio eletrónico transfronteiriço, Airbnb B&B business, etc. A indústria precisa compreender e classificar as revisões dos utilizadores em várias línguas. Esta necessidade conduziu a um desenvolvimento aprofundado de estatísticas e estratégias de marketing, e a classificação de textos multilingues é particularmente importante neste cenário.

Desta forma, concentrar-nos-emos na interpretação da metodologia do modelo de classificação de texto multilingue da tradução automática no departamento de PNL Baidu. Colhemos para o efeito conjuntos de dados multilingues de comentários e críticas, manchetes de notícias e outros dados para classificação manual, utilizamos os resultados dessa classificação para o aperfeiçoamento do modelo de classificação de texto multilingue e produzimos os dados de avaliação da qualidade do modelo de classificação de texto multilingue da Baidu. Discutiremos se o pré-treino e o aperfeiçoamento do modelo podem melhorar substancialmente a qualidade e o desempenho da classificação de texto multilingue. Finalmente, com base no modelo de classificação de texto multilingue de tradução automática, derivamos o método de aplicação do paradigma de pré-formação e afinação no atual modelo de IA de aprendizagem profunda de ponta sob o sistema de PNL, e verificamos a robustez e os resultados positivos do paradigma de pré-treino e afinação no campo de pesquisa de aprendizagem profunda.

Palavras-chave: Tradução automática; Classificação de Texto Multilingue; Modelo de IA; Avaliação da Qualidade; Aprendizagem Profunda; Pré-treino e Técnicas de Afinação de Parâmetros

Index

Chapter 1

1 Introduction and motivation

1.1 Background and main goals

During 2022, we have conducted an internship in Baidu's NLP department, taking the multilingual pre-training model ERNIE-M as a starting point to learn a series of theories related to the study of multilingual understanding tasks, based on deep learning, and to explore its application scenarios. We take multilingual text classification as an example, investigate its use cases in three fields: pornographic and anti-political content recognition, news classification, and sentiment analysis, describe in detail the process of pre-training plus fine-tuning in conjunction with the use cases, and evaluate the model’s quality. Finally, the multilingual language understanding task summarizes the development paradigm of pre-training plus fine-tuning in the field of NLP deep learning and tries to validate it in the field of intelligent search, proving that the development paradigm of pre-training plus fine-tuning can effectively improve the models.

With the development of science and technology, ultra-fast information transfer and high level of resource sharing have greatly facilitated our work and life, at the same time, information of orders of magnitude is constantly being produced in milliseconds, and the information explosion and information overload have brought people new challenges in the Internet era, and the ability to efficiently process information has become the key to drive our society.

In recent years, due to its series of powerful feature extraction networks, deep learning has liberated NLP tasks from traditional manual feature engineering, allowing models to effectively capture the patterns and features embedded in the original input data and even generalize in several scenarios. This neural network-based deep learning approach has achieved great success in fields such as Computer Vision and Automatic Speech Recognition, and has also made a lot of progress in the field of Natural Language Processing represented by Machine Translation, amongst others.

However, with the increase of task complexity and the demand of constant enhancements, deep learning models show a trend of increasing complexity. The more labeled training sets are available, the better the correctness and robustness of the model will be. However, when the labeled data is small, it will lead to deep learning models that are easy to overfit and cannot generalize well, which is not conducive to the improvement of user models and the solution of practical problems.

Taking the text classification task, one of the key fundamental tasks of NLP, as an example, it is unrealistic to rely on manual classification of huge amounts of text information. In recent years, classification tasks have become mainstream with the help of machine learning techniques, where computers can acquire empirical skills through continuous learning of labeled data sets and give a correct classification label even for unknown problems.

Multilingual text classification is a further exploration based on the capability of text classification. Since texts in different languages differ greatly in terms of length, grammatical structure, cultural background, etc., to classify texts in different languages, multiple monolingual classifiers need to be trained, and then these monolingual text classifiers are used to classify the languages supported by the system. There are abundant English labeled datasets, and the use of distinct language datasets, such as Chinese and Portuguese, can still be helpful to train models, but the scarcity of minority language labeled datasets seriously affects the application of multilingual text classification.

To solve these problems, a new mainstream trend in NLP has emerged: using the idea of transfer learning, a training task is divided into two stages: pre-training plus fine-tuning, that is, first using a large amount of unsupervised corpus for the pre-trained model to get a generic language representation, and then fine-tuning for a small amount of labeled corpus for specific NLP tasks (text classification, text matching, sequence annotation, etc.).

Baidu released its multilingual pre-training model ERNIE-M in early 2021 (Ouyang et al., 2021), which enables a model to understand 96 languages simultaneously. This technology has set new world best results on 5 types of typical multilingual understanding tasks. ERNIE-M also topped the authoritative multilingual

understanding list XTREME, surpassing models proposed by Microsoft, Google, Facebook and others. This is a relevant breakthrough for multilingual text classification tasks. Based on this model, we only need to train the model with a small amount of text data with annotations in a single language and can use the model to classify any text in 96 languages in a given application scenario (Ouyang et al., 2021).

In this thesis, We explore the principles of the cross-language understanding pre-training model ERNIE-M and, based on it, the applicability of the model for multilingual text classification tasks after fine-tuning. We try to validate it experimentally in three real-world scenarios, namely, false comments, news headline classification and sentiment analysis in comments. It is finally shown that ERNIE-M has the ability to classify multilingual text.

1.2 Thesis Structure

The structure of this thesis is as follows:

Chapter 1: Introduction and Motivation This chapter describes the context of this study, the importance of multilingual text categorization research and the chapter organization.

Chapter 2: Overview of Baidu. Initially, the overall status of the Baidu firm is presented, focusing on the introduction of Baidu NLP department and Baidu translation, and enumerating in detail the diverse products and functions of Baidu translation.

Chapter 3: State of the Art. It consists mostly of NLP, machine translation, text classification, cross-language comprehension, and multilingual text classification, which are five aspects that are gradually expanded along the chapter.

Chapter 4: Experimental validation of the Baidu multilingual text categorization model. In the multilingual text classification scenario based on the cross-lingual understanding model, we first summarize and investigate the application cases of the cross-lingual understanding model. The use cases are condensed into three domains of expertise: detecting specific text, extracting text themes, and sentiment analysis. On

the basis of these three fields, three specific use case scenarios are chosen to collect relevant data: false comment identification, news title categorization, and sentiment analysis on comments. After data annotation, the Baidu pre-training model ERNIE-M is utilized for training fine-tuning, and the experimental outcomes are then assessed. It demonstrates the usefulness of Baidu's ERNIE-M model for cross-language comprehension in the multilingual text classification scenarios analyzed.

Chapter 5: Summarization on the entire thesis and look ahead to the next steps.

Chapter 2

2 Baidu´s Presentation

Baidu is a leading AI company with a strong Internet foundation. Baidu is one of the few companies in the world to provide full-stack AI technology, including AI chips, software architecture and applications, and has been named one of the top four AI companies in the world by international organizations. Baidu's mission is to "make the complicated world simpler through technology"(Company Overview) and to become the world's top high-tech company that understands users best and helps people improve.

Founded in Zhongguancun, Beijing on January 1, 2000, Baidu's founder, Robin Li, holds the patent for "hyperlink analysis" technology, making China one of only four countries in the world to possess the core search engine technology, in addition to the United States, Russia, and South Korea. Baidu responds to billions of search requests from more than 100 countries and regions every day, and is the most important portal for Internet users to access information and services in Chinese, serving 1 billion Internet users.

Based on the search engine, Baidu has evolved artificial intelligence technologies such as voice, image, knowledge graph, and natural language processing; in the last 10 years, Baidu has invested in deep learning, conversational AI operating systems, autonomous driving, AI chips and other cutting-edge fields, making Baidu a leading AI company with a strong Internet foundation.

2.1 Baidu overview

On January 1, 2000, Robin Li and his partner, Xu Yong, founded Baidu Inc (Nasdaq: BIDU) in Zhongguancun, Beijing, China. The name "Baidu" is derived from a line in the Chinese Southern Song Dynasty lyricist Xin Qiji's "The Green Jade Case - New Year's Eve": "When I look for him in the crowd, I suddenly look back, but the person is at the end of the lamp." It reflects Baidu's confidence in its own technology and its persistent pursuit of Chinese information retrieval technology. Baidu's corporate logo is a "bear paw", which has the imagery of "hunters using bear paws to

find traces" and is very similar to Baidu's "analytical search technology", thus constituting Baidu's search concept.

Figure 1: Baidu's Icon

As of January 2022, Baidu was ranked third in the world and first in China by Alexa and accounted for 0.58% of the world's search engine market share, ranking fourth in the world and first in China. (Company Overview | Baidu Inc, 2020)[1]

When Baidu was founded, its main product resembled a set of 'behind-the-scenes tools' to provide search technology services to major portals such as *Silicon Valley Power, Sina, Sohu,* etc. By providing search technology services, Baidu quickly captured 80% of China's search engine market and became the most important search technology provider. By the end of 2001, Baidu launched an independent search engine, directly serving the C-user, and business focus began to shift from "business-oriented" to "user-oriented" (Johnson, 2021).[2] At the same time, Baidu followed the example of Overture in the United States and created the first "bidding ranking" in China, which determines the ranking of advertisers in the website based on the amount of money paid. This has enabled Baidu to generate huge amounts of traffic and search advertising revenue through its search engine products.

At the end of 2003, Baidu created the largest Bullletin Board System (BBS) in the history of the Chinese Internet, which is a combination of a search engine to build a communication platform, any user can create a posting on any keyword, post relevant information, others can find them through the search engine, which allows people, who are interested in the same topic, to come together to communicate and help each other. It is a low barrier (no need to sign in to become a bar user), easy to

[1] http://home.baidu.com/about/about.html
[2] https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

operate, and has a large number of participants (everyone can use keywords to create a bar), making it extremely influential in China. Baidu's posting bar has gradually become the world's largest Chinese community in terms of traffic and has also opened up the web 2.0 era of user-contributed content for Baidu. If Baidu's search engine was doing information distribution before, the emergence of the posting bar marked the beginning of Baidu's efforts in information aggregation. Based on this, Baidu followed up with Baidu Know and Baidu Encyclopedia, the former being an interactive knowledge Q&A sharing platform and the latter being a Chinese information collection platform covering various fields of knowledge, which together attracted a lot of traffic for Baidu and accelerated the development of Baidu's community very quickly.

On August 5, 2005, Baidu was listed on NASDAQ in the U.S., and on its first day of trading, it rose as much as 354%, setting a record for the first day of overseas stock gains on the NASDAQ (Dfdaily, 2011). In the midst of the booming Internet industry, Baidu began to seek diversified layout development. From 2005 to 2009, Baidu had 21 product lines, including not only star products such as Baidu Encyclopedia, Baidu Know, Baidu Post, but also Baidu Space, Baidu Video, Baidu Map, Baidu IM software "Baidu HI" and a series of new products born under the background of "diversification" strategy. "These products cover all aspects of people's lives. Among these 21 product lines, there are 7 or 8 products with more than 100 million users.

However, the development of Baidu's diversification has not been smooth, and Baidu has met its Waterloo in several important areas, such as e-commerce, social networking and games. Take e-commerce as an example, in 2007, Baidu set up the e-commerce division, but an important part of the e-commerce product is to maintain the service from supply chain to merchants to users, which requires a lot of refined operational investment. Baidu's consistent style of "raising products with traffic" made the development of Baidu in the field of e-commerce difficult, coupled with the lack of differentiation with Alibaba, which already occupies a large market share in the C2C field, eventually announced in 2011 to close the e-commerce function, to local life platform transformation.

In 2008, CCTV publicly criticized Baidu. In 2008, CCTV publicly criticized Baidu, saying that Baidu's ranking was a major drawback and had reduced the search experience of Internet users. Baidu's stock price plunged in response to CCTV's voice, and major media outlets reported on the many problems with Baidu's bidding system. Baidu has since begun a profound internal reflection, and six months later, the "Phoenix Nest" system was officially launched. On the basis of the original bidding ranking rules and the introduction of "quality" (including click-through rate, creative quality, account performance, etc.), an entry placed information, its placement price and "quality" will jointly determine its display ranking, this rule replaced the original This rule replaces the original bidding ranking, and fully manages all page positions of search promotion. This has improved the search experience of users to a certain extent, and also contributed to the growth of Baidu's advertising revenue.

Looking back at Baidu's development before 2009, Baidu built a strong content ecosystem using search and community through its core products such as Baidu Search and Baidu Post, plus the lucrative profits brought to Baidu by the honeycomb system and bidding rankings, which made Baidu expand from a single search engine company to one of China's Internet giants in just a few years.

In August 2009, Baidu proposed the concept of box computing, which allows users to complete a series of information search, consumption, and service activities through the search box on Baidu pages. For example, users can enter the weather in the search box, and Baidu will automatically identify and match the best content provider or service provider to respond and process. With this concept, Baidu has realized the change from distributing information traffic to distributing application traffic.

In January 2010, Google announced its withdrawal from China, and Baidu lost its biggest competitor in China. In 2011, Baidu launched a new home page and a mobile terminal platform, turning the original standalone search box into a full application model. After 2012, with the increasing market share of smartphones, the Internet era has gradually changed from the traditional PC to mobile. Search engines are no longer the main entrance in the mobile Internet era, replaced by the consumption of various applications and distribution platforms, fragmented information and offline services, etc. The strategic advantage of frame computing is

doubly hindered in the mobile Internet era. Therefore, the layout of Baidu began to shift to the mobile side, building mobile Baidu, maps, mobile distribution-based three major traffic entrances. And one by one, it acquired Nuomi.com and launched O2O products such as takeaway, wallet, travel, medical and real estate, trying to transform into a full media platform with search engine as the core.

By the second half of 2015, China's O2O platform ushered in the winter, local life services software Meituan merged with VW Dianping, Ali invested in hungry, Baidu takeaway eventually packaged Baidu traffic entrance, sold to hungry at a price of $800 million. Along with the gradual marginalization of Baidu Nuomi business, Baidu O2O to seize the mobile portal, the transformation of the full media platform answer sheet is not too ideal.

In June 2016, Baidu in the search box home page online "Feed stream", that is, continuously updated and presented to the user content information flow. On the basis of the original Baidu app, information flow combined with it, which greatly improved user stickiness and daily active user volume. It can be said that the launch of information flow was an extension of Baidu's original core search business, which allowed Baidu to gain a firm foothold in the mobile Internet era and laid the foundation for Baidu's further growth later. 2018 Baidu financial report showed that the overall information flow user hours of Baidu increased by 112% year-on-year (Jiedian, 2019).

Under the huge wave of the entire mobile Internet era, search is no longer the only traffic entrance in the original PC era, and more and more vertical mobile apps are gradually breaking this tradition. For example, people can use Meituan, hungry to buy takeout; use Ctrip, Ma Hive to travel-related advice consumption; when people want to check the weather can download a series of mobile apps to meet the search needs, such as Caiyun weather. More and more daily life content has greatly enriched the search result resource base of today's headlines app, and more people started to search for daily information and other content on today's headlines. The mobile Internet environment, the Internet world became gradually closed, and even they chose to close the information and refuse Baidu's crawl, Baidu's traffic began to gradually reduce, the core advertising business received threats, this year's Internet giant urgently needed to consider a new direction of development.

In 2017, Baidu invited Lu Qi to serve as Baidu's revenue operations officer, reporting directly to Robin Li and responsible for the company's operations, sales, technology, and products. Lu Qi directly cut off the then non-core takeaway and medical advertising businesses, focusing on Baidu's core technologies, trying to take off the search engine label, putting forward the 'ALL IN AI' strategy for the first time, and starting to transform into an AI ecological enterprise.

About the development of Baidu AI, in fact, since January 2010, Baidu began to lay out artificial intelligence technology, led by Wang Haifeng founded the "natural language processing department", trying to empower search through artificial intelligence technology. Immediately afterwards, Baidu continued to comprehensively lay out AI technology, and successively carried out research and development of artificial intelligence technologies such as voice, image, machine learning, data mining, knowledge mapping, recommendation personalization and so on. After about two years, it has basically completed a more complete layout.

From 2012, Baidu began the research and development and application of deep learning technology, and in 2012 that year in Baidu's voice, image, and other systems online. At this time, Baidu initially formed its AI strategy and set up the Deep Learning Institute and AI Lab. It was because it saw the amazing performance of deep learning technology in practical applications that Baidu established the world's first deep learning research institute in January 2013, and also in this year, Baidu was the first in the world to apply deep learning to a large-scale online search engine. 2014, Baidu began to step into the field of intelligent driving, and has launched a deep voice system, Baidu Brain, Do Secret, Apollo autonomous driving platform and other technologies. In 2014, Baidu started to enter the field of intelligent driving, and launched its deep voice system, Baidu Brain, Dou Secret, Apollo autonomous driving platform and other technology systems. In 2015, Baidu was the first in the world to launch a neural network-based machine translation system. This is also the key area to be studied in this paper. In 2016, Baidu officially opened Baidu Brain (Baidu Brain is an AI open platform integrating natural language processing, deep learning, knowledge mapping, speech, vision, and other technologies) to the public. Baidu also launched PaddlePaddle, an open-source deep learning platform that integrates the core framework, tools, and service platform of deep learning; and for the first time, Baidu

released a trinity cloud computing strategy of "cloud computing + big data + artificial intelligence", which is committed to providing a good ecological service environment for Baidu cloud users.

Before 2016, Baidu's business was mainly focused on the mobile Internet sector, Baidu has since formed an ecological layout with AI as the core. In 2019, cloud computing was defined as the base of intelligent infrastructure, while AI and underlying technical capabilities were instilled into the base, thus becoming a "power plant" to empower various industries. At the beginning of 2021, Baidu once again solidified its AI strategy, transitioning from the integration of "AI+Cloud" to a new phase of cloud intelligence.

As an important bearer and exporter of Baidu's AI implementation, Baidu Intelligent Cloud has established a new strategy of "taking cloud computing as the foundation, AI as the grip, and focusing on important tracks", choosing smart city, smart finance, smart medical, smart manufacturing, smart energy, and other important tracks, and repeatedly signing orders of hundreds of millions of dollars.

Baidu intelligent cloud scale landing undoubtedly corroborates the view that when artificial intelligence from the seed stage to flowering, Baidu is in a new round of spiral upward process. According to the "2020 AI China Patent Technology Analysis Report" released by China National Industrial Information Security Development Research Center and the Electronic Intellectual Property Center of the Ministry of Industry and Information Technology, Baidu is in the first place with 9,364 patent applications and 2,682 patent licenses respectively in terms of AI patent applications and licenses, which is also the third consecutive year that Baidu has been in the first place in the total number of applications in the AI patent analysis report. This is also the third consecutive year that Baidu has ranked first in the total number of applications in the AI field. The report also pointed out that among the 10 fields of AI patent applications, Baidu was the first in the seven fields of natural language processing, intelligent recommendation, deep learning, traffic data, intelligent speech, knowledge mapping and autonomous driving.

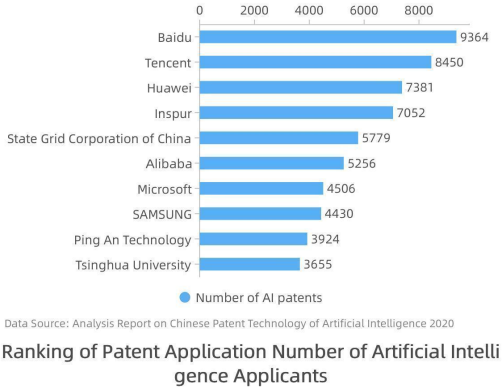

Figure 2: Ranking of Patent Application Number of Artificial Intelligence Applicant

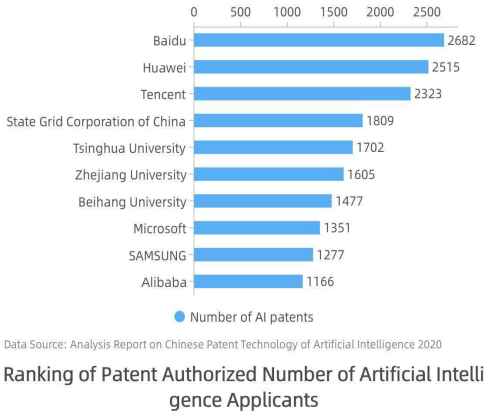

Figure 3: Ranking of Patent Authorized Number of Artificial Intelligence Applicants[3]

[3]https://www.ncsti.gov.cn/kjdt/ztbd/xydrgzn/lbt_848/202011/t20201119_24763.html

Not only does the number of patents lead the way, but many of Baidu's AI technologies have also been implemented in the daily life of millions. Baidu CTO Wang Haifeng has said that in recent years, with the development of technology, AI has entered the industrial mass production stage as large-scale industrial applications are landed. Based on artificial intelligence technology, Baidu has created various platforms, industry applications and industry solutions, which have now been implemented in urban, agricultural, medical, service, manufacturing, and other fields.

Currently, major investment banks and industry insiders generally believe that Baidu's business will have greater growth potential in the future and is a veritable AI ecological company. 2021 March, Baidu's second IPO in Hong Kong, Robin Li said in his IPO speech that returning to Hong Kong for a second IPO is a second departure for Baidu and a second venture for Baidu. In the future, Baidu will have a broader space for development.

2.2 Baidu’s NLP team

Natural language processing is a cross-discipline of computer science, artificial intelligence, and linguistics that aims to solve the problem of interaction between computers and human language. This includes the analysis, understanding, generation, retrieval, transformation, and translation of natural language.

Since the birth of Baidu, NLP technology has played a crucial role in this, and basic NLP technology such as Chinese word separation has been an essential part of the entire search engine since Baidu searched the first result for the first user. Along with the rapid development of Baidu, Baidu NLP is also developing in parallel, or even faster. The acceleration of this development started in the spring of 2010, when Dr. Haifeng Wang, a top international natural language processing expert, joined Baidu, the work of Baidu NLP was reorganized and re-planned, in addition to the traditional directions of word separation, proper name recognition, query demand analysis, query rewriting, etc. were strengthened, and new technologies such as machine translation, machine learning, semantic search, semantic understanding, intelligent interaction, deep Q&A, chapter understanding, etc. were also expanded. In

addition to the traditional directions, such as machine translation, machine learning, semantic search, semantic understanding, intelligent interaction, deep question and answer, chapter understanding, etc., new technical directions have been expanded. Under his leadership, the NLP team has grown from just 10 people to a team of more than 200 people today, with diverse talents in algorithm development, system implementation, academic research, linguistics, product design and architecture, front-end, client-side engineering development, etc., fully supporting Baidu's various product applications.

Baidu NLP is characterized by "deep, forward-looking, application, and innovation"(Baidu NLP, 2010), maintaining deep and forward-looking academic research, but also focusing on the transformation of product applications in industry, especially good at making innovations in the combination of technology and products.

In recent years, the NLP department has led the industry's first deep learning technology in search and machine translation; machine translation has become the first national science and technology progress award for Internet enterprises; incubated the artificial intelligence products DouSecret and Xiaodu Robot, which are the first AI technology to start, accumulate deep knowledge and stand at the tip of the wave; for the first time in the industry, online learning technology has been applied to industrial-grade products, serving millions of users; personalized user modeling technology has blossomed in many Baidu product lines; traditional NLP technology directions such as deep Q&A, syntax/lexology, and digest generation have been fully covered and launched in search and other products to improve the technical barriers and user experience of products.

Wang Haifeng, the only Chinese chairman in the history of Association for Computational Linguistics (ACL), often uses the phrase "look up to the stars and keep your feet on the ground" to encourage colleagues in the NLP department, which makes every step of the NLP team approaching the stars a solid and stable one. Let computers understand human language, create products with real intelligence, change the lives of hundreds of millions of users, and ultimately make people's lives better and the world better. This is the common ideal of Baidu NLP people, and this ideal is becoming reality step by step through their persistent efforts.

2.3 Baidu Translation

Baidu Translate is an online translation service released by Baidu, relying on massive Internet data resources and the advantages of leading natural language processing technology, dedicated to helping users across languages to access information and services more conveniently and quickly.

It currently supports 200 languages worldwide, including Chinese (simplified), English, Portuguese, etc., covering about 40,000 translation directions, and supports over 400,000 enterprises and individual developers through its open platform, responding to hundreds of millions of translation requests on average daily, and is the translation product with the largest market share in China. It has made significant breakthroughs in several translation technologies, released the world's first Neural Network Translation (NMT) online system, and won China's National Science and Technology Progress Award in 2015.

2.3.1 Baidu translation product forms and functions

Baidu Translation has various product forms such as webpage, app, Baidu applet, etc. In addition, it also provides open cloud interface service for developers and responds to 100-billion-character translation requests daily. In addition to text and webpage translation, Baidu has launched multimodal translation functions such as document translation, picture translation, photo translation and voice translation, as well as rich foreign language resources, such as massive example sentences and dictionaries, and foreign language learning functions, such as practical spoken language, English follow-up reading, English short videos and AI word memorization to meet users' diverse translation needs and learning needs.

2.3.2 Baidu Translation Open Platform

Baidu Translation Open Platform can provide developers with online general text translation services in 200+ languages; vertical translation services with accurate terminology translation results and sentence styles that meet industry characteristics; in addition, it can also build customized translation systems based on the bilingual pair corpus provided by customers to accurately meet the translation needs of pendant

scenarios; and help customers deploy local models and model independent training In addition, it can also help customers deploy local models and autonomous model training capabilities to meet the requirements of translation accuracy, security and reliability. So far, it has served more than 400,000 developers, covering dozens of industries such as e-commerce, education, smart hardware and social networking, and responded to an average of 100-billion-character translation requests per day.

2.3.3 Baidu AI Simultaneous Translation

In terms of capability, Baidu AI simultaneous interpretation is a full-scene, multimodal, cross-platform intelligent machine simultaneous interpretation solution provided by Baidu. It uses a neural network translation model that can fuse contexts and can model predictions based on semantic units, which enables Baidu AI simultaneous interpretation to provide smooth and accurate simultaneous interpretation with low latency. It has successfully provided services for large-scale conferences such as the Baidu AI Developer Conference.

In terms of product form, for users with conference simultaneous interpretation needs, Baidu AI Simultaneous Interpretation Conference Edition is launched. It supports offline and online meeting forms, and provides functions such as bilingual subtitle casting, cell phone listening and watching, industry terminology customization, and viewing meeting records; for scenarios such as watching video, listening to audio, and online meeting communication, Baidu AI Simultaneous Assistant is launched, which can pick up computer voices and generate bilingual subtitles for simultaneous interpretation in real time to assist in cross-language information acquisition and communication.

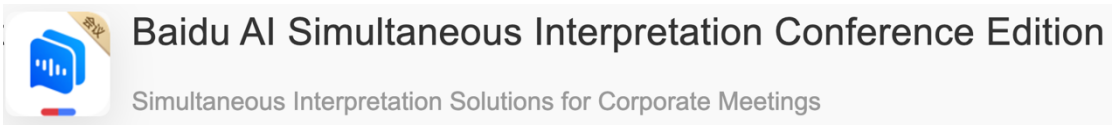

Figure 4: Icon of Baidu AI Simultaneous Interpretation Conference Edition

For users with audio transcription and subtitle translation needs, AI video translation is introduced to generate bilingual subtitles with one click; for personal learning and internal reference needs of enterprises, combining the advantages of

machine speed and human accuracy, AI translation + human proofreading service is provided; for commercial release and broadcast-grade film and TV translation production needs, one-to-one customized solutions are provided.

2.3.4 Baidu Web Translation

The web version of Baidu translation includes PC web translation and wise translation.

PC Web translation is the translation provided by accessing fanyi.baidu.com web page through computer, which supports translation between multiple languages. The input box supports the function of automatic checking of the original language type, and can translate documents, web pages and pictures, and supports translation of online literature, biomedicine, water conservancy machinery, electronic science, and technology vertical fields. The results of translation of text include corresponding target voice pronunciation, bilingual example sentences, Oxford and Collins dictionaries, English-English interpretation, featured video explanation, word root suffix, synonym analysis, Baidu encyclopedia, Baidu know (a question and answer webpage) and many other resources. The translation of web pages returns the entire translated page.

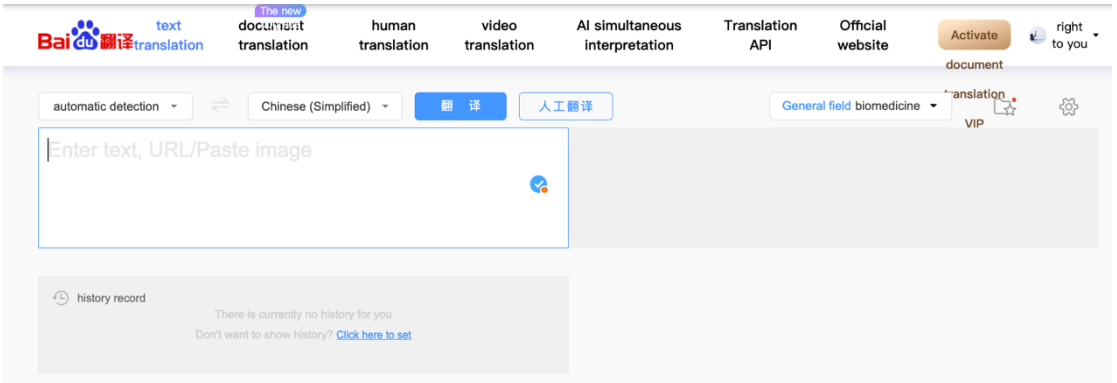

Figure 5: Interface of Baidu Web Translation

PC web translation is a product with rich functions. In terms of document translation, it supports word, pdf, ppt, excel format documents, supports mutual translation between Chinese and English, etc., retains the style and layout of the document to the maximum extent, uploads the whole translation with one click, the

original text and translation are viewed against each other, currently supports unlimited word translation, and can export the whole text for free. In terms of web page translation, Baidu Translate PC web page supports web page translation function, users only need to enter the URL in the input box, select the translation language, and they can translate the web page content, convenient to browse foreign language websites. At the same time, it supports Chrome, Firefox and other 8 browser web page translation plug-ins, which can identify the language of the page after installation, one-click web page translation, and support crossword translation. Baidu translation PC web page also supports image translation function, users can paste the image into the input box, Baidu translation can automatically extract the text from the image for translation. Baidu translation supports Chrome screenshot translation plug-in, users in any interface of Chrome using the mouse screenshot, the user can translate the text on the picture, bilingual cross-reference.

Wise Translation is a translation provided on cell phones by visiting the fanyi.baidu.com web page, supporting translation between multiple languages. Its functions are relatively simple compared to the web version, including translation results, pronunciation, example sentences, encyclopedia, and Chinese-English interpretation.

2.3.5 Baidu Translate App

Baidu Translate App is an English learning software that combines translation, dictionary, bilingual texts, audio and video content, word memorization and English-speaking assessment.

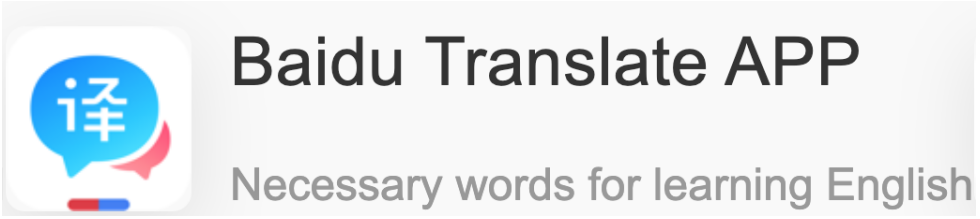

Figure 6: Icon of Baidu Translate App

Here we mainly analyze the translation function of the Baidu translation app, including ordinary translation, voice/conversation translation and OCR translation.

Ordinary translation is the basic translation capability of Baidu, which supports mutual translation of texts in more than 200 languages around the world and covers more than 40,000 translation directions. In terms of conversation translation, it supports CH-EN, CH-JA, CH-KOR, CH-TH, CH-DE, CH-RU, CH-FRA, CH-SPA, JA-EN translation, and users can translate into the target language by speaking directly to the phone. It also has real-time voice translation, which supports voice input of 21 popular languages, and even has English oral scoring and sound correction to help improve spoken language. In terms of OCR translation, OCR technology is applied to the translation field, which translates into the target language by recognizing images or languages. It supports three translation modes, namely photo translation, AR translation and real-time word pickup, and currently covers 17 popular languages.

2.3.6 Baidu Machine Translation

Machine translation is currently one of the most challenging research topics in artificial intelligence research. Making computers understand a language and convert it into multiple languages is not only the dream of centuries of technology, but also a free way of communication that the general public desires.

In June 2011, Baidu launched the machine translation system based on Internet big data, and in May 2015, Baidu released the world's first Internet neural network translation system, which integrates the methods of statistics and deep learning and imitates the process of understanding language and generating translations for automatic translation with the help of a large number of computer-simulated neurons, which makes the quality of machine translation greatly improved.

Baidu's neural network translation system predates Google by more than a year. The core of Baidu's deep learning-based translation system is a deep neural net composed of a large number of nodes (neurons). An utterance in a language is quantified to a certain extent and then transmitted through layers to become an expression that can be "understood" by a computer.

It is interesting to note that when a huge system is built, it is like a newborn child that does not know anything. The system gradually improves its translation ability

mainly by learning a lot of mutual translation between Chinese and English. Baidu uses a large network database to discover a large number of sentences from a large number of Chinese and English sentences and uses them as "learning materials".

Because the construction of deep neural networks is extremely complicated, the "learning" process is very long and consumes a lot of energy. If the number of "teaching materials" is calculated by the number of words in the Encyclopedia Britannica, then the amount of knowledge learned in a week is equal to one million copies of the Encyclopedia Britannica. Through technical research, Baidu's mechanical translation team has significantly reduced the learning cycle.

Relying on the advantages of massive Internet resources and natural language processing technology, Baidu has developed high-quality translation knowledge acquisition technology, breaking through the bottleneck of small scale and high cost of traditional methods in translation knowledge acquisition; multi-strategy translation model based on Internet big data, capable of responding to users' complex and diverse translation in multiple fields and genres, including new words on the Internet, scientific and technological literature, e-commerce, ancient languages, Cantonese, etc. in real time The pivot language based translation method makes it possible to translate small languages with limited resources.

As a comprehensive, multi-service Internet company, technology innovation seeks to be universally accessible to as many users as possible in the easiest way possible. Baidu has combined its leading image recognition, voice recognition and other technologies, and has now developed into a mature system that integrates many functions in multiple application scenarios.

At present, the main product forms of Baidu machine translation include text translation, vertical field translation, translation customization training, document translation, voice translation, image translation, English speaking assessment, AI simultaneous interpretation and translation privatization deployment.

Among them, text translation supports mutual translation in more than 200 languages and users only need to pass in the content to be translated and specify the source language (supporting automatic language detection) and target language to be

translated to obtain the corresponding translation results and can also intervene in the results. The dictionary version of Text Translator comes with millions of Chinese and English dictionary resources and speech synthesis resources in the translation results, which can efficiently help educational application developers. The text translation function can be applied in a variety of industries. In foreign language teaching and learning scenarios, it can help teachers and students communicate with each other, help foreign teachers conduct post-class reviews, and assist students in reading and writing through real-time sentence translation, word interpretation, and speech synthesis to improve learning efficiency and quality in all aspects. It can also be applied to cell phone systems to realize services such as cell phone system word pickup translation and dialogue text translation, providing convenient translation functions for cell phone application developers. Cross-border e-commerce and smart hardware industries are also two major application industries for text translation, which can translate basic website information such as product names and detail pages in cross-border commerce services to help enterprises develop international markets, and can also be applied to hardware systems such as translation machines, learning machines and smart watches to provide users with text translation, dictionaries and speech synthesis, realizing convenient and accurate multilingual mutual translation functions.

In terms of translation in vertical fields, Baidu Machine Translation can make targeted optimization based on specific application scenarios such as biomedicine, electronic technology, water conservancy machinery, finance and economics, and online literature to make terminology and sentence translation more authentic. For fields not covered at the moment, Baidu Machine Translation also supports data participation in model training by providing double statements of specialized terms, as well as providing terminology dictionaries for optimization intervention. It is widely used in education learning, online literature and biomedical fields. In education learning, it can improve translation quality by applying chapter translation technology and combining contextual information for problems such as translation of human names, industry terms, word suffixes and word deformation in foreign teaching dialogue scenarios. It can also apply adaptive technology and chapter technology to customize training for problems such as consistency of human names and place names, complex sentence processing and skill name translation in online literature

scenarios, and the translation effect can reach the human translation level of junior translators after evaluation. In the biomedical industry, Baidu Translation can build a medical translation engine for a large number of specialized terms in biomedical scenarios, combined with massive corpus data accumulation, to precisely meet the personalized translation needs of customers.

In terms of translation customization training, Baidu Machine Translation can quickly build a set of customized translation systems for specific fields based on the domain bilingual pair corpus data provided by users independently and realize pre-consultation custom training. On this platform, users can upload the corpus independently, without the need to have the foundation of algorithm and model, without the assistance of technical personnel, and complete the whole process of model training, debugging and deployment with zero threshold, and the platform supports private deployment.

In terms of translation privatization deployment, Baidu Machine Translation can rely on its profound technical reserve to provide customers with translation capability privatization services covering text translation, document translation, image translation, corpus, translation engine optimization, model independent training, and so on. It can be deployed to enterprise local servers or private cloud servers according to different user needs, fully meeting customers' requirements for translation accuracy, data security and service reliability.

The release of Baidu's online translation system based on deep learning shows us the real hope of solving the classic problem of artificial intelligence, which is machine translation. Perhaps in the future, people all over the world will be able to communicate freely and access information and services on a global scale. Even with different languages, people will be able to exchange ideas and concepts and pass on ideas and cultures.

2.4 Baidu Ernie-M

In April 2020, Carnegie Mellon University, Google, and DeepMind jointly presented The Cross-lingual TRansfer Evaluation of Multilingual Encoders

(XTREME)[4], an authoritative large-scale multilingual evaluation list covering forty languages, which quickly became the golden evaluation set for cross-lingual pre-trained models (Hu et al., 2020). XTREME consists of nine subtasks in four categories: text classification, sequence annotation, sentence recall, and question and answer, in which the models are tuned on English training data and then inferred on their respective test sets in 40 languages from 12 other language families. A higher score means that the model is better able to transfer the knowledge learned from the dominant language (English) to smaller language applications.

As major Internet companies compete for business abroad, more and more mature businesses require rapid deployment of small language versions. The hot application demand has led to the research of cross-language pre-training models, and since the release of XREME, top international academic institutions and technology companies such as New York University, Google and Microsoft have been competing fiercely. mBERT is only a BERT that replaces the training corpus, and is not really designed for cross-language migration, meaning, it is not really a pre-trained language model designed for cross-language transfer. Standing on its shoulders, successors have taken the XTREME list higher and higher. In 2020, Microsoft's T-ULRv2 model won the championship (Hagen, 2020). But the record was held for less than 2 months before Baidu released its own model, ERNIE-M, to break it.

On January 1, 2021, Baidu Research Institute introduced a new training method, ERNIE-M (Ouyang et al., 2021), which is a multilingual model that can understand 96 languages and improve the cross-linguistic transferability of the model on data-sparse languages as well. Experimental results show that ERNIE-M obtained state of the art results on all five cross-language downstream tasks, topping the list with a score of 80.9 and setting a new record.

[4] https://github.com/google-research/xtreme

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
2	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
3	Anonymous3	Anonymous3	Anonymous3	Jan 3, 2021	79.9	88.2	74.6	71.7	89.0
4	Polyglot	MLNLC	ByteDance	Nov 13, 2020	77.8	87.8	72.9	67.4	88.3
5	VECO	DAMO NLP Team	Alibaba	Sep 29, 2020	77.2	87.0	70.4	68.0	88.1
6	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
7	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
8	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
9	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7

Figure 11: Rank of ERNIE-M on XTREME, Jan 2021[5]

2.4.1 ERNIE-M Background

There are currently a number of technologies that try to learn the semantics of different languages using monolingual corpora, and then use bilingual corpora to align the semantics of different languages. However, the bilingual corpus of most languages is very sparse, which means that the advantages of the multilingual model are limited due to the lack of bilingual corpora. Using back translation as a mechanism, Baidu researchers propose a mechanism based on the idea that in order to overcome the limitation of bilingual corpus sizes on the learning effects of multilingual models, and to improve the effectiveness of cross-language understanding, there has to be a mechanism that breaks through the limitations. ERNIE-M is a pre-trained model that learns semantic alignment relationships between languages based on monolingual corpora and significantly improves five typical cross-language models. These include cross-language natural language inference, semantic retrieval, semantic similarity, named entity recognition, and reading comprehension.

[5] https://sites.research.google/xtreme

2.4.2 ERNIE-M Principle

It is designed with the objective of enhancing multilingual semantic representations through the implementation of two phases of individualized pre-training tasks as part of the ERNIE-M training process. In the first stage of the model, cross-linguistic comprehension is acquired by learning from a limited amount of bilingual corpora in order to establish initial linguistic alignment relationships, whereas in the second stage, the model is enhanced with cross-linguistic comprehension by learning from a large amount of monolingual corpora using the concept of back translation.

As part of the first phase of learning, ERNIE-M proposed the Cross-attention Masked Language Modeling (CAMLM) pre-training algorithm, which captured information about inter-language alignment using a small bilingual corpus. As part of the CAMLM project, researchers represent parallel sentence pairs in the form of <source sentence, target sentence>. In order to learn multilingual semantic representations, CAMLM is required to restore the masked words from the target sentences without exploiting the context of the source sentence.

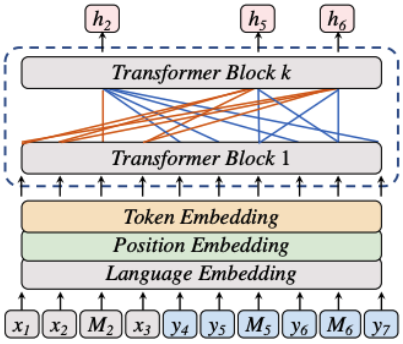

Figure 12: Overview of CAMLM training[6]

As shown in Figure, x refers to one language, y refers to another language, and M refers to the token to be predicted, such that the sentence pairs constitute a parallel corpus. For example, given a parallel PT-EN sentence pair with input <Se amanhã [MASK][MASK], Will it be rainy tomorrow>, the model must recover the MASK token as <vai chover> in the source sentence only by the meaning of the target

[6] https://arxiv.org/abs/2012.15674

sentence, thus learning the semantic representation between the two languages, so that the model initially align relationships between the languages.

The second stage is Back-translation Masked Language Modeling (BTMLM) to align cross-linguistic semantics with the monolingual corpus. Specifically, the researchers trained the model using BTMLM, which is built based on transferability learned through CAMLM, to generate pseudo-parallel sentences from monolingual utterances. The generated sentence pairs are then used as input to the model, which can rely not only on the original input sentences but also on the generated pseudo-parallel sentences when restoring masked words to further align cross-linguistic semantics and thus enhance multilingual representation.

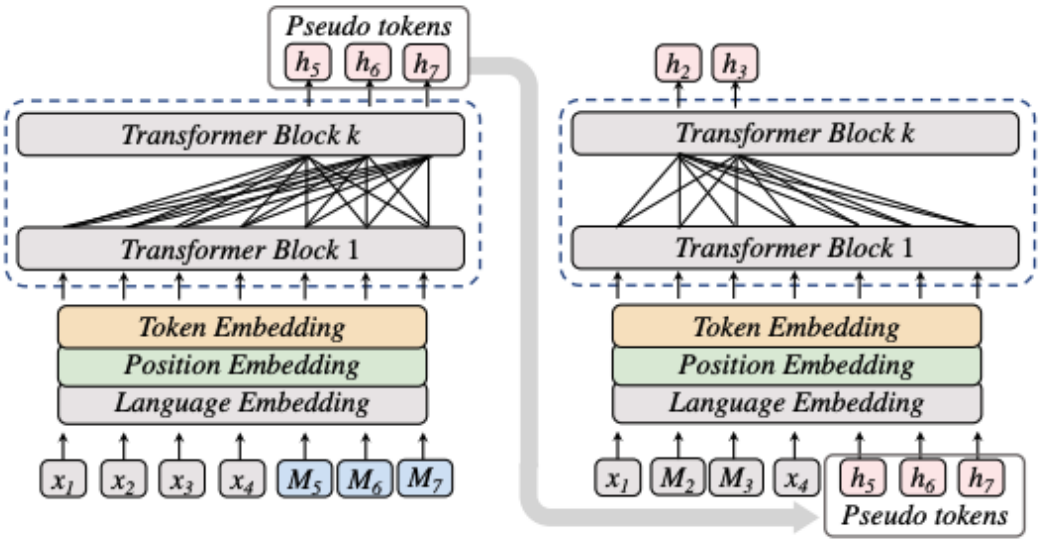

Figure 13: Overview of BTMLM training[7]

As shown in figure 13, add several Mask tokens to the back of sentences in one language (such as: x1, x2, x3, and so on), and then use the model to convert the M token into the corresponding tokens in another language, thus obtaining a pseudo-translated character (pseudo token). In other words, M expects to get the translation of the previous x1 x2 x3 in another language after running the model. After that, the obtained pseudo-characters are spliced to the back of the original sentence, and at the same time, some of the preceding tokens are masked out, and then

[7] https://arxiv.org/abs/2012.15674

input into the model for the supervised MLM task. For example, the input single sentence is <Eu gosto muito de comer maçã>, the model will first generate a pseudo-bilingual parallel sentence according to the input sentence <Eu gosto muito de comer maçã> <Eu gosto muito de comer maçã, eat apple >. Then learn the generated pseudo-parallel sentences <Eu gosto muito de comer [MASK][MASK], eat apples>. In this way, ERNIE-M leverages monolingual corpora to better model semantic alignment relationships.

2.4.3 Experimental results

There are two types of corpora used to train the ERNIE-M model: monolingual and parallel corpora, which is based on the PaddlePaddle[8] framework. This parallel corpus consists of a total of approximately 1.5 trillion characters of common words in 96 languages, including Chinese, English, French, Afrikaans, Albanian, Amharic, Sanskrit, Arabic, Armenian, Assamese, Azerbaijani, etc. A wide range of publicly available datasets were used to test the efficacy of ERNIE-M on five distinct tasks, cross-language natural language inference, reading comprehension, named entity recognition, semantic similarity, and cross-language retrieval, and all obtained optimal results.

ERNIE-M was evaluated by the Baidu researchers in two ways in order to determine its effectiveness.

1. Cross-lingual Transfer: This is an approach in which a model trained in English is directly tested on other languages to test whether the model is able to understand the other language. For example, if the model is asked to understand that "This restaurant has a comfortable environment" is a positive sentiment, the model needs to determine that "I am very happy." is also a positive sentiment. In practical applications, if there is a lack of labeled data in a specific language, this technique can help solve the problem by training multilingual models with labeled data in other languages, reducing the difficulty of building small language systems, even if there is no labeled data available in the language of interest.

[8] https://www.paddlepaddle.org.cn/en

2. Multi-language Fine-tuning: This approach uses the labeled data of all languages to train the model in multiple tasks, and then verifies whether the model can utilize the labeled data for other languages as well to further enhance the understanding of the language when the labeled data of the language is available.

The experiments of ERNIE-M also verified the effect of the model in other application fields, including cross-language retrieval, natural language inference, reading comprehension, and named entity recognition, which are summarized as following.

In the Cross-Lingual Information Retrieval task, semantically identical sentences are retrieved from bilingual corpora in order to find out what their content is. As can be seen in figure 14. With ERNIE-M, users will be able to retrieve results in other languages, such as Portuguese, French, German, etc., by searching for them in one language, such as English. The cross-language retrieval task of ERNIE-M achieves an accuracy rate of 87.9%, according to its performance on Tatoeba[9].

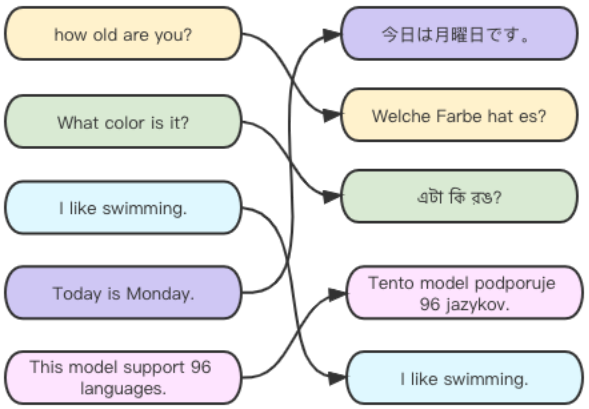

Figure 14: ERNIE-M on Cross-Lingual Information Retrieval task

In natural language understanding, natural language inference serves as a benchmark task. This is seen as one of the most challenging tasks since it aims to determine what logical relationship may exist between two sentences. Table 2 shows two examples of it. Multilingual dataset Cross-lingual Natural Language Inference

[9] https://paperswithcode.com/dataset/tatoeba

(XNLI)[10] contains 15 languages, which includes major languages like English and French, as well as minor languages like Swahili, which are part of the XNLI dataset.

Language	Sentence 1	Sentence 2	Label
English	You don't have to stay there.	You can leave.	Related
Portuguese	Maria tem medo de água.	Maria gosta muito de nadar.	Contradictory

Table 2: Example of Natural Language Inference

ERNIE-M verified its effectiveness in both Cross-lingual Transfer and Multi-language Fine-tuning. The researchers fine-tuned ERNIE-M training in English and tested it on Chinese, German, and Urdu, and were able to achieve an average accuracy of 82.0%. The accuracy can be further improved to 84.2% if the training corpus of all languages is used (Ouyang et al., 2021).

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune cross-lingual model on English training set (Cross-lingual Transfer)																
XLM (Lample and Conneau, 2019)	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder (Huang et al., 2019)	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
XLM-R (Conneau et al., 2019)	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
INFOXLM (Chi et al., 2020b)	**86.4**	**80.6**	80.8	78.9	77.8	78.9	77.6	75.6	74.0	77.0	73.7	76.7	72.0	66.4	67.1	76.2
ERNIE-M	85.5	80.1	**81.2**	**79.2**	**79.1**	**80.4**	**78.1**	**76.8**	**76.3**	**78.3**	**75.8**	**77.4**	**72.9**	**69.5**	**68.8**	**77.3**
XLM-R_{LARGE} (Conneau et al., 2019)	89.1	84.1	85.1	83.9	82.9	84.0	81.2	79.6	79.8	80.8	78.1	80.2	76.9	73.9	73.8	80.9
$INFOXLM_{LARGE}$ (Chi et al., 2020b)	**89.7**	84.5	85.5	84.1	83.4	84.2	81.3	80.9	80.4	80.8	78.9	80.9	77.9	74.8	73.7	81.4
$VECO_{LARGE}$ (Luo et al., 2020)	88.2	79.2	83.1	82.9	81.2	84.2	**82.8**	76.2	80.3	74.3	77.0	78.4	71.3	**80.4**	**79.1**	79.9
ERNIE-M_{LARGE}	89.3	**85.1**	**85.7**	**84.4**	**83.7**	**84.5**	82.0	**81.2**	**81.2**	**81.9**	**79.2**	**81.0**	**78.6**	76.2	75.4	**82.0**
Fine-tune cross-lingual model on all training sets (Translate-Train-All)																
XLM (Lample and Conneau, 2019)	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Unicoder (Huang et al., 2019)	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
XLM-R (Conneau et al., 2019)	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
INFOXLM (Chi et al., 2020b)	86.1	82.0	82.8	81.8	80.9	82.0	80.2	79.0	78.8	80.5	78.3	80.5	77.4	73.0	71.6	79.7
ERNIE-M	**86.2**	**82.5**	**83.8**	**82.6**	**82.4**	**83.4**	**80.2**	**80.6**	**80.5**	**81.1**	**79.2**	**80.5**	**77.7**	**75.0**	**73.3**	**80.6**
XLM-R_{LARGE} (Conneau et al., 2019)	89.1	85.1	86.6	85.7	85.3	85.9	83.5	83.2	83.1	83.7	81.5	**83.7**	**81.6**	78.0	78.1	83.6
$VECO_{LARGE}$ (Luo et al., 2020)	88.9	82.4	86.0	84.7	85.3	86.2	**85.8**	80.1	83.0	77.2	80.9	82.8	75.3	**83.1**	**83.0**	83.0
ERNIE-M_{LARGE}	**89.5**	**86.5**	**86.9**	**86.1**	**86.0**	**86.8**	84.1	**83.8**	**84.1**	**84.5**	**82.1**	83.5	81.1	79.4	77.9	**84.2**

Table 3: Evaluation results on XNLI cross-lingual natural language inference[11]

The goal of the Cross-lingual Question Answering task is to answer specific questions based on the text. To evaluate the effectiveness of ERNIE-M on the reading comprehension task, ERNIE-M was evaluated on the MultiLingual Question Answering (MLQA) dataset (Lewis et al., 2020) proposed by Facebook. In this task, the model needs to be trained on English first and then tested on datasets in other

[10] https://cims.nyu.edu/~sbowman/xnli/
[11] https://arxiv.org/abs/2012.15674

languages. This task can evaluate the effectiveness of the model on cross-language quizzing tasks and help in the construction of cross-language quizzing systems. The effect of this task is shown in table 5. When ERNIE-M is trained in only one language, 50.2% of questions in different languages can be completely answered correctly (Ouyang, 2021).

Model	en	es	de	ar	hi	vi	zh	Avg
mBERT	77.7/65.2	64.3/46.6	57.9/44.3	45.7/29.8	43.8/29.7	57.1/38.6	57.5/37.3	57.7/41.6
XLM	74.9/62.4	68.0/49.8	62.2/47.6	54.8/36.3	48.8/27.3	61.4/41.8	61.1/39.6	61.6/43.5
XLM-R	77.1/64.6	67.4/49.6	60.9/46.7	54.9/36.6	59.4/42.9	64.5/44.7	61.8/39.3	63.7/46.3
INFOXLM	81.3/68.2	69.9/51.9	64.2/49.6	60.1/40.9	65.0/47.5	70.0/48.6	64.7/**41.2**	67.9/49.7
ERNIE-M	**81.6/68.5**	**70.9/52.6**	**65.8/50.7**	**61.8/41.9**	**65.4/47.5**	**70.0/49.2**	**65.6**/41.0	**68.7/50.2**
XLM-R Large	80.6/67.8	74.1/56.0	68.5/53.6	63.1/43.5	62.9/51.6	71.3/50.9	68.0/45.4	70.7/52.7
INFOXLM Large	**84.5/71.6**	**75.1/57.3**	**71.2/56.2**	**67.6/47.6**	72.5/54.2	**75.2/54.1**	69.2/45.4	73.6/55.2
ERNIE-M Large	84.4/71.5	74.8/56.6	70.8/55.9	67.4/47.2	**72.6/54.7**	75.0/53.7	**71.1/47.5**	**73.7/55.3**

Table 5: Accuracy of MLQA data under each model[12]

The goal of the named entity recognition task is to identify information such as names of people, places, time and institutions in texts. It can help people to extract valuable information from a large number of texts quickly.

Model	en	nl	es	de	Avg
Fine-tune on English dataset					
mBERT	91.97	77.57	74.96	69.56	78.52
XLM-R	92.25	**78.08**	76.53	**69.60**	79.11
ERNIE-M	**92.78**	78.01	**79.37**	68.08	**79.56**
XLM-R Large	92.92	80.80	78.64	71.40	80.94
ERNIE-M Large	**93.28**	**81.45**	**78.83**	**72.99**	**81.64**
Fine-tune on all dataset					
XLM-R	91.08	89.09	87.28	83.17	87.66
ERNIE-M	**93.04**	**91.73**	**88.33**	**84.20**	**89.32**
XLM-R Large	92.00	91.60	**89.52**	84.60	89.43
ERNIE-M Large	**94.01**	**93.81**	89.23	**86.20**	**90.81**

[12] https://developer.baidu.com/article/detail.html?id=292593

Table 6*: F1-Score of CoNLL data under each model*[13]

As shown in Table 6, using a multilingual model can help with the task of information extraction on less resourced languages. ERNIE-M was evaluated on the CoNLL (Sang & de Meulder, 2003) dataset and the effect was verified in both Cross-lingual Transfer and Multi-language Fine-tuning modes. The researchers fine-tuned ERNIE-M in English and tested it on Dutch, Spanish and German, achieving an average F1 of 81.6%, which can be further increased to 90.8% when using the training corpus of all languages.

2.4.4 Concluding remarks

The achievement of cross-language machine understanding is one of the major challenges in artificial intelligence today. Eliminating the gap between different languages is a significant challenge, but it is becoming increasingly important. A multilingual pre-training algorithm, ERNIE-M, developed by Baidu, is the first pre-training algorithm to learn semantic alignment relations from a monolingual corpus, breaking the limitation of the influence of bilingual corpus size on multilingual models and providing a new idea for the research of cross-language semantic understanding. In addition to offering a wide range of potential applications, ERNIE-M can be extended to support other languages, thus allowing us to better analyze each language as a whole using the artificial intelligence system developed based on Chinese. It is also important to note that ERNIE-M technology can also help linguists and archaeologists to better understand endangered or lost languages as well as to preserve the cultures of ancient cultures.

A wide variety of applications and meanings can be attributed to ERNIE-M. Due to the fact that most of the world's current AI systems are based on English, ERNIE-M begins from the perspective of a Chinese speaker. A Chinese-based artificial intelligence system can now be used to improve products and services for users across the world using this technology, which can be applied to other languages.

However, according to existing papers and studies, usability validation on the ERNIE-M fine-tuning task is only performed for Cross-lingual Natural Language

[13] https://developer.baidu.com/article/detail.html?id=292593

Inference, Cross-lingual Named Entity Recognition, Cross-lingual Question Answering, Cross-lingual Parameter Identification, and Cross-lingual Sentence Retrieval tasks. An imperative component of natural language processing is the understanding, integrating, and the exploitation of multilingual texts as a part of natural language processing. Currently, there are many scenarios where it can be used and can be of substantial use for R&D (Research and development). Consequently, it is important to confirm that ERNIE-M can perform well in tasks related to the classification of multilingual text and that it can be put to practical use in terms of the development and practical application of distinct products.

Chapter 3

3 State of the Art

Natural language processing is a research area with several applications on how to use computers to understand and manipulate natural language (Chowdhury, 2005), and it is a pivotal area of exploration in the field of computer research and artificial intelligence research.

Bill Manaris, an American computer scientist, defined NLP in his article, which provides a comprehensive description of the nature and disciplinary orientation of NLP, and this generally agreed upon by current scholars:

> *NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems* (Manaris, 1998: 5).

With the booming development of Internet technology and the continuous penetration of information technology into offline fields, rich and diverse natural language data has been accumulated in more and more different application scenarios. The expansion of application scenarios poses new challenges to natural language processing technologies and capabilities. Academic innovations in natural language processing technologies have also given rise to new applications in industry, such as intelligent voice assistants, question and answer, and reasoning systems in professional and open domains. The application of NLP technologies in industry is at a new high point and faces unprecedented opportunities.

Multilingual tasks have been an area of great interest in natural language processing research. At the beginning of artificial intelligence, machine translation was listed as one of several classical problems in the field of artificial intelligence (Slocum, 1985). Machine translation, the jewel in the crown of artificial intelligence, has been studied for many years, producing a large number of academic results and

industry impact. In addition to machine translation tasks, academics are continuously working on other multilingual tasks such as bilingual lexicon extraction, multilingual sentiment analysis, multilingual text classification, multilingual lexical annotation, bilingual and multilingual word vector training, etc.

One of the implications of multilingual tasks is to help human users to communicate across languages, either between humans, or between humans and machines, and to obtain multilingual information in other languages. For example, multilingual searches can enable users to effectively retrieve content in other languages in their native languages in search tools, and obtain news, blogs, e-commerce information, etc. expressed in other languages; multilingual sentiment analysis can be used to understand the emotional tendencies expressed by speakers of other languages, such as the satisfaction of users in other countries with a certain product (Bhatt et al., 2015). Machine translation tasks can also be used to communicate directly with users of different languages through automatic translation.

This internship focuses on the application of machine translation methods to a multilingual text classification task. The multilingual text classification task attempts to use only the annotated training set on the source language to estimate or train a text classifier that can be used to classify a test set on the target language. This classifies the target language text without the annotated training data on the target language, or helps the user to obtain a specific class of text, solving the problem of insufficient manual annotation data on low resourced languages.

3.1 Machine Translation

According to current statistics, more than 7,000 languages exist in the world today (Austin, 2011). With the increasing development of economic globalization, people communicate more and more frequently. How to solve the language communication barrier is a common topic for everyone. Since ancient times, human translation has been the mainstream translation method. However, with the increasing maturity of computer technology and the rapid development of the Internet, machine translation technology gradually enters the historical stage, which is one of the earliest issues raised in the field of artificial intelligence and one of the core issues in the field

of natural language processing. As the name implies, machine translation is a process of translating one natural language into another natural language by using the efficient power of computers.

The origins of machine translation can be traced back to the late ninth century, when Arab cryptographers developed systematic language translation techniques used in modern machine translation, such as frequency analysis, probabilistic statistical information, and cryptanalysis (Cho et al., 2014). The idea of machine translation can even be traced back long ago, beginning in the late 1620s when René Descartes proposed a universal language in which the same ideas in different languages share a single symbol (Koehn et al., 2007).

In 1956, the first conference on machine translation marked a new phase in the field of MT. Since then, scientists around the world have been exploring machine translation technology. Both the Association for Machine Translation and Computational Linguistics and the Automatic Language Processing Advisory Committee (Alpac) were established in the United States, but there were no major breakthroughs in MT technology in the following decade. 1972 saw the Defense Research and Engineering Agency submit a report showing that their self-developed sign MT system had successfully translated an English military manual into Vietnamese, re-establishing the feasibility of MT. Since then, a number of researchers have made several achievements in MT technology.

In the late 1980s, breakthroughs in the quality of computer hardware also brought a reduction in computing costs. The emergence of various machine translation methods marked a greater progress in machine translation, and various machine translation competitions began. MT gradually developed from a research topic to practical application. From its origin to its development, the subject has become more and more popular in view of the great research prospect and commercial value of MT.

3.1.1 Rule-based systems

Traditional machine translation methods can be broadly classified into three categories: rule-based, example-based and statistical-based. rule-based machine translation dominated the field of machine translation before 1988.

The main rules used in rule-based machine translation include lexical rules, lexical conversion rules and syntactic generation rules, which generate dictionary query, syntactic analysis and text processing related techniques for the source language. And the conversion from the source language to the target language requires the help of dictionary queries.

To sum up, rule-based machine translation can maintain the structure of the original language and is more effective for the translation of linguistic phenomena or the original language with more standardized structure. However, due to the gradual expansion of the information volume of language systems, when using the rule-based approach, manual establishment of rules has problems such as large workload, subjectivity, and difficulty in guaranteeing consistency, which cannot adapt to large-scale new vocabulary and new sentence patterns.

3.1.2 Example-Based Machine Translation

Example-based machine translation (EBMT) was present in the 1980s and it is different from the rule-based approach, which relies more on a bilingual corpus to obtain the target language by comparing and matching with the corpus. The basic idea is to translate the input sentences into EBMT system matches, find the maximum source language sentence similarity with the bilingual corpus of input sentences, and generate the corresponding target language translations in a database of matched sentences in the standard translation examples (Nagao, 1996). However, the limitation of this approach is that the matching rate is very low and the labor cost remains high.

3.1.3 Statistical Machine Translation

Before the full application of deep learning to NLP tasks, statistical machine translation (SMT) was the most intensively studied and applied MT approach and has

dominated machine translation tasks. The concept of SMT was introduced by Warren B. Weaver in 1949 and was revived by researchers at IBM's Thomas Watson Research Center in the late 1980s and early 1990s. These researchers reintroduced statistical machine translation and improved its translation effectiveness, with better results in terms of translation quality.

The design idea of SMT is derived from information theory. Based on the probability distribution of strings in the target language corresponding to strings in the source language, the string with the maximum probability match is taken to translate the sentence. SMT contains three main parts: statistical language model, translation model and text decoder (Koehn et al., 2007). The core working mechanism is to train the translation model with a large amount of bilingual parallel corpus, based on which the decoder completes the translation task. Machine translation in mathematical terms can be explained as follows: given a source language **s** and a target language **t**, try to find the target language **t** that maximizes the following equation:

$$\text{argmax } \mathbf{t}, (P(\mathbf{s}|\mathbf{t}) * P(\mathbf{t}))$$

Compared with rule-based and example-based machine translation, SMT can make more efficient use of human and data resources, which include parallel corpora and monolingual corpora in readable formats; the time, money, and human costs consumed in manually developing language rules are expensive, and the same set of rules cannot be generalized for other languages. SMT uses a language model with a translation model which is more universal and robust.

SMT also has some areas for improvement. First of all, language models and translation models are more efficient and economical to use, but specific errors are difficult to predict and correct; since translation systems cannot store all native strings and their translations, long passages or documents are usually translated sentence by sentence; language models usually use smooth n-gram models, and similar methods are applied to translation models, but again sentence length and word order in languages are complex matters. The complexity increases by the differences on sentence length and word order across languages; therefore, SMT can neither translate in conjunction with contextual semantic scenarios, nor handle discourse rules such as word ordering, lexis, and syntax as well; meanwhile, since SMT works according to

the rule of counting the frequency of phrase occurrence in parallel corpus and selecting the best matching words, it does not reflect well the word similarity between them.

3.1.4 Neural Machine Translation

In 2015, Bahdanau's team first proposed to embed the attention mechanism into the end-to-end machine translation network model to solve the problem of "fixed-length vectors" in neural networks. The performance of Neural Machine Translation (NMT) is significantly improved because the focus of NMT memory can be determined by combining the contextual relevance.

Neural machine translation has achieved remarkable results in just a few years. Junczys-Dowmunt et al. (2016) tested NMT and SMT on a parallel UN corpus of 15 language pairs and 30 translation directions. The results show that NMT is equal to or better than SMT in 30 translation directions, and furthermore, most of the top-ranked translation systems in several world-class machine translation competitions are based on neural networks. The advantage of NMT over SMT is that multiple features can be trained jointly without prior knowledge and it can optimize sentence structure to obtain better translation results (Bahdanau et al., 2014).

Overall, NMT can solve the problems of word order errors, syntactic errors and morphological errors commonly found in SMT. On the other hand, NMT also faces some problems and challenges that need to be solved, such as the time-consuming training process and decoding process; there are cases that the translation results exceed the words in the word list; the same word may appear inconsistently under different corpus of translation styles; poor interpretability due to the "black box" mechanism of neural networks. The "black box" mechanism of neural networks leads to poor interpretability and therefore to ethics considerations in its usage.

3.1.5 Difficulties and Challenges

Machine translation technology is constantly evolving, and at the same time, it faces many challenges.

The first challenge faced by MT is the complexity of natural languages. Programming languages, such as the Java language, are very easy for machines to handle, because they are artificially defined and very rule-based. However natural languages naturally evolve with culture and human development and are an important tool for human communication and thought expression. It is dynamically evolving and is a microcosm of human culture.

The second challenge facing MT is the scarcity of data. There are still some machine translation methods that rely on large amounts of parallel data, which is relatively abundant for languages with a large number of speakers, like Chinese, English, and Japanese. However, for the majority of the world's languages, parallel data sets are very scarce. For example, Swahili in Africa has very few written scripts itself, not to mention its parallel correspondence with other languages, such as English.

3.2 Text Classification

As the network platform allows for rapid transmission speeds, it also allows for easy publication and for strong user interaction, therefore, a large amount of electronic data can be collected. The era of "information poverty" has been transformed into the era of "information overload".

Among the huge amounts of data, text-based data is one of the most basic and common data types, such as the conversation information of daily mobile communication software and the comments on shopping platforms. Text data occupies fewer space than video, audio, or image types of data and it is usually more descriptive than these data types, which is the primary purpose for which text data is used. This makes it an urgent problem to process the required information from the large amount of text data quickly and accurately.

Traditional solutions usually use humans to identify and classify data, which requires a lot of labor and takes time. The computer-led text classification technology can recognize and automatically train the classification of text data on the Internet, which helps to achieve effective human-computer interaction and promote the development of artificial intelligence.

Text classification by a computer can save manpower, facilitate information filtering and push and customize personalized services and is often used to handle tasks such as sentiment analysis, topic labeling, news classification, question answering systems, dialogue act classification, natural language inference, relationship classification, event prediction, public opinion analysis, spam recognition, pornographic and violent content recognition, etc. The research of text classification technology has bright application prospects and significant practical significance in the context of the current big data era. This is conducive to the efficient management and effective use of information.

Text classification can be traced back to 1958, when Luhn first defined "word frequency," counted the distribution information of the number of occurrences and positions of word phrases in an article and used it as a feature attribute of the text, and finally presented the statements with higher statistical values of the feature attribute as the summary of the article, establishing a statistical-driven automatic text summarization model (Luhn, 1958). In 1975, Salton et al. proposed a vector space model using a method based on spatial density calculation to select the most appropriate indexed vocabulary for a collection of documents (Salton et al., 1975). Classification methods constructed based on knowledge engineering techniques dominated the field of text classification in the late 1980s. The construction of classification systems is guided by hand-written classification rules built on the accumulation of knowledge (Wang et al., 2015).

By the 1990s, with the rapid development of the Internet, the number of electronic texts such as web pages, e-mails, BBS and BLOG on the Internet increased exponentially. Text classification has replaced knowledge engineering-based classification methods as the mainstream. Collobert and Weston published a neural network model for word vector training in 2008, which involved the concept of "word vector" (Collobert & Weston, 2008). Huang in 2012 proposed a method that can use the relationship between contexts to train word vectors (Huang et al., 2012). Mikolvo proposed the Word2vec word vector training model in 2013 (Mikolov et al., 2013). The proposal of Word2vec is a milestone. Until today, Word2vec is still a commonly used word vector model. Zhang and Wallace first used a CNN network model to classify textual data in 2015, and this method performed very well in multiple text

classification tasks (Zhang & Wallace, 2015). In July 2017, Shi and Bai et al. proposed a novel neural network architecture, the CRNN model, the performance of which was verified by practical problem processing (Shi et al., 2017). In 2018, Devlin proposed a fully bidirectional language model Bert model (Devlin et al., 2018), which can take into account the text sequence information, contextual relationship information, and grammatical context information in the entire sentence. In 2020, Saigal proposed a least squares double support vector machine (LS-TWSVM) method for text classification (Saigal & Khanna, 2020), with excellent results.

Figure 7: Text Classification Cloud[14]

3.2.1 Text Classification Overview

The so-called classification divides the objects from the same population into two or more category concepts. Under a given system, the categories of text associations are automatically determined based on the content of the text. From a mathematical point of view, the essence of classification is a mapping process. It maps uncategorized texts to existing categories. The text classification can be a binary-class classification, a multi-class classification or a multi-label classification. Mathematically expressed as follows:

[14] https://zhuanlan.zhihu.com/p/107721682

$$f: A \rightarrow B$$

In this formula, A is the set of texts to be classified, and B is the set of categories in the classification system.

The decision function f in the text classification system depends on the learning method of the classifier and different learning methods will produce different decision functions. One strategy for tackling such problems is for the computer to learn the correspondence of the input function from the examples, a process known as the learning method. If there is a given training data and results in the example, it is called supervised learning; if the data in the example does not contain the output, it is called unsupervised learning. Text classification is a typical supervised learning problem. Using the data set of the category given in advance in the training set, the relationship between the characteristics of the data set and its corresponding category label is mined and a classification model is established.

3.2.2 Traditional text classification methods

The main process of traditional text methods is to manually design some features, extract features from original texts, and then specify classifiers such as Linear Regression and Support Vector Machines. The common processes are: text preprocessing, feature extraction, classifier selection, and multiple Adaboost[15] training (Freund & Schapire, 1997). Traditionally used feature extraction methods are frequency method, TF-IDF, mutual information method, N-Gram.

By using the frequency method, the frequency distribution of each text is recorded. The distribution is then inputted into a machine learning model for training a suitable classification model. This will allow us to classify this type of data. It should be pointed out that when counting the distribution, a reasonable assumption can be made, since the impact of words with relatively small frequencies on text classification is relatively small. Therefore, we can reasonably assume a threshold, filter out words whose frequency is less than the threshold, and reduce the dimension of the feature space.

[15] https://www.sciencedirect.com/science/article/pii/S002200009791504X

Compared with the frequency method, TF-IDF has further considerations. The number of occurrences of words can reflect the characteristics of the text to a certain extent, namely TF. TF-IDF increases the so-called inverse text frequency. If a word appears more frequently in a certain category, but relatively fewer times in all texts, then this word has a stronger probability to distinguish texts. TF-IDF is a comprehensive consideration of frequency and inverse text frequency.

The mutual information method is also a statistical-based method that calculates the degree of correlation between the words appearing in the text and the text category.

The method based on N-Gram is to form a group of texts through a window of size N. Then make statistics on these groups, filter out the groups with low frequency, form these groups into a feature space, and pass them into the classifier for classification.

3.2.3 Deep learning methods

The concept of deep learning originated from the study of artificial neural networks and is a generic term for a class of learning methods based on artificial neural networks. It is a structure for combining low-level features into more abstract high-level features to create a distributed feature representation. In recent years, with the improvement of computer hardware and software equipment and the development of big data, deep learning has re-entered people's vision and gained wide attention in various fields, and has made significant achievements in the fields of speech and image. At present, text classification methods based on deep learning have become the mainstream research methods in natural language processing.

Since the features of text are difficult to extract, and these features do not represent the semantics and syntax of texts well, a large part of useful information is lost, leading to the limitations of traditional text classification methods. Since the 2010s, text classification has gradually transitioned from shallow learning to deep learning. In deep learning, this part of feature extraction is left to the neural networks to do automatically. This allows them to exchange better and more comprehensive text features, at the cost of a higher computational cost. However, compared with

traditional text classification methods, it is relatively weaker in terms of interpretation and controllability. Common deep learning-based text classification models include FastText, CNN, RNN, BERT, etc.

FastText is a text classification and word training tool from Facebook AI Research. As shown in figure 8, FastText is characterized by a simple model with only one hidden layer and an output layer. Therefore, training is very fast and can be achieved at the minute level on an ordinary CPU. This is several orders of magnitude faster than the training of deep models. At the same time, on several standard test datasets, FastText is comparable or close to some existing deep learning methods in terms of text classification accuracy.

.

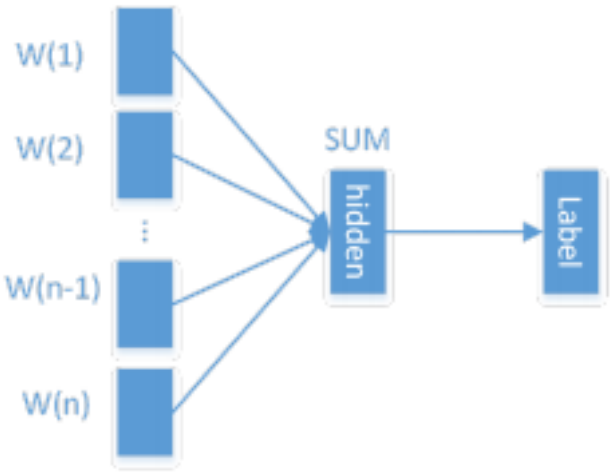

Figure 8: FastText process[16]

When CNN is used for text classification, the pre-convolutional convolutional layer of CNN is able to extract a large number of linguistic features, then pooling for feature decay, and finally SoftMax for normalization to obtain the final desired classification.

Although TextCNN does well in many tasks, CNN lacks the ability to model long sequence information due to its fixed filter/size field of view, and adjusting the super parameters of the filter/size is time-consuming.

[16] https://qianshuang.github.io/2018/08/25/word2vec-&-fasttext/

RNN's mathematical basis can be seen as a Markov chain, where the subsequent value is derived from the probability of having the former and some parameters. A RNN structure is well-suited to time series problems and is effective at capturing long-term time series; however, the plain RNN structure has gradient vanishing and explosion problems, wherein the gradient explosion can be solved by reduction, while the gradient vanishing problem appears to be stretched. So the LSTM network structure is derived, which can solve the gradient vanishing problem well. Since standard recurrent neural networks (RNNs) process sequences in temporal order, they tend to ignore future contextual information. So the Bi-RNN network structure was designed.

CNN, RNN and FastText have their own advantages. CNN is more obvious for feature extraction. RNN is better for word association, and FastText has the biggest advantage of superior performance and speed compared to the previous two models.

In 2018, Google introduced the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). Bert can be fine-tuned to be used for a wide variety of tasks by simply adding an additional output layer, without the need for task-specific model structure adjustment. The emergence of BERT, which can generate contextual word vectors, was an influential turning point in the development of text classification and other NLP techniques. The model offers better performance in several NLP tasks, including text classification.

Overall, it seems that shallow models improve text classification performance mainly by improving feature extraction schemes and classifier design. This is compared to deep learning models that enhance performance by using representation learning methods, model structure, and adding data and knowledge. Existing algorithmic models, both shallow and deep learning, have been tried for text classification, including integration methods. However, to improve model accuracy, the main approach is still to increase data, and it is worthwhile to investigate how to trade-off between increasing data and computational resources, and prediction performance. Secondly, the noise resistance of the results needs to be improved. How to achieve a reasonable evaluation of the depth model is also a technical challenge.

3.3 Cross-lingual Language Understanding

With the deepening of economic globalization and "Internet+", the international business of global companies is in full swing and the internationalization of products has become an inevitable trend. Artificial intelligence technology is also evolving, and artificial intelligence systems such as search engines, smart speakers, and intelligent customer service continue to bring better experiences to people. However, the construction of these systems often relies on a large amount of labeled data, and many systems are trained using a single language and cannot be directly applied to other languages, which is undoubtedly a huge challenge for machine understanding of small languages with sparse labeled corpora (e.g., Swahili, Urdu). How to build a unified language model to understand multiple languages is a hot research topic in the field of natural language processing in recent years.

Although most languages are data sparse, many of them share a large amount of underlying structures. There have been many studies in the past trying to exploit this shared structure in order to overcome the data sparsity problem. Research on the cross-language versions of pre-trained language models supported by BERT has grown even more rapidly in the last few years since the development of pre-trained language models.

3.3.1 Traditional cross-linguistic comprehension methods

It has always been a classic problem in the field of Natural Language Processing to deal with the problem of cross-lingual tasks. Prior to the invention of the pre-trained language models, it was considered a research topic within MT. During the course of MT, the core concept, which is known as alignment, refers to finding a word or phrase in the source language that can be matched in the target language in order to achieve translation. For example, we can say that "Lisboa" is aligned to "Lisbon", or that the two words are aligned to each other.

A classic MT model relies on a large number of parallel corpora (sentence pairs that contain both the original and translated text) in order to establish this alignment relationship between the two. In spite of this, for some small languages, it is very

difficult to collect relevant data for the purposes of research. As a means of overcoming this constraint, unsupervised machine translation (UNMT) technology is continuing to grow and develop, and has finally reached a historic milestone in 2018. In that year, Facebook AI Research (FAIR) extended "word alignment" to "word vector alignment", open sourced the unsupervised word-level translation model MUSE and, based on this, finally realized "sentence vector alignment" (Lample et al., 2018).

Related research on UNMT has demonstrated that all languages can share the same latent space. That is, two sentences with similar semantics from different languages, after encoding, their representation vectors can be aligned with each other. As long as the model can manage to model the alignment relationship at this level with a high degree of accuracy, it will be possible to minimize the distance between the two representation spaces of the source language and the target language, and the source language sentences will be able to be translated smoothly into the target language.

In the absence of parallel corpus support, however, and relying only on monolingual data to find the abstract semantic alignment relations, where does the model start to find those relationships when it is stripped of the parallel corpus support?

The Multilingual Unsupervised and Supervised Embedding method (MUSE) uses a mix of word lists from both languages as inputs for the embedding process as well as the Byte Pair Encoding method (BPE) for the subword cutting process (Joulin et al., 2018). In order to align the symbols across languages, punctuation and numbers, or the same subwords, it is necessary to use symbols shared across languages. It is a key assumption of word vector models that words with semantics that are similar to those that occur in the same contexts have equivalent meanings. There is a strong correlation between the semantic representation spaces of two identical text sources because they have similar word statistical distributions, which reinforces their intersection.

As a result of the introduction of pre-trained language models since BERT, UNMT has been able to reach a new level of performance as a powerful NLP tool. As

a result of this, the alignment relationships across languages are shown in a more accurate manner.

3.3.2 Cross-language pre-training models

3.3.2.1 Pre-training and fine-tuning models Introduction

Over the past few years, the terms pre-training and fine-tuning have been frequently used. Pre-training and fine-tuning have become new paradigms in natural language processing, and large-scale pre-training models have achieved state of the art results on a wide variety of tasks. It is possible to alleviate the low resource issue that occurs across tasks as well as across languages with the use of pre-training models.

What does the terms "pre-training" and "fine-tuning" mean? Consider the classification of multilingual text as an example. Consider the situation where a network model is being built for the purpose of performing a specific task of multilingual text classification. In the first step, the parameters are randomly initialized, and then the network is trained and continuously tuned until the loss of the network reduces. During the training process, the parameters initialized at the beginning are continuously changed. When the results are satisfactory, the parameters of the trained model can be saved. This is performed as such so that the trained model can be used to obtain better results for a similar task the next time. This process is called pre-training.

Later, when a similar multilingual text classification task is received, the parameters of the previously saved model can be directly used as the initialization parameters for this task, and then some modifications are made during the training process based on the results. In this case, a pre-trained model is used, and the process is fine-tuning.

This concludes that pre-training is referred to as a pre-trained model or the process of pre-training a model. In terms of fine-tuning, it means adjusting the parameters of a pre-trained model to match the data you have collected and applying it to your own dataset in order to improve the model's performance.

When dealing with multilingual text classification, it is difficult to design a text classification network that specifically supports each language, making it difficult to train a text classification network. Mostly, this is due to the lack of small language datasets with annotated and bilingual corpora. This is very relevant since the cost of annotation for rare languages is very high, which is the main reason for this problem. The potential risk of overfitting can happen if the dataset is not large enough and a reliable model is to be used.

So, the general operation is to train a model and then use that model as an initialization or feature extractor for similar tasks. For example, models such as Ernie (Sun et al., 2019), VGG (Simonyan & Zisserman, 2014), Inception (Szegedy et al., 2015), etc. provide their own training parameters so that one applies them to the fine-tuning stage. This saves time and computational resources while achieving better results very quickly.

3.3.2.2 mBERT

In November 2018, Google released Multilingual BERT, or mBERT for short, supporting 102 languages (Devlin et al., 2018). As a matter of fact, mBERT is not generally considered to be a comprehensive cross-language modeling experiment. The difference between mBERT and BERT is that it is pre-trained with a mixed-language corpus, and that is the only difference between the two. In addition, neither the model structure nor the optimization goals are customized in any way for the purpose of carrying out a cross-language migration.

There are two major baselines in mBERT that are very powerful. Translate-Train and Translate-Test, both use external MT systems that provide cross-language assistance. The former translates the training data (finetune) into the target language and the latter translates the test data into the target language. In this way, training and testing can be performed in the same language.

In addition, mBERT enforces strict language separation during training and testing, and hence it is a true "zero-shot" migration model. In terms of cross-language tasks, it is naturally less effective than the baseline model. BERT's greatest strength is that it allows for migration of knowledge across languages by simply replacing the

training data without explicitly aligning it with external data, which is what makes the model so strong.

In 2019, Pires analyzed Multilingual BERT's multilingual representational capabilities and came to several conclusions (Pires et al., 2019):

Firstly, BERT's multilingual representational capabilities are not only based on shared word lists, but also work very well for zero-shot tasks. The closer the language is to the native language, the better the results will be with regards to multilingual representations.

Secondly, for languages with different language orders (for example, subject-verb-object or adjective-noun), mBERT does not work well. The Multilingual BERT representation contains both representations that are common to multiple languages, as well as representations that are specific to different languages. This conclusion was also pointed out by Wu and Dredze (2019) in the language classification task: Multilingual BERT requires a certain language-specific representation in order to be able to select language words from the vocabulary since it is required to complete the language model task (Wu & Dredze, 2019).

3.3.2.3 XLMs

In 2019, Lample and Conneau proposed XLMs based on multilingual pre-training models (Lample & Conneau, 2019). First, some sentences are sampled from a monolingual corpus. The number can be increased for languages with scarce resources and reduced for languages with abundant resources, and all languages are represented through uniform byte pair encoding. To accomplish the learning objective, three language modeling goals are used. The first two are based on monolingual corpus and the last one is based on bilingual alignment data. The first one is Causal Language Modeling (CLM), which predicts the next word based on the previous word. As shown in figure 9. The second is Masked Language Modeling (MLM), which is similar to BERT, but uses a stream of words instead of sentence pairs. The third one is Translation Language Modeling (TLM), which can randomly mask out some of the words in both languages and then make predictions.

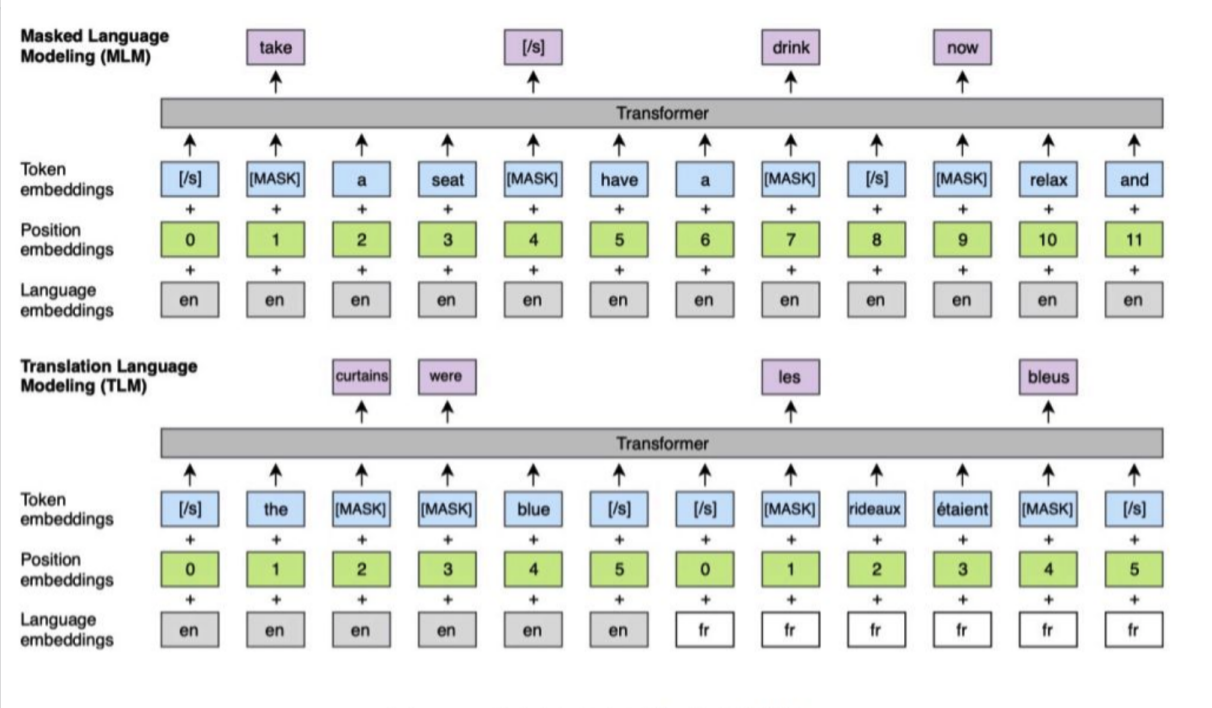

Figure 9: Cross-language model pre-training[17]

3.4 Multilingual text classification

There is growing recognition that cross-language text classification technology is of increasing significance, both in terms of its application and in terms of its contribution to research development. In addition to eliminating the difficulties associated with information retrieval and text classification due to language differences, it can make it easier for people to understand information, facilitate the exchange and sharing of knowledge, and contribute to the development of scientific research, the economy, and society in general.

Traditionally, text classification techniques and system tools have been designed for a single language, such as English. When texts are classified in different languages, it is necessary to train multiple monolingual classifiers. These monolingual text classifiers are then used to classify the languages supported by the system, which is a resource-consuming procedure, as multiple monolingual classifiers will need to be trained. Cross-language text classification is a technique that can be used to classify multilingual texts with a unified classifier, which can result in a reduction of classification costs, increased classification efficiency and improved classification

[17] https://arxiv.org/abs/1901.07291

results. Moreover, cross-lingual text classification is an important branch of the text classification discipline that plays an instrumental role in the further improvement of the theoretical structure of text classification.

In comparison to monolingual text classification, multilingual text classification is a relatively new field that has only been studied for a few years, with the earliest research results dating back to 2003. There was a study published in 2003 by Bel et al., which introduced the concept of cross-lingual text classification into multilingual and cross-lingual text classification for the first time (Bel et al., 2003). In 2010, Liu Zhihong proposed an automatic text classification system under a multilingual and multi-category system (Liu, 2010), and in 2011, Lu Weixiong proposed a multilingual text classification platform based on support vector machines (Lu, 2011). In 2016, Zhu Juan proposed a multilingual text classification system based on Bayesian algorithms (Zhu, 2016).

3.4.1 Model and theoretical framework

Currently, there are three main solutions and strategies for multilingual text classification: based on corpus, machine translation and bilingual word embedding.

3.4.1.1 Corpus-based approach

Multilingual corpora can be divided into parallel corpus (Peng, 2014) and comparative corpus (Wang, 2013). Parallel corpus refers to the same textual information described in different languages, i.e., multilingual textual information is translated into each other on the same content. Comparative corpus refers to information on the same topic described in different languages, i.e., multilingual textual information described similarly under the same topic.

Potthast et al. proposed a CL-ESA multilingual text classification algorithm based on a parallel corpus, which exploits the semantic alignment between concepts of multilingual Wikipedia entries and represents cross-lingual text similarity by computing the similarity between semantic vectors of the entries in their respective language spaces (Potthast, et al., 2008). Gliozzo proposed a multilingual domain model from a comparable corpus to classify texts whose target language is Italian, and achieved better classification results (Gliozzo & Strapparava, 2005). Xingzuo

proposed a multilingual text representation and classification method based on word vector clustering by mining the characteristics of word vectors in multilingual parallel corpora, and achieved better classification results in Chinese English-French multilingual text classification (Liu, 2016). Lu and Tan learn monolingual classifiers based on sentence-aligned parallel corpora in a cross-lingual sentiment classification task and then combine them for classification (Lu et al., 2011).

3.4.1.2 Method based on machine translation

Machine translation-based algorithms are a simple and straightforward approach that maps sets of texts from different languages to the same language space by means of existing translation software. MT based approaches can be divided into full text translation (Rou, 2015), terminology translation and feature translation approaches.

Full text translation is the translation of an entire text with the help of a machine translation tool. When there is an established machine translation tool between the source language and the target language, a suitable one can be selected to translate the text set of the other language into the text set of that language according to the requirements. When there is no available machine translation software between the two languages, you can unify the language space with the help of an intermediate language.

Terminology translation is to extract the important terms in each category to form a terminology list, and then translate this terminology list. Feature translation is to perform text classification on the source language text, and then translate the feature items actually used in the classification process into the target language.

In 2003, Koster conducted experiments on Spanish and English corpora by both term translation and feature translation to validate the method (Koster, 2003). In 2005, Rigutini combined machine translation and EM algorithms for cross-language text classification on English and Italian texts (Rigutini et al., 2005). After that, Hanneman improved the correct rate of multilingual text classification by constructing syntax-based full-text translation algorithms (Hanneman & Lavie, 2011), and Prettenhofer achieved cross-lingual sentiment classification by translating feature words (Prettenhofer & Stein, 2011).

3.4.1.3 Bilingual word embedding based approach

In recent years, with the development of deep learning technology, neural network-based feature representation models have gradually formed a new research direction. Luong proposed a bilingual Skip-Gram model (BiSkip) for cross-lingual text classification, which extends the Skip-Gram with Negative Sampling (SGNS) model from monolingual to bilingual, which enables cross-language prediction of words and can effectively obtain high-quality bilingual representations (Luong et al., 2015). Mikolov proposed a distributed word vector representation model and demonstrated that word-to-word correlation can be represented by similarity between vectors (Mikolov, Sutskever, et al., 2013). It was found that word vectors trained by different languages were independent but their distribution patterns were very similar. In the same year, he transformed word vectors between different languages by linear transformation and gradient descent algorithm to obtain bilingual word embedding vectors, which improved the accuracy of multilingual intertranslation (Mikolov, Le, et al., 2013). In 2015, Vulić and Moens used a bilingual corpus based on word embedding for training and obtained multilingual word vectors (Vulić & Moens, 2015). Ouyang proposed a dependency-based bilingual word embedding model by constructing syntactic dependency trees of sentences and performed cross-language text classification on English-German, English-French and English-Spanish texts (Ouyang, 2018).

The field of multilingual text classification is still immaturely developed and there are still some problems that can be optimized.

The corpus-based approach has high requirements on the construction of a corpus with comprehensive coverage and text alignment for newly added test sets, which can cause great limitations in the experimental process and is not conducive to expansion.

Although the method based on MT is simple, the differences in cultural backgrounds of different languages can lead to thematic drift in the translation process, and the effect of classification is heavily dependent on the accuracy of machine translation, resulting in lower efficiency (Shi et al., 2016). The bilingual word embedding approach can obtain accurate semantic information and get specific

feature representations, but it is more difficult to extend to multiple languages (Liu, 2018).

At present, most of the existing text classification systems are oriented to single language, and multilingual text classification systems are rare in China, while using traditional multilingual text classification methods, the work cost is high. How to apply deep learning methods to develop multilingual text classification systems is one of the key points of my internship in this field.

3.4.2 Evaluation Indicators

For multilingual text classification tasks, Accuracy(A), Precision (P), Recall (R), and F1-Measure are usually used as performance metrics to measure the classification effectiveness, two of which are shown in Table 1.

Actual category	Classifier discriminant results	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Table 1 Binomial classification results

The above table represents the confusion matrix for a binary classification task, where TP indicates that the class of the samples is a positive class and the total number of sample points whose classification prediction is also positive class; FN, FP, TN are defined similarly to TP.

Accuracy: the proportion of sample points in the set of sample points that are correctly classified, takes a value between 0 and 1, the larger the value the better:

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

Precision: the proportion of sample points whose true class is also positive in the set of all samples whose classification prediction results in positive class:

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

Recall: The proportion of sample points that are correctly classified in the set of samples where all true classes are positive, taking a value between 0 and 1, with larger values being better:

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

F1-score: the summed average of Precision and Recall. For classification models, the larger the F1-score, the better, but it should not exceed 1:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

ROC: Receiver Operating Characteristic Curve, whose vertical coordinate is the true rate (TPR) and horizontal coordinate is the false positive rate (FPR). Where:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

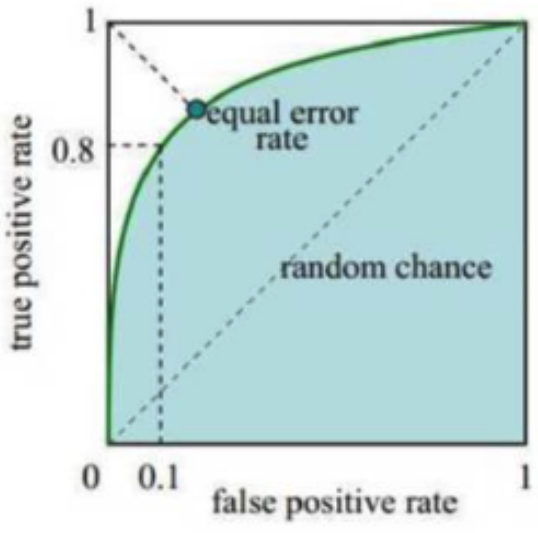

Figure 10: Schematic diagram of ROC curve[18]

[18] https://dreamocean.github.io/2017/07/25/index/

There are four points in the ROC curve that are worth noting.

(0,1): FN=0, FP=0, indicating that all samples are correctly classified.

(1,0): TN=0, TP=0, indicating that all samples are misclassified.

(0,0): FP=0, TP=0, indicating that all samples are classified as negative classes.

(1,1): TN=0, FN=0, means all samples are classified as positive class.

The more the ROC curve is skewed to the upper left corner of the classifier, the better the results are.

Area Under Curve (AUC) represents the area under the ROC curve. The advantage of AUC is that the area under the ROC curve can be specifically calculated as a value. The value of AUC is positively correlated with the effect of the classifier, and the value of AUC is taken in the range of [0.5,1].

For the evaluation metrics of the model dealing with multi-categorization problems, the selection method is basically the same as that of binary classification, only some preprocessing is needed. Assuming that there are n categories in a data set, firstly, the first category is regarded as a positive category, the second and third categories up to the nth category are regarded as negative categories, and various indicators are calculated; then the second category is regarded as positive category, the first category, the third category, the fourth category up to the nth category are regarded as negative categories, and various indicators are calculated; and so on, until the nth category is considered as positive category and the first n-1 categories are considered as negative category, calculate the various indicators needed, and finally perform the average.

Calculate the precision, recall, and F1 values for class i, respectively (treating class i as a positive class and the others as negative classes).

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1_i = 2 * \frac{Precision_i * Recall_i}{Precision_i + Recall_i}$$

There is a recall rate and a success rate corresponding to each category in the classification results, so the overall performance of the classification can be assessed based on the results of each category in the classification. There are two usual methods: Macro-averaging and Micro-averaging. In the macro-averaging process, the correct and recall rates are first calculated for each category, and then the correct and recall rates are averaged separately to obtain the total correct and recall rates. The micro-averaging concept means that the total value of correct and recall values is calculated directly based on the formulas for correct and recall values. It could be said that macro-averaging can reflect more on the effect of classifying some special classes, whereas micro-averaging is more influenced by the effect of the classifier on classifying some common classes (which typically have a larger corpus of data). In order to compare multiple algorithms, the micro-averaging algorithm is usually used.

(1) Macro-averaging

Calculate the precision, recall and F1 values for class i, respectively, and then average.

$$Precision_{macro} = \frac{\sum_{i=1}^{N} Precision_i}{|N|}$$

$$Recall_{macro} = \frac{\sum_{i=1}^{N} Recall_i}{|N|}$$

$$F1_{macro} = 2 * \frac{Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}}$$

(2) Micro-averaging

$$Precision_{micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}$$

$$F1_{micro} = 2 * \frac{Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

(3) Weighted averaging

The weighting is to deal with the problem that the macro-average Macro is more affected by categories with small sample size. Where, w_i denotes the number of i-th sessions as a proportion of the total number of samples.

$$Precision_{weighted} = \frac{\sum_{i=1}^{N} Precision_i * w_i}{|N|}$$

$$Recall_{weighted} = \frac{\sum_{i=1}^{N} Recall_i * w_i}{|N|}$$

$$F1_{weighted} = 2 * \frac{Precision_{weighted} * Recall_{weighted}}{Precision_{weighted} + Recall_{weighted}}$$

Chapter 4

4 Use Case Analysis of Multilingual Text Classification

The NLP department of Baidu, where the internship took place, released in 2021 ERNIE-M, a cross-language understanding model based on back-translation that aims to learn semantic alignment relationships between languages by analyzing a monolingual corpus, and which is significantly more effective than before for five typical cross-language comprehension tasks, including natural language inference, semantic retrieval, semantic similarity, named entity recognition, and reading comprehension. It is important to note that in addition to this application scenario, the pre-trained model can also handle multilingual text single-label classification tasks by pre-training and fine-tuning the data used to create the model.

During the internship, we studied multilingual text classification models and discovered that the three current areas of invalid comment recognition, news classification, and sentiment analysis are those that are being used more often for internationalized products at the moment. As a result, we have prepared data, trained data, deployed models, and optimized the results in Baidu AI's Easy DL platform for these three scenarios, and finally summarized the experimental findings.

Throughout the remainder of this chapter, we will introduce the background of the pre-training model Ernie-M that has been used in the classification of multilingual texts, its implementation principles, and its efficacy in a variety of cross-language understanding scenarios. After that, we will present the work conducted during the internship, which was related to the classification of multilingual texts. On the basis of the pre-trained model, we investigated the generality and robustness of ERNIE-M in the multilingual text classification market, applied to three broad scenarios. In an experiment, there are mainly three steps involved: data collection, model training, and analysis of the results.

4.1 Research on multilingual text classification application scenarios

Multilingual text classification has a number of significant applications in a variety of different domains, each of which has their own unique set of benefits. Its

primary application capabilities include accurately analysing users' emotional tendencies expressed through text; extracting and categorising topics from identified text; and identifying advertisements, politically sensitive content, pornographic content, uncivilised language, and low-quality spam text. These scenarios are based on the three capabilities, which in turn can also be subspecified.

I: Specified text recognition: To determine whether text in each language contains specific types of content, in actual use scenarios, which is typically used for auditing text content by identifying specific types of content. Specified text recognition is used to determine whether text in each language contains specific types of content.

1. Detection of pornographic text includes the ability to determine mechanically whether or not a given paragraph of text in each language contains pornographic vocabulary.

2. The feature that detects whether or not the text in articles, reviews, emails, and other scenarios in each language contains violent or abusive content is known as violent text detection.

3. Detecting advertisements involves determining whether the text in various contexts, such as articles, reviews, emails, and other scenarios, is an advertisement.

4. Invalid comment detection: this function checks to see whether the reviews in each language contain any invalid reviews that were written in order to drive up the heat by utilizing the guidelines provided by the platform.

5. Identification of spam, such as the automatic recognition that a certain language's e-mail is not spam.

6. Recognizing politically sensitive material requires first determining, in an automated manner, whether or not a given chunk of text in each language contains politically sensitive material.

II: The Extraction of Text Topics: This task involves the automatic categorization of texts in each language using a classification system that has been predetermined.

1. Classification of the news involves automatically determining different news categories, such as political, economic, lifestyle, and sporting events, amongst others.

2. Automatic allocation of telephone inquiries. In the case of the government affairs scenario, the categories of user feedback, such as environmental pollution, water pollution, noise problems, road congestion problems, and so on, are automatically identified and assigned to the corresponding processing departments for processing based on the type of problem. For example, in the case of environmental pollution, water pollution, noise problems, and road congestion problems, etc., the telephone inquiries are automatically allocated.

3. Classification of tax issues involves determining the nature of the problem based on the questions posed by the user, such as whether it is an APP issue or a website issue.

4. Customer service / chat topic classification consists of the automatic identification of customer service chat with users in the process of user feedback problem categories, such as "is the return of goods," "logistics problems," and "product quality concerns."

5. Case description classification: Based on the information entered into the case by the police, the system can automatically determine the type of case that is being investigated, such as SMS fraud, network fraud, acquaintance fraud, and so on.

6. Email auto-response: In cross-border e-commerce scenarios, automatically identify the category of user feedback issues, such as product quality issues, delivery issues, product description issues, and so on, and automatically reply to the content based on the issue category. For example, "product quality issues" could be identified as "product delivery issues," "product description issues" could be identified as "product description issues," and so on.

7. Classification of events by type: Consists on automatically ascertaining the types of events, such as those including news items from the financial sector that are involved in the automatic identification of the types of events: appointment, resignation, increase, reduction, and meeting held.

8. Identification of fraudulent SMS involves performing an automatic check to establish whether or not the SMS that was delivered was fraudulent.

III: Sentiment analysis: Assess the inclinations indicated by users in each language throughout the text to determine how they feel about a certain topic.

1. Identifying positive and negative product reviews involves doing an analysis that determines, on an automated basis, whether a user's review of a product is a favorable or negative review.

2. Willingness to buy recognition is the process of determining if a user is willing to purchase a particular product based on information that the user has provided, such as tweets that the user has published.

3. Customer service chat sentiment analysis: Based on the user and customer service chat content for sentiment analysis, assess the emotional attitude of users and customers, with the results being able to be utilized as one of the dimensions of the evaluation between customer service.

4. Prediction of a proclivity toward suicide behaviour is accomplished by analysing the user's social media postings to determine whether or not the user is depressed or suicidal, among other things.

There are many additional applications of multilingual text categorization that can be created, in addition to the ones that have been listed above; however, due to the limited amount of time available for investigation, only the scenarios that have been given above are discussed in this part. A single scenario will be extracted from each of the three major categories that we will discuss in the following sections, namely fake reviews detection, news classification, and positive and negative product review recognition. We will then use Baidu's ERNIE-M model to train multilingual text classification on Baidu's AI platform EsayDL using Baidu's EsayDL Platform.

4.2 Fake Reviews Detection

4.2.1 Project Description

The amount of information available online is continuously expanding in tandem with the rapid development of the internet. Due to the fact that users have the freedom to express themselves in situations such as e-commerce, live-streaming, the scope of the comment content that can be entered is not limited. As a result, it is unavoidable that high-risk content such as pornography, violent content, and invalid comment content will be entered. Many information companies have established content auditing departments in order to monitor the network environment and filter out the reviews that are truly meaningful. These departments are responsible for auditing the reviews that users upload to the website and determining whether or not they contain content that is illegal or invalid. This helps to reduce the risk of business violations while also ensuring that reviews are effective and that users have a positive experience.

It is difficult to obtain the necessary raw data since the offensive words containing pornography, violence have been filtered on a variety of platforms. This makes it more difficult to identify offensive content. Consequently, multilingual invalid reviews will serve as the application scenario that will be tested throughout this thesis.

The first thing that has to be done is to identify the various kinds of invalid reviews. In the commenting situation, the types of evaluations that occur the most frequently are useless symbolic reviews, such as reviews that just contain numbers or only contain punctuation. The second kind of review consists of some written reviews that are pointless. On many different platforms, people who publish reviews that are longer than the required number of words can be rewarded. This is done in an effort to encourage users to comment on content. For instance, the "Meituan" app for local life services frequently contains some false reviews that have been written inside of it in order to earn cash back from the seller. Invalid reviews are considerably more difficult to identify when they are attached to the original reviews, which do not have any tags. It must first go through the manual review process. The third type of comment is one

that displays a significant gap between the rating and the sentiment conveyed by the commenter. It's important to note that this form of comment is also deemed to be an invalid kind of comment. Because of this, the kind of comment in question needs to be analyzed to establish if the feeling conveyed by it is favorable or unfavorable, and its rating must also be taken into consideration before the appropriate comparison can be evaluated. This also requires human auditing in multiple languages.

The vast majority of early comment detection is accomplished through straightforward and ruthless manual review, which is labor that takes up a lot of time. However, as the number of reviews increased, the cost of manual review became higher and higher, and as a result, the use of machine filtering according to rules, in conjunction with the manual method of identification, became the standard practice. The machine-to-manual method generally entails the use of automatic filters to filter out the majority of the text which can be neutral, after that the remaining text is subjected to manual review, which can significantly cut down on the amount of time and effort required for the process. The cost of manual review is still quite high, since rule-based filtering has continued to show limitations. It is of the utmost importance to develop AI models capable of identifying fake comments. There is currently no product on the market that provides a solution for universal text categorization that is capable of satisfying the requirement to recognise multilingual text content. In some languages with uncommon resources, it is not possible to use common models to determine whether or not the reviews are genuine; therefore, special training of models is required for this scenario.

Building up an AI team expressly for this activity, in order to meet business objectives as an independent firm, entails setting up a team, obtaining machine resources, and matching operation and maintenance people, at a total cost of millions of dollars. With the help of the EasyDL-Multilingual Text Classification (Single Label) task, the creation of an invalid comment detection model was finished during the internship in just one week. This model had the potential to be applied to several business scenarios.

4.2.2 Data Preparation

The data for this experiment comes from 5902 multilingual reviews of Airbnb[19] that were posted on Google Play between June 1 and June 30, 2022 and were exported from AppBot[20]. These reviews were written in languages such as Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, and Dutch. The data includes references to 34 different languages, including English and Finnish. Only 200 of them are labeled and used to train the model, while the remaining are used for testing the model. The 200 data that are used to train the model were randomly chosen.

Reviews	Language	Sentiment
Ansich würde ich gerne Airbnb nutzen, nur wenn man mal ein Konto hatte und eine Telefonnummer wiederlegt ist und man das Passwort nicht mehr weiß hat man keine Chance das man das Konto wiederherstellt	German	Negative
Muy efectiva la app	Spanish	Positive
Excelente aplicación, nunca nos ha fallado.	Spanish	Positive
Super app. No me ha quedado mal en ningún viaje.	Spanish	Mixed
Es presta vez q estoy usando esta página y quiero alquilar un departamento en Ambato Y quiero ver sie pueden ayudar gracias	Spanish	Mixed
Airbnb used to be such a good deal and experience. Reasonable prices have turned expensive and on top of that it's incredibly frustrating to see a price on the app and find a $40-$150 cleaning fee and $30 service fees hidden away until you are at the checkout. Adding all these hidden fees so late makes it harder to find a nice place to stay. I find myself more and more often using other services and hotels, without all the hidden fees, as it's easier and doesn't waste nearly as much of my time.	English	Negative
좋아요	Korean	Positive
Mai più con Airbnb, host inesistente, numero sbgliato	Italian	Negative
Works for me	English	Positive
Buen cerbicio	Spanish	Positive
واو	Arabic	Positive

[19] https://www.airbnb.com/
[20] https://appbot.co/

Table 4: Example of original dataset

On the EasyDL platform of the Baidu AI system, we next produce the invalid review detection dataset names as Fake_Review_Detection.

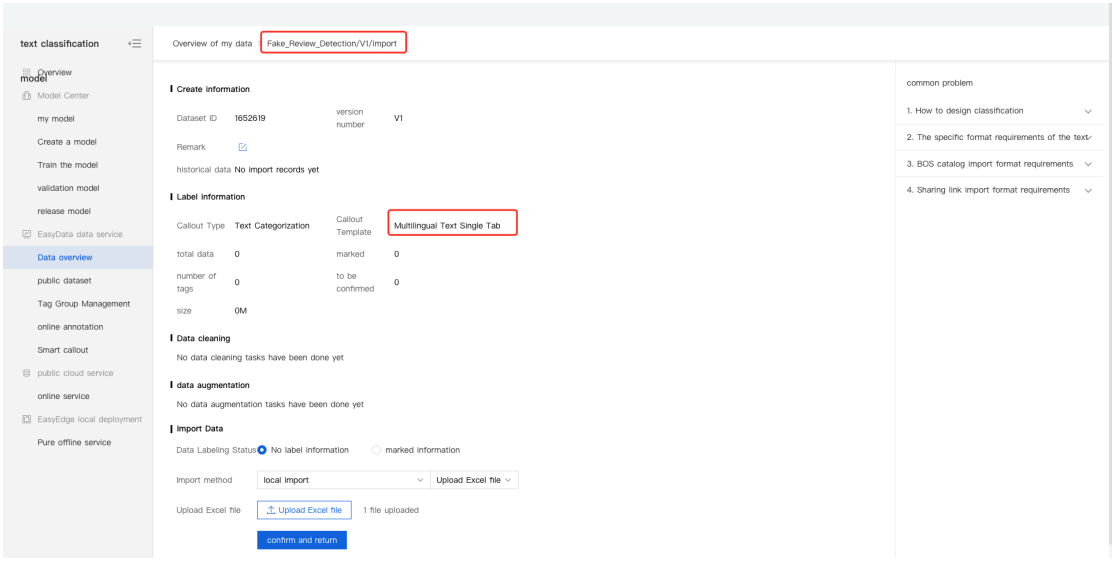

Figure 15: Create dataset of Fake_Review_Detection on EasyDL

Since the original data is not annotated, we need to set two tags FAKE and TRUE on Baidu AI platform EasyDL for manual annotation.

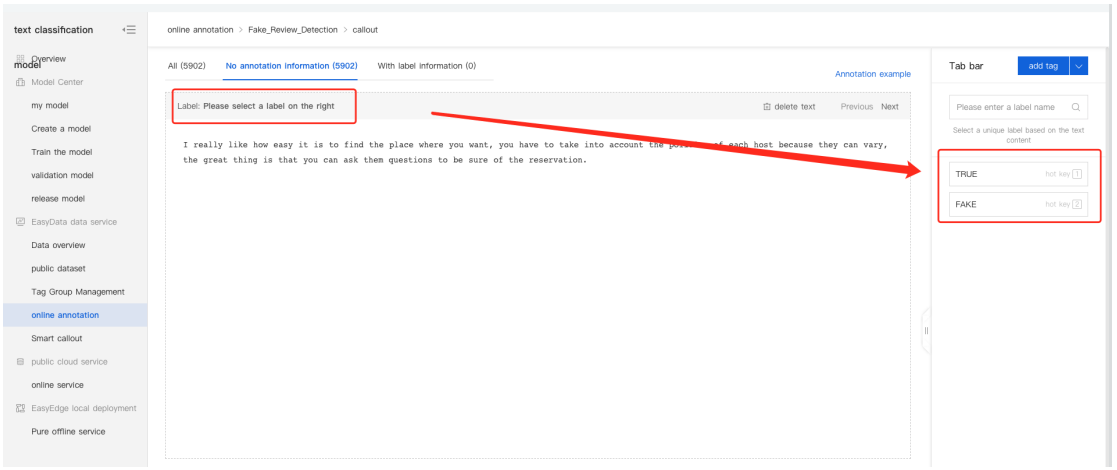

Figure 16: Online annotation of reviews on EasyDL

Due to the multiplicity of source languages, we used the Google Translate plugin to translate the to-be-annotated reviews into English, in order to facilitate the

annotation process. Here, we randomly annotated 200 pieces of multi-lingual reviews, of which 193 were true reviews and 7 were fake.

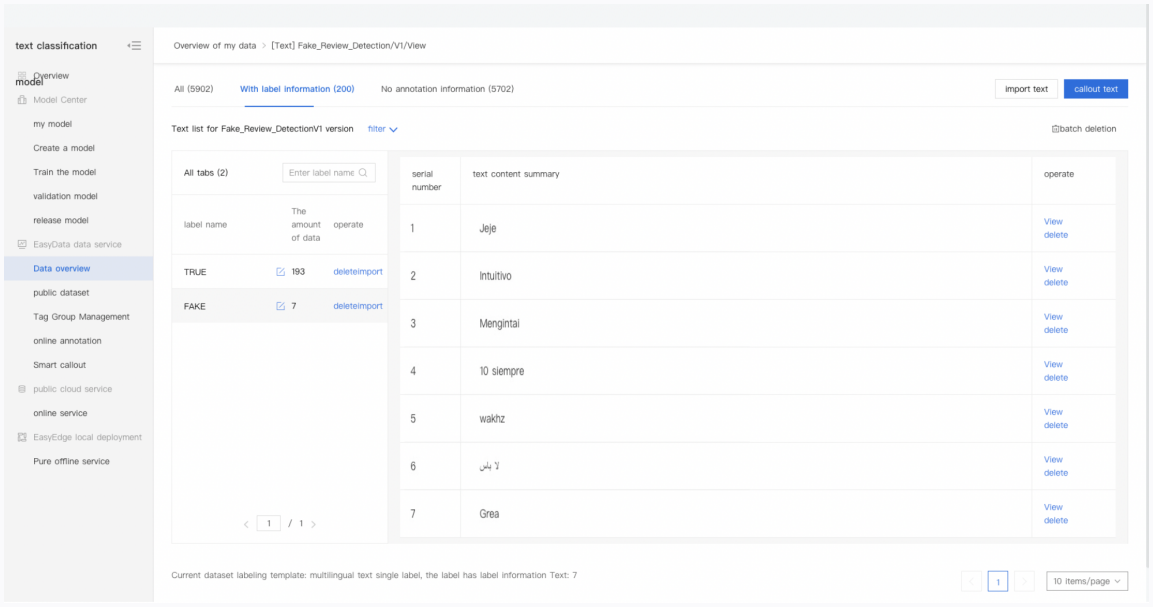

Figure 17: Dataset overview of Fake_Review_Detection

4.2.3 Model Training

After labeling the dataset, we created the FakeReviewDetection model, added a dataset containing both FAKE and TRUE labels, and clicked [Train Model] to initiate the training. Since we only use it for exploratory testing, it requires fewer resources, therefore we train it on the public cloud deployment directly. When using it, businesses can incorporate the online API into their enterprise app or website.

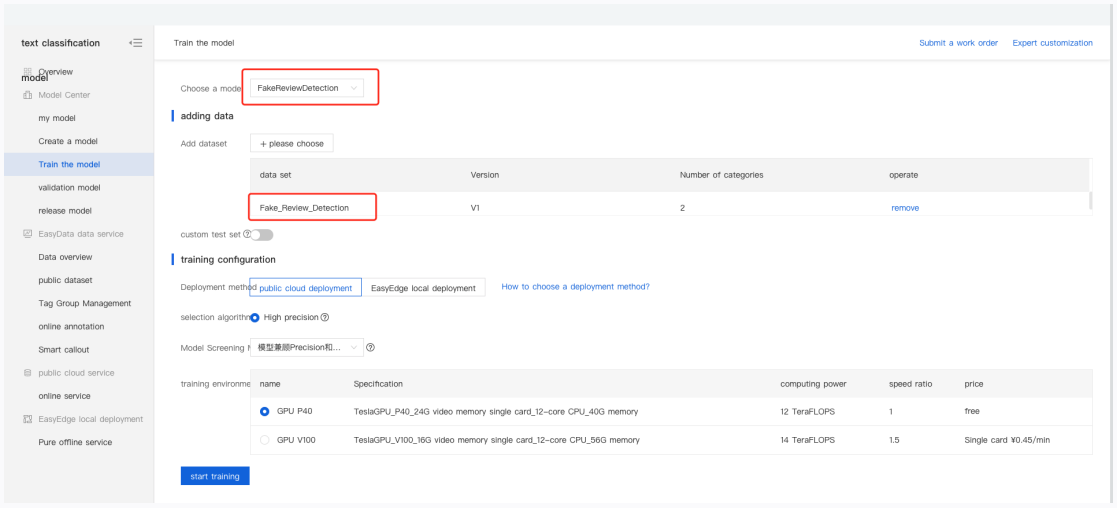

Figure 18: Train the model FakeReviewDetection on public cloud

Since we are using it as an experiment with no special latency requirements, we select a high-precision approach for training, so we may receive high-precision findings, even with a smaller training dataset. When selecting the filtering metrics for the trials, we do not have any specific criteria, thus we select the accuracy and recall balancing method, which is the platform default metrics.

4.2.4 Effect Analysis

When the model is trained, EasyDL generates an overall evaluation report that includes accuracy, F1-score, precision and recall.

The results of the test indicate that the model has a good level of accuracy, but nonetheless, its F1 score, recall, and accuracy are only average. Because the two tabs were broken apart and looked at separately, a better metric for TRUE reviews was made.

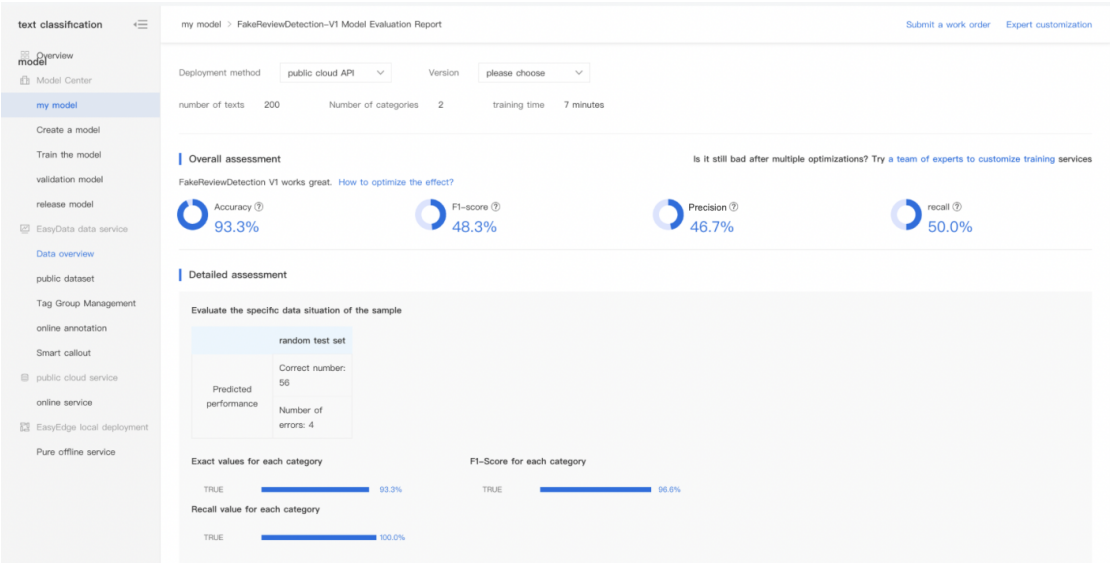

Figure 19: FakeReviewDetection model assessment

It can be seen from the above data that the number of FAKE reviews with annotations was too small during the preliminary data preparation, resulting in the algorithm not being able to learn the features of FAKE reviews accurately. Therefore, the data under the Fake reviews tag needs to be supplemented to ensure at least 50

items. However, there are relatively few fake reviews posted by users on Google Play, so it was decided to choose a new data type, news classification, to verify whether the classification effect of ERNIE-M is good enough when the term pre-trained test data with annotation is rich enough. However, with these experiments, we could understand the impact of having unbalanced or fewer examples of certain categories and the effects of such in the metrics performance.

4.3 News Headlines Categories Classification

4.3.1 Project Description

In recent years, the rapid growth of digitalization in the news industry and the popularity of news network platforms have substantially satiated the public's demand to know what is happening around the globe. In recent years, the news has become an increasingly essential source of information and knowledge about current events. In addition, the demand of people to know what is happening around the globe has been mostly satisfied. The volume of text data, such as news stories and reviews, as well as the voices of internet users, are continuously expanding on online platforms. Due to the fact that correctly classifying this text data enables better organization and utilization of this information, it is of the utmost importance to complete the categorization of news as soon and precisely as possible.

Since the beginning of traditional newspaper media, manual classification has always been an option. Due to the widespread use of networking and the frequent distribution of new information, it is impracticable to rely on human classification of the large volume of multilingual text data. As a result of the popularity of networking and the high frequency of news updates, there is an enormous accumulation of news information on news network platforms. In addition, the computer is capable of acquiring knowledge and skills through continuous learning as well as appropriately categorizing unknown challenges. Existing news classification models only handle a single language or a limited number of languages, and the capacity to categorize news headlines in several languages is currently insufficient. Consequently, there is an urgent need for a pre-trained model that can classify news with headlines in different

languages. In turn, this helps consumers utilize the information more effectively and gain quick access to news that interests them.

4.3.2 Data Preparation

The experimental data in this paper uses Misra Rishabh's News Category Dataset[21], which contains about 200,000 English news headlines obtained from HuffPost from 2012 to 2018, delineated into 41 candidate classification categories.

Category	Count	Category	Count
POLITICS	32739	CRIME	3405
WELLNESS	17827	MEDIA	2815
ENTERTAINMENT	16058	WEIRD NEWS	2670
TRAVEL	9887	GREEN	2622
STYLE & BEAUTY	9649	WORLDPOST	2579
PARENTING	8677	RELIGION	2556
HEALTHY LIVING	6694	STYLE	2254
QUEER VOICES	6314	SCIENCE	2178
FOOD & DRINK	6226	WORLD NEWS	2177
BUSINESS	5937	TECH	2082
COMEDY	5175	MONEY	1707
SPORTS	4884	ARTS	1509
BLACK VOICES	4528	FIFTY	1401
HOME & LIVING	4195	GOOD NEWS	1398
PARENTS	3955	ARTS & CULTURE	1339
THE WORLDPOST	3664	ENVIRONMENT	1323
WEDDINGS	3651	COLLEGE	1144
WOMEN	3490	LATINO VOICES	1129
IMPACT	3459	CULTURE & ARTS	1030
DIVORCE	3426	EDUCATION	1004

[21] https://www.kaggle.com/datasets/rmisra/news-category-dataset?resource=download

Table 7: Category distribution of the dataset

In this experiment, the data of the first 10 categories were selected for testing, including: POLITICS, WELLNESS, ENTERTAINMENT, TRAVEL, STYLE & BEAUTY, PARENTING, HEALTHY LIVING, QUEER VOICES, FOOD & DRINK, and BUSINESS. DRINK, BUSINESS.

Category	News Headline
ENTERTAINMENT	What To Watch On Amazon Prime That\u2019s New This Week
ENTERTAINMENT	Justin Timberlake Visits Texas School Shooting Victims
POLITICS	Edward Snowden: There's No One Trump Loves More Than Vladimir Putin
POLITICS	Ireland Votes To Repeal Abortion Amendment In Landslide Referendum
ENTERTAINMENT	Twitter #PutStarWarsInOtherFilms And It Was Universally Entertaining
POLITICS	Jack Johnson Was Pardoned, But Taboo Sex Is Still Being Criminalized
ENTERTAINMENT	Kanye West Spent $85,000 On Photo Of Whitney Houston's Bathroom For Album Cover
POLITICS	How The Chinese Exclusion Act Can Help Us Understand Immigration Politics Today
ENTERTAINMENT	People Are Rattled By How Much This Nigerian Man Looks Like Chadwick Boseman

Table 8: Example of News Category Dataset

ERNIE-M performs multilingual text classification in any language and then classifies the multilingual text. Therefore, for each of the above 10 categories, 50 English headlines were selected as the training data for the news headline classification model. Since the 500 headlines are already labeled with categories, the training can be launched directly after selecting the labeled text when uploading as the figure below.

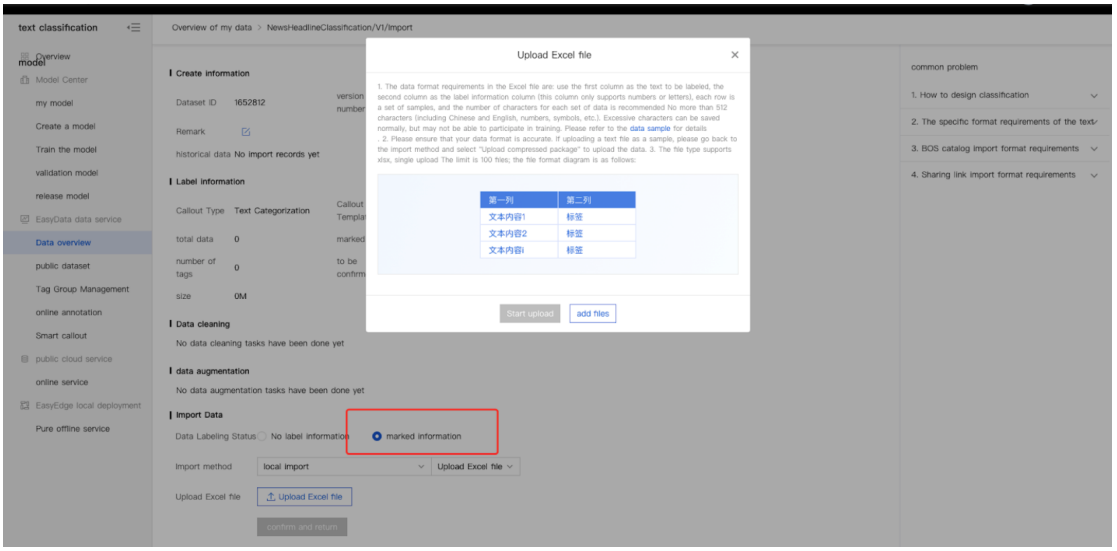

Figure 20: Create dataset with marked information

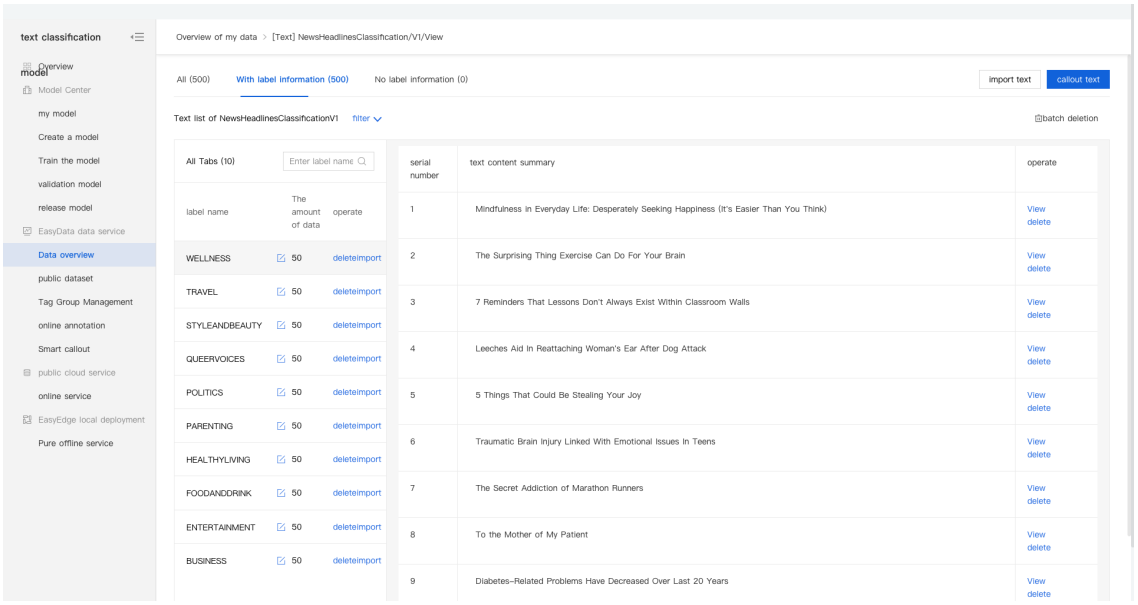

Figure 21: Data overview of NewsHeadlinesClassification

This batch of data is pure English text, in order to verify the effectiveness of the model for multilingual text classification. For each of these 10 groups of news headline types, I selected 1000 pieces of data and translated them into 200 pieces of Portuguese, 200 pieces of Spanish, 200 pieces of French, 200 pieces of Swedish, and 200 pieces of Japanese using Google Translate[22]. Thus, there are 10,000 multilingual

[22] If using Baidu translation, the model may be able to learn the semantic pairs of relationships between these data as they are translated, resulting in a better cross-linguistic understanding of the tested ernie-m model than the actual ability.

news headline text data as test data, which are 2000 Portuguese, 2000 Spanish, 2000 French, 2000 Swedish, and 2000 Japanese.

News Headline	Category
Omdefiniering av framgång: Företagsstegen spelar ingen roll längre	WELLNESS
Kan du inte somna? Prova detta	WELLNESS
Vad är värre: Äta inga grönsaker eller äta ostiga, smöriga grönsaker?	WELLNESS
ARVD: Min gåva	WELLNESS
メガ ミリオンズ ジャックポットは火曜日の抽選で 4 億 4,900 万ドルに増加	BUSINESS
あなたがアンビバートであることを示す 9 つのサイン	BUSINESS
1% と 99% の間の溝は拡大し続ける	BUSINESS
ウォール街は最初の自動運転テスラの死についてあまり心配していません	BUSINESS

Table 9: Example of translated News Headlines

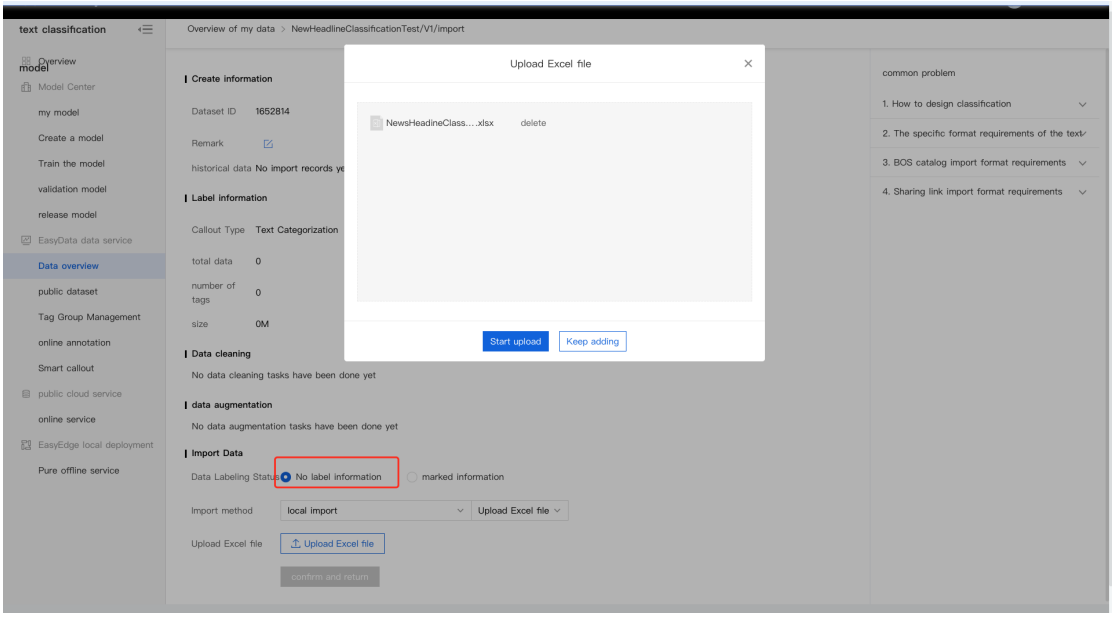

Figure 22: Create another test dataset using information without labels

4.3.3 Model Training

Once the data preparation and model creation is complete, we trained the model. Since we only use it for experimental testing, which takes up less resources, we directly train it on the public cloud deployment. We prepared 10,000 unlabeled

multilingual news headline data, as the test set, so that we can get more objective model effect evaluation results.

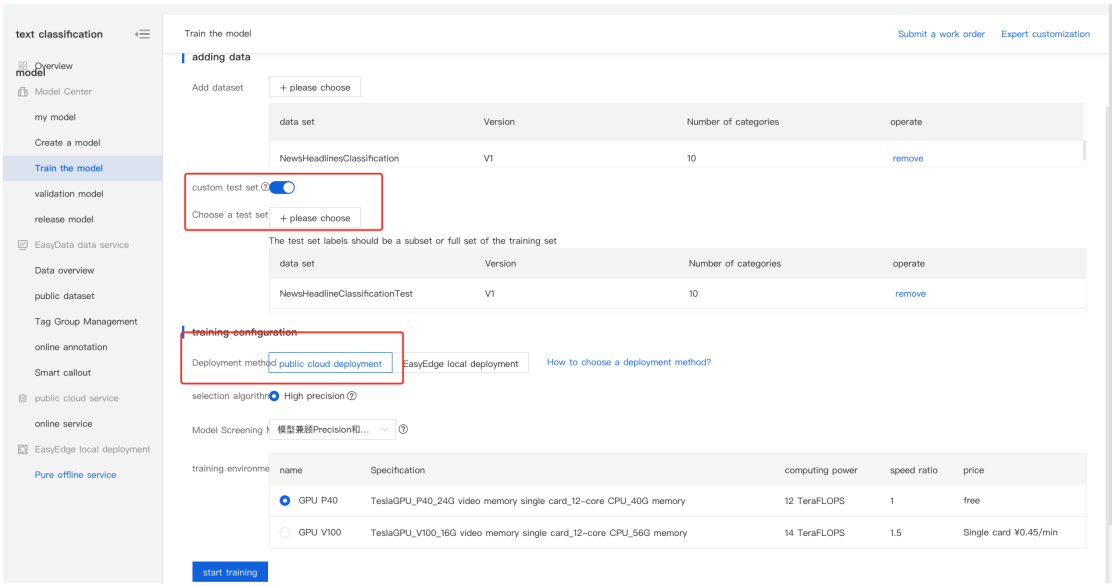

Figure 23: Fine-tuning the model with labeled dataset and test it with raw data

4.3.4 Effect Analysis

The same accuracy, precision, recall and F1-score are selected here as the metrics to measure the goodness of the model. When the model has been trained, the overall evaluation of the model can be reported by the following figure.

Figure 24: Overall assessment of NewsHeadlineClassification model

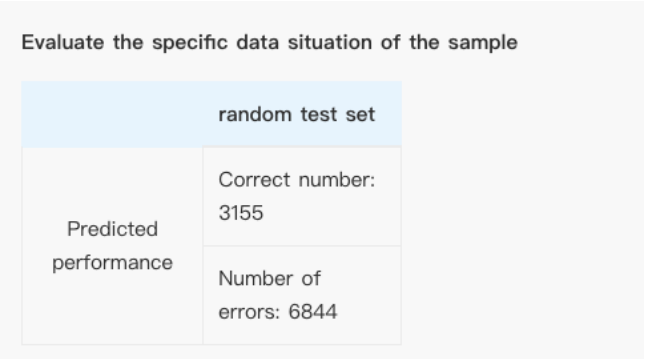

Figure 25: Predicted performance with specific data of NewsHeadlineClassification model

According to the results of this evaluation, the overall model effect is not very satisfactory. It can be broken down to see the specific values of each label.

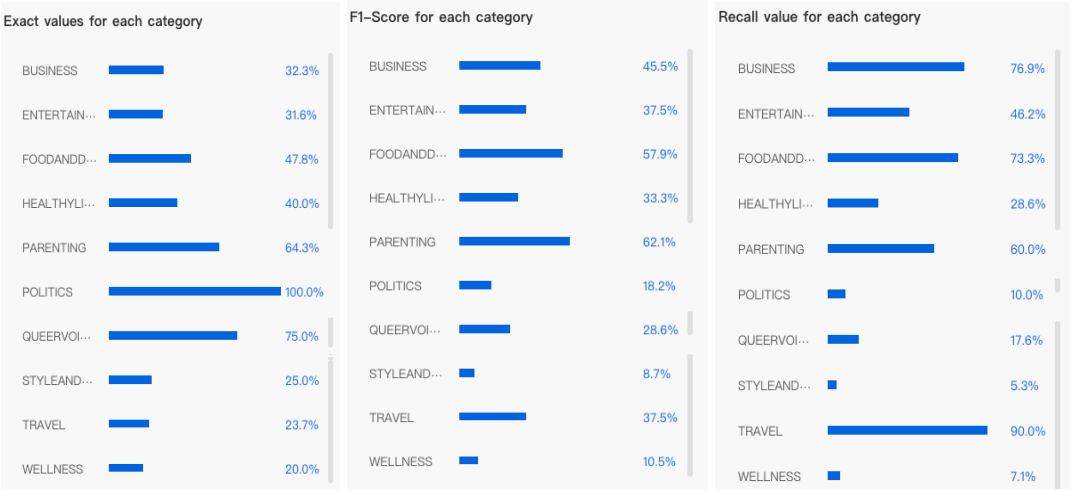

Figure 26: Accuracy, F1-Score and Recall for each category

It can be seen that POLICY has the highest accuracy, PARENTING has the highest F1-score, and TRAVEL has the highest recall. It implies that the model learns more precisely for these three types of data. However, there are still many classifications with low evaluation result values that need to be checked.

As shown in Figure 27, 'First-Class Passenger Tries To Open Emergency Door Mid-Flight, Shouts 'I Am God!'' with the category being TRAVEL was randomly selected in the model calibration, and after model prediction, the category of TRAVEL has the highest probability among all possible labels.

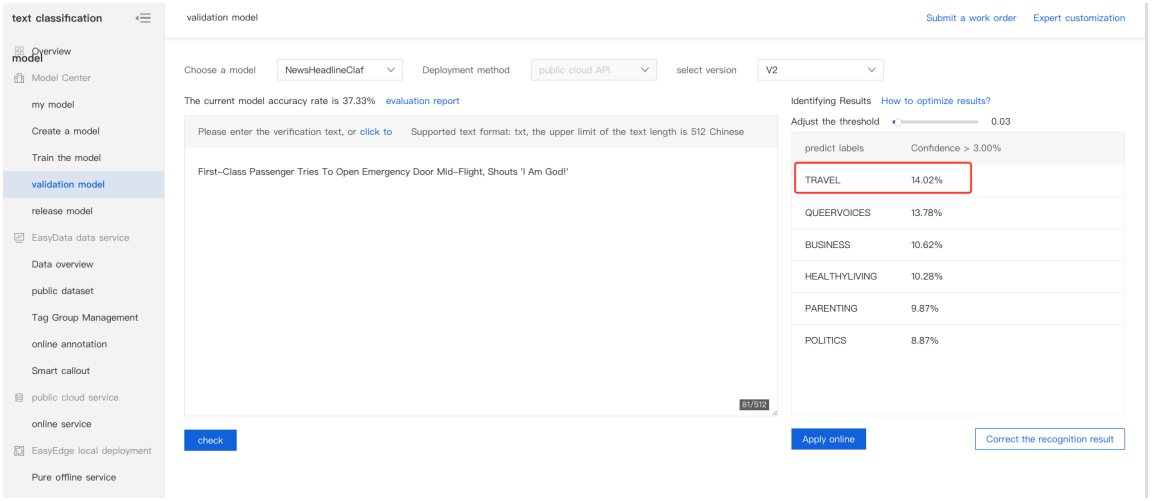

Figure 27: NewsHeadlineClassification model validation

The English prediction is accurate and it needs to be verified whether the model accuracy is degraded due to the test language. Therefore, we used Google Translate to translate this sentence into Portuguese, Japanese, and Swedish respectively for verification, i.e., 'Passageiro de primeira classe tenta abrir porta de emergência a meia luz, grita 'Eu sou Deus'! ', 'ファーストクラスの乗客が飛行中に非用ドアを开けようとし、「私は神だ！」と叫んだ。 and call んだ.' ' and 'Passagerare i första klass försöker öppna nöddörren mitt under flygningen och skriker "Jag är Gud! ', concluding as follows.

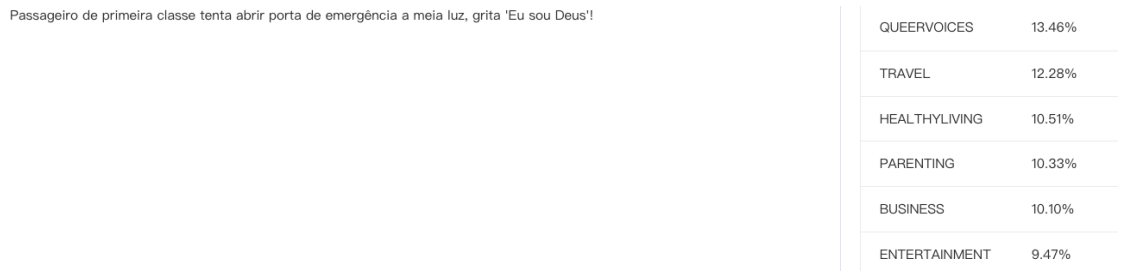

Figure 28: Predicted labels with exact confidence - Portuguese

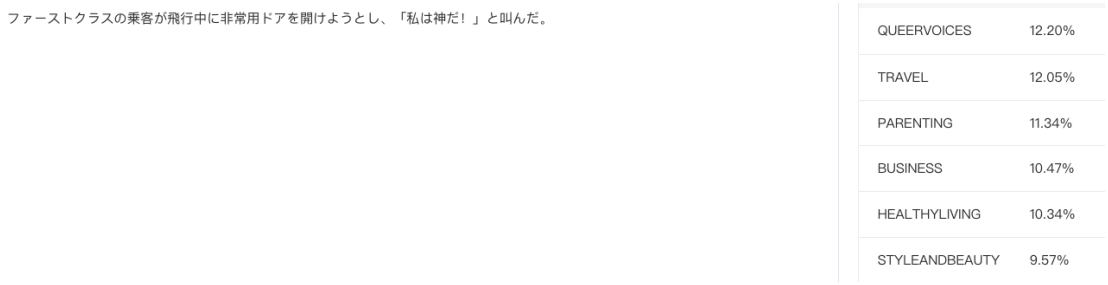

Figure 29: Predicted labels with exact confidence - Japanese

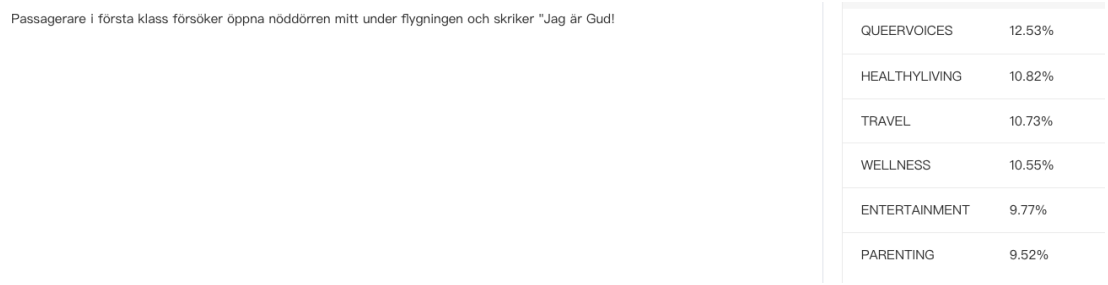

Figure 30: Predicted labels with exact confidence - Swedish

In all three languages tested, Portuguese, Japanese, and Swedish, the title 'First-Class Passenger Tries To Open Emergency Door Mid-Flight, Shouts 'I Am God!' was first identified as Queer Voices, while Travel appears in the second or third priority. The case demonstrates that for different languages, models trained with the same data produce different results for different languages. The most common language, English, works best, while other languages are able to be classified, but not as well.

The reason for this is that when the initial data was being prepared, all of the labeled models were in English, and the cross-language understanding was based solely on the ability of the pre-trained model, without any labeled data to fine-tune it.

Based on this, we trained and tested with only English data, and the results are as follows:

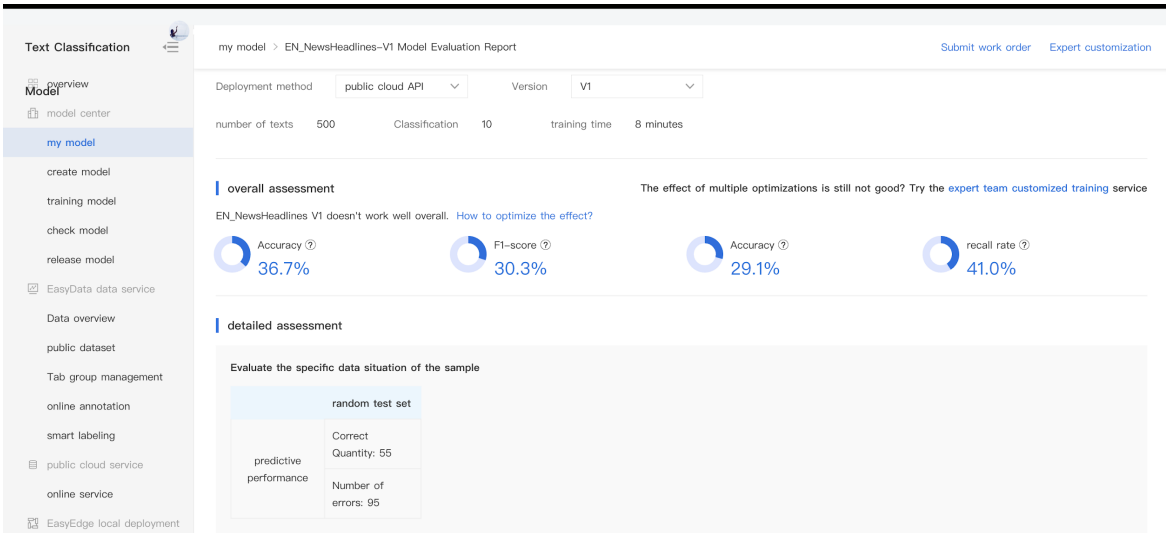

Figure 31: Assessment of the fine-tuning and testing the model with only one language

The tests conducted have shown that the language that is chosen for the pre-training can have an effect on the results of the experiments. During the process of fine-tuning the model, it is essential to not only ensure that a sufficient number of labeled individual tags were utilized for the training of the model, but also that the language was sufficiently rich. Because of this, it is essential to determine whether or not the quality of the model improves when the quantity of labeled data that is used for training is both sufficient data and multilingual.

4.4 Analysis of positive and negative reviews

4.4.1 Project Description

Sentiment analysis is one of the most classic applications in the field of NLP and has been flourishing, especially since the development of the Internet, it has greatly increased everyone's involvement in online shopping, using online products, etc. For instance, many users make this decision, which involves the aggregation and comparison of many apps, Airbnb, Booking, Skyscanner, and so on; when the software's features are comparable, we will check out the reviews left by previous customers and make our decision based on the total number of positive reviews. There is information about the preferences of users contained in the product reviews that are provided by users on top of third-party sites. Through the application of models for

sentiment analysis, it is possible to glean the feelings and preferences of users regarding product quality from reviews of product reviews. Analyzing the app's quality, popularity, and potential future commercialization area allows for the creation of an evaluation scale that can be used to improve both the app's functionality and its services. As a result of this, third-party platforms such as Google Play can make further use of the intelligent recommendation system to recommend products to users that they like more and help users select products that better meet their needs. This has the potential to increase the stickiness of users and tap some potential profits for the platform.

4.4.2 Data Preparation

Consistent with FAKE REVIEWS DETECTION, the data for this experiment uses 5902 multilingual reviews of Airbnb on Google Play from June 1 to June 30, 2022, exported from AppBot, which contains data from Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, and 34 other languages. Four candidate classification categories were delineated: Positive 4314, Neutral 284, Negative 1081, and Mixed 223.

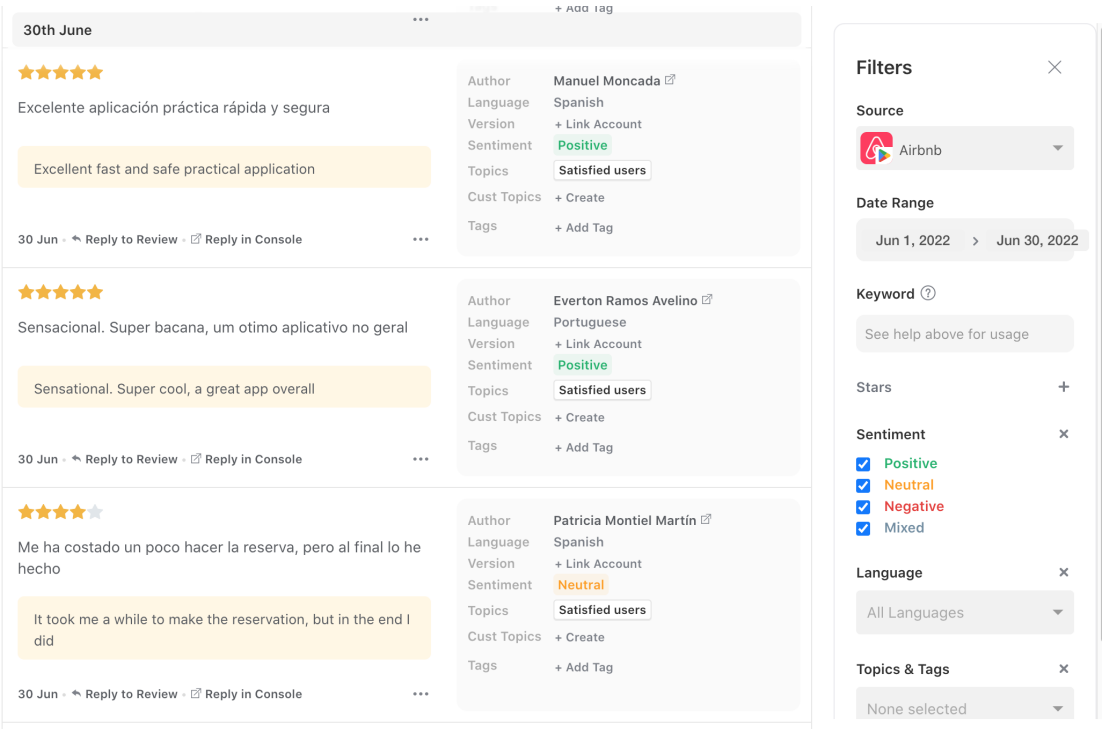

Figure 32: Reviews of Airbnb on Appbot

reviews	sentiments
Excelente opción para buscar alojamientos en tus vacaciones o salidas de fin de semana	Positive
사용 방법이.까다롭게 바껴 불편합니다. 신분증 앞뒷면까지 요구하고 사진도 요구하고	Neutral
Looking for an air bnb is harder than necessary! You have to go through every possible housing choice. I just want somewhere to stay. I don't care if it's a hut or yurt or cabin or whatever! The search needs to be like zillow where it just searches everything unless unmarked. Also, if I am searching in Texas and go from a lakefront option to the cabin option, it sends me half way around the world and I have to find Texas on the map again! I just don't feel its worth the hassle.	Negative
ME ENCANTA. Se adapta al.presupuesto, lugar, da recomendaciones. Al leer otras opiniones se puede solicitar el hospedaje con más claridad.Envian todo por escrito y en caso de viajar es más que útil porque lo piden en las migraciones. Tener un lugar de llegada es excelente y las evaluaciones que se publican son excelentes para tomar decisiones. Excelente app	Positive
No se que sucede con la aplicación pero me marca estancias disponibles de 8000 € 7000€ no tiene sentido antes de la actualización funcionaba bien. No la recomiendo.	Negative

Table 10: Example of multilingual reviews from Appbot

4.4.3 Model Training

The data is uploaded to the EasyDL AI platform in Excel format. Again, since this data is already classified by sentiments when exported from AppBot, the training can be launched directly after uploading to the platform. Figure 33 is an example of this process.

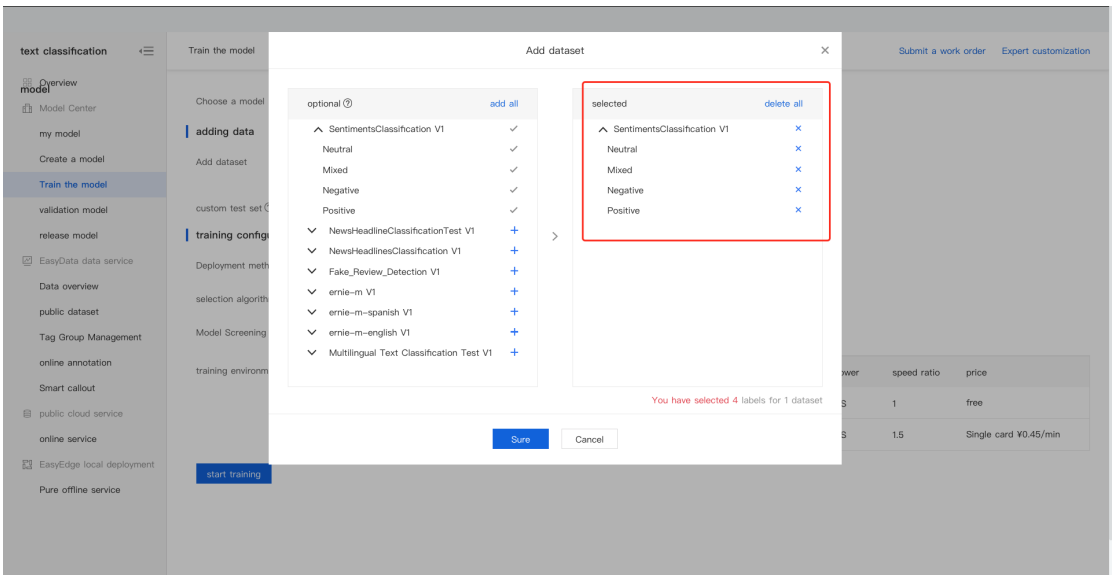

Figure 33: Training the model with labeled dataset for sentiments

4.4.4 Effect Analysis

As for the previous experiments, accuracy, precision, recall and F1-score are selected as the metrics to measure the goodness of the model.

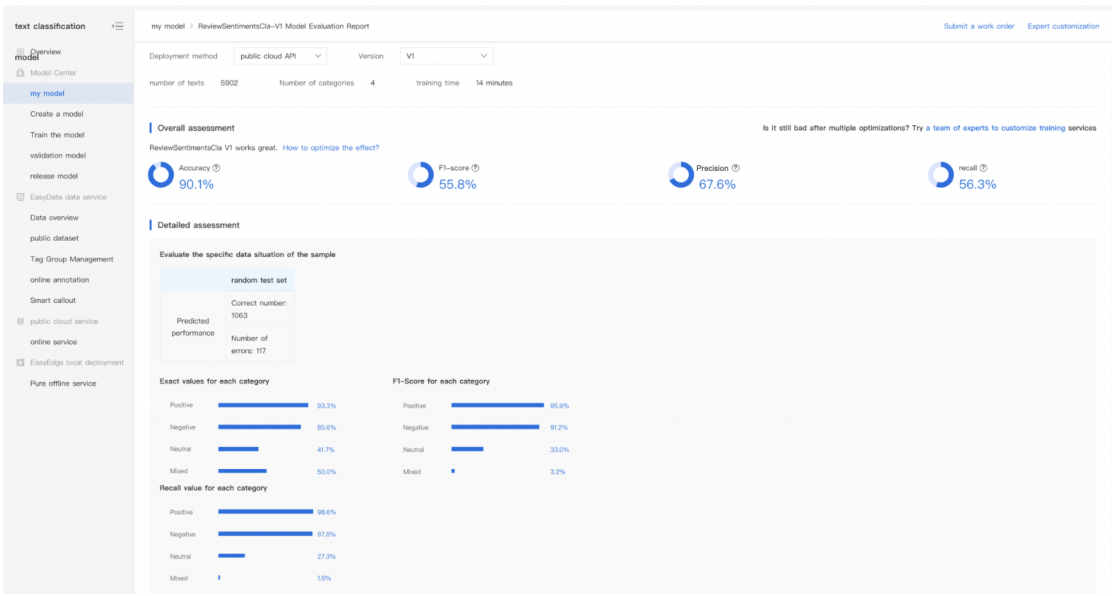

Figure 34: Assessment of ReviewSentimentsClassification model

According to the evaluation report, the accuracy rate of the model was 90.1%, and the other three indicators were also higher than the other two models. The

performance of the effect as a whole is of the highest values. For each tag to be subdivided, it is found that if we only look at Positive and Negative, the model's scores of each item are almost over 90%, and even the recall rate of both is above 97%. This indicates that the model is very accurate in judging the two labels with multilingual texts. However, the classification accuracy for Neutral and Mixed is not high. We analyzed the raw data in order to solve the issue of the model's poor classification accuracy for Neutral and Mixed.

1	reviews	languag	sentimer
37	Good app but issues with payment function: My wifi was too weak to allow authentification through my banking app and the payment flow doesn't let me try my credit card again with better signal. It automatically goes to reject although my card is perfectly fine. What's more， the error message I got is really weird ("Action required: Your payment failed. Pay ***€ by Jan 2 to keep your *** reservation. January 1， 1970." NB: we're in June and my booking is for August	English	Mixed
40	Buena app en general， pero en la busqueda de mi destino puse claro especificamente PLAYA DEL CARMEN， y me tiro a Cancun...🤦♂	Spanish	Neutral
52	Ho utilizzato Airbnb in più occasioni nel corso degli anni e nel complesso mi sono trovata piuttosto bene. Unico neo， non avere la certezza che l'alloggio corrisponda alle aspettative. Dalle foto può apparire graziosamente rifinito， molto pulito e spazioso， ma la realtà potrebbe non corrispondere del tutto. Lo stesso per quanto riguarda la posizione， che nella maggior parte dei casi non viene specificata sino a poco prima della partenza. In questi casi le recensioni aiutano molto nella scelta.	Italian	Neutral
58	I have loved Airbnb since they first started. I've never had a bad experience yet. They are always very responsive to messages if there is a problem with the rental or help in any way they can.	English	Neutral
63	Es mi primera vez ocupandolo， de regreso cuento que tal la atención y todo!	Spanish	Neutral
68	How can I make the camera to focus when taking pictures of my identity card? I just won't focus.	English	Mixed
73	Excelente segura. Nada más como comentan otras personas. En la aplicación aparece un importe y en el cargo final del banco otra.	Spanish	Mixed
74	Jusqu 'à présent je suis contente des services de airbnb trouvant des prix adaptés.	French	Neutral
98	La valoración por parte de ambas razones	Spanish	Mixed
119	사용 방법이.까다롭게 바껴 불편합니다. 신분증 앞뒷면까지 요구하고 사진도 요구하고	Korean	Mixed
123	It was my first time using airbnb... I went to el paso texas.. I rented a place through the airbnb app. It was very easy to use.. will definitely be using again.	English	Neutral
150	C'est pénible pour les prises de cliché. Sinon appli correct	French	Mixed

Table 11: Examples of data with 'Mixed' or 'Neutral' sentiment labels

It was found that the distinction between Mixed and Neutral was not clear enough in the original data, e.g., 'Ho utilizzato Airbnb in più occasioni nel corso degli anni e nel complesso mi sono trovata piuttosto Unico neo, non avere la certezza che l'alloggio corrisponda alle aspettative. dalle foto può apparire graziosamente rifinito, molto pulito e spazioso, ma la realtà potrebbe non corrispondere del tutto. Lo stesso per quanto riguarda la posizione, che nella maggior parte dei casi non viene In questi casi le recensioni aiutano molto nella scelta.' The original label of this data is

'Neutral', however it mentions 'nel complesso mi sono trovata piuttosto bene.' arguing that this app is good, while at the same time saying 'non avere la certezza che l'alloggio corrisponda alle aspettative', i.e. there is a discrepancy between the photos and the actual situation. This suggests that the evaluation should be Mixed rather than Neutral, and that there are many more such datasets with vague labels for training and inaccurate raw classification. It is speculated that this is the reason why the model cannot classify these two types of labels very accurately.

To verify the above speculation, we divided all the commented sentiments into three categories: Positive, Negative, and Neutral; i.e., the Neutral label now contains the original "Neutral" and "Mixed"; and then retrained the model with the modified labeled data.

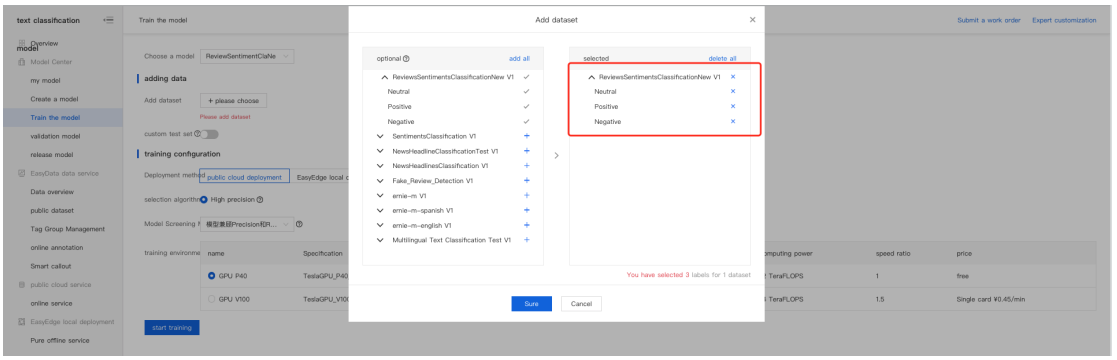

Figure 35: Fine-tuning the model after adjusting the dataset labels

The overall assessment results were 93.3% accurate, as shown in Figure 36.

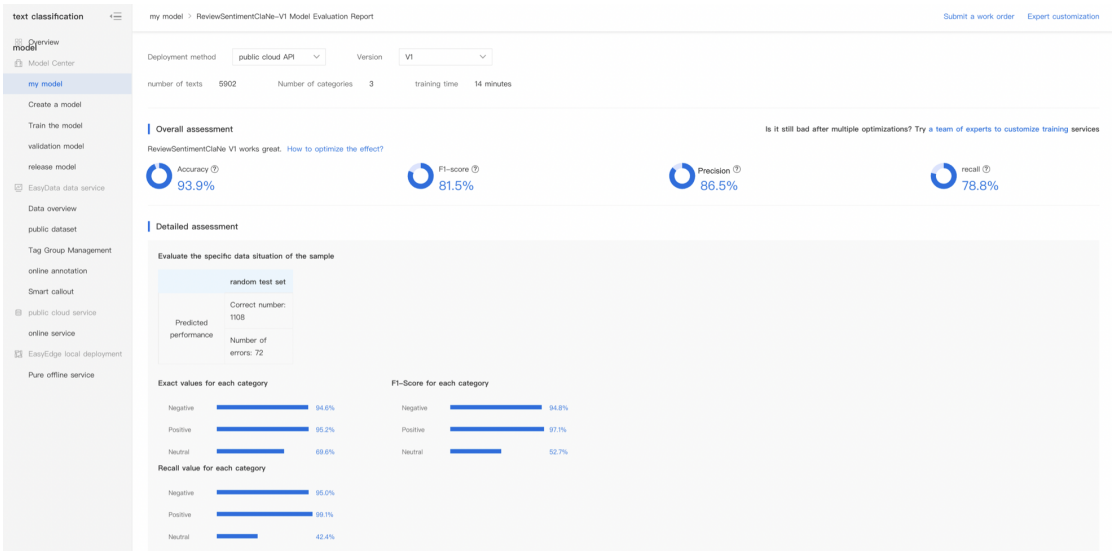

Figure 36: Assessment of the ReviewSentimentClassificationNew model

When the amount of labeled data used for fine-tuning is sufficient and the data is multilingual, a new set of test data demonstrates that the multilingual text classification ability of the model is significant after fine-tuning the pre-trained model ERNIE-M. This is the case when the model is being tested.

In this section, 20 different reviews written in 9 different languages, including French, Spanish, and Arabic, are taken and analyzed to determine the effectiveness of the model.

The results of the model classification are found to be very accurate for evaluations that have more explicit attitudes, according to the data from the overall evaluation. Taking the example of reviews with clearer emotional tendencies 'Site de réservation super Pratique et qd y an un souci quelconque, les personnes qui répondent au standard sont supers compé tentes et très sympa 👍 Je recommande ! The result of the model classification is positive, and there is confidence in the result up to 98.95%.

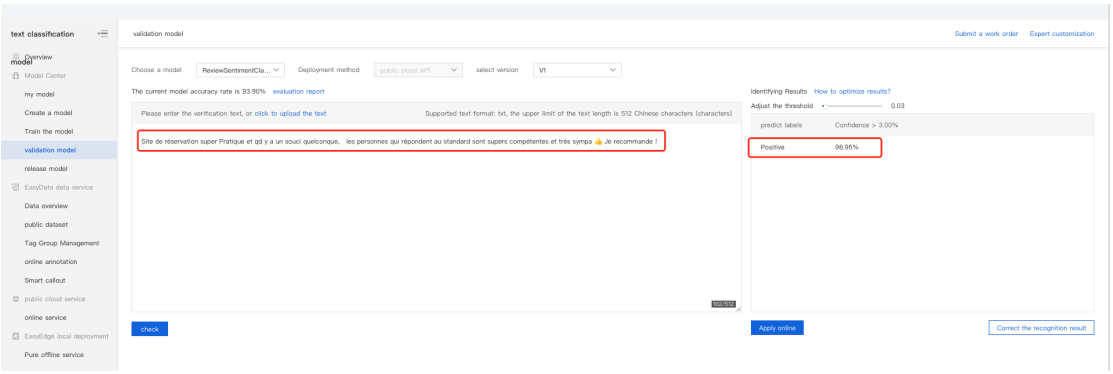

Figure 37: Validation of the ReviewSentimentClassificationNew model - Data with clear emotional tendency

The comprehensive evaluation of the model showed a slight lack of effect of the data for the Neutral category. Two Neutral categories were present in these 20 raw data used for model calibration.

'사용 방법이. 까다 날게임 바껴 불편합니다. 신분증 앞뒷면까지 요구하고 사진도 요구하고' (How to use. It is difficult to change and it is inconvenient. They ask for both the front and back of your They ask for both the front and back of your ID and a photo.) ERNIE-M classification results are 74.43% Negative, 25.16%

Neutral. according to 'It is difficult to change and it is inconvenient' can be analyzed. This is a more Negative evaluation.

'It was my first time using airbnb... I went to el paso texas. I rented a place through the airbnb app. it was very easy to use. It was very easy to use. will definitely be using again.' ERNIE-M classification results were 65.16% Positive, 34.60% Neutral. according to 'very easy to use' and It was very easy to use... will definitely be using again.

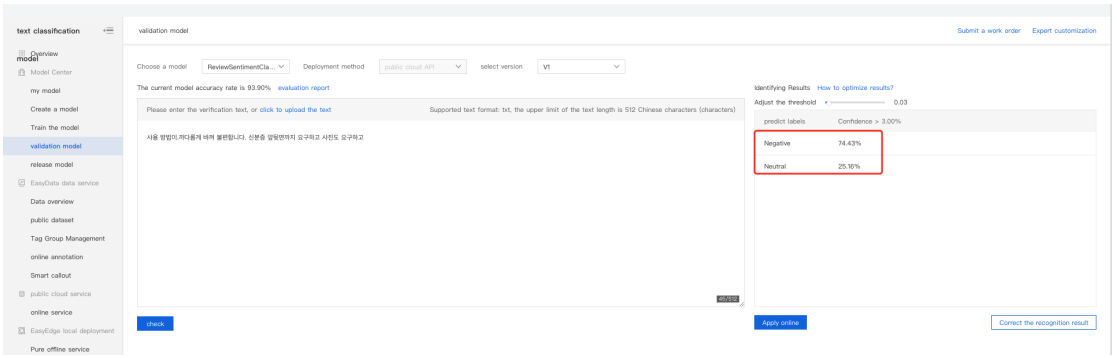

Figure 38: Validation of the ReviewSentimentClassificationNew model - Data without clear emotional tendency

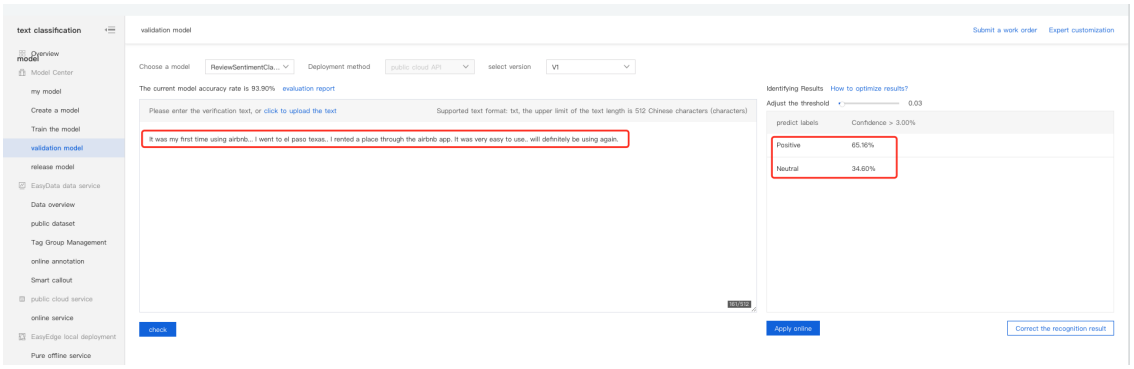

Figure 39: Validation of the ReviewSentimentClassificationNew model - Data with mixed emotional tendencies

Reviews	Language	Appbot Sentime	Ernie-m Classification	Manual classification
Site de réservation super Pratique et qd y a un souci quelconque, les personnes qui répondent au standard sont supers compétentes et très sympa 👍 Je recommande !	French	Positive	Positive	Positive
Fantástico si te gusta la aventura en el viaje y la convivencia...	Spanish	Positive	Positive	Positive
تجربه رائعه جداً	Arabic	Positive	Positive	Positive
Quick and not a expense travel now with airbnb services provider	Chinese (Traditional)	Positive	Positive	Positive
Me resultó muy útil en mi única experiencia en Colombia!...😊	Spanish	Positive	Positive	Positive
사용 방법이.까다롭게 바껴 불편합니다. 신분증 앞뒷면까지 요구하고 사진도 요구하고	Korean	Neutral	Negative 74.43% Neutral 25.16%	Negative
Very practical and easy to use Love the Google Maps link to get there	English	Positive	Positive	Positive
App super prático , dinâmico. Vc encontra acomodações muito boas por um preço justo	Portuguese	Positive	Positive	Positive
No lo uso.	Spanish	Negative	Negative	Negative
It was my first time using airbnb... I went to el paso texas.. I rented a place through the airbnb app. It was very easy to use.. will definitely be using again.	English	Neutral	Positive 65.16% Neutral 34.60%	Positive
Terrible aplicación me robó practicamente dinero , hice una reservación pero la anfitriona no respondió a la solicitud y en el momento que la app indico haría la devolución , no devolvió nada y me debitaron en mi cuánta , terrible aplicación.	Spanish	Negative	Negative	Negative
Melhores experiências da minha vida foi pelo airbnb	Portuguese	Positive	Positive	Positive
Excelente app , recomendable	Spanish	Positive	Positive	Positive
Very buggy! Couldn't log in to the app regardless of the method used	English	Negative	Negative	Negative
Muy buena	Spanish	Positive	Positive	Positive
L'app è diventata inutilizzabile , è complicatissimo scegliere le date e i filtri sono inutili (giusto per fare un esempio , a cosa serve il filtro per rilevatore di monossido se poi non si può nemmeno scegliere di cercare camere con idromassaggio o piscina privati invece di doverli condividere con un intero condominio). Si è trasformata in una vetrina per gli inserzionisti che pagano di più per mettere in evidenza il loro alloggio	Italian	Negative	Negative	Negative
오류가 잦다	Korean	Negative	Negative	Negative
ottima	Italian	Positive	Positive	Positive
Great air bnbs. Used aboard before. Our 1st stay in UK Stroud fantastic. Just booked another Airbnb Isle of Wright.	English	Positive	Positive	Positive
Kan ikke skifte til værtsskab	Danish	Negative	Negative	Negative

Figure 40: Comparison of the classification results of three different ways

The classification result of the word review ERNIE-M is even better than the original Appbot's Sentiment label.

4.5 Experiments' Conclusions

For the purpose of validating a multilingual text classification system, this experiment makes use of Ernie-M as a cross-language pre-training model. First of all, we conduct research on the use scenarios of multilingual text classification, and we find that it is primarily applied in three areas. These areas are cross-lingual text recognition, text topic extraction, and sentiment analysis, and we find that it is a powerful tool in all of these applications. In addition, we find that it is primarily applied in these areas because it is a multilingual text classification. On the basis of this information, we have chosen three different scenarios, namely "invalid comment recognition," "news headline classification," and "positive and negative comment analysis". In the first experiment, we used data that had not been labeled, and we trained the model on the Baidu AI platform called EasyDL, using just a small dataset of labeled data. However, the outcomes were not as satisfactory as originally anticipated, due mostly to the small dataset being used. In the second experiment, after reviewing all of the factors contributing to the issue, we increased the quantity of

labeled data used for training the model; however, the results of the model were still not satisfactory. The reason for this was due to the fact that the model could only be trained with data pertaining to English, so it was unable to correctly classify any other languages, it was quite specific to the language being used. In the third experiment of multi-lingual sentiments classification, the model was given fifty annotations written in multiple languages for each label. As a result, the learned model performed better than the classification results obtained from the primary source. As a result of fine-tuning the cross-language pre-training model ERNIE-M to correspond to multilingual text classification scenarios, we were able to demonstrate that this model can be applied to practical application scenarios with high quality after it has been fine-tuned to reflect multilingual text classification scenarios. This was done by showing that this model can be used in real-world situations.

Chapter 5

5 Conclusions and Future Work

This internship has greatly enhanced our understanding of Baidu as a company. We combined what has been learned along our student's path with practice, in order to have a better understanding of the field of deep learning, mostly focusing on pre-training and fine-tuning. Based on this, we also gained a more in-depth knowledge of machine translation. Most importantly, through research and experiments on cross-lingual text classification, we were able to verify the question of "whether the cross-lingual understanding model ERNIE-M using deep learning with machine translation is usable in the field of multilingual text classification" and the different effects of different fine-tuning and datasets properties on the model's effectiveness, proving that ERNIE-M has satisfactory confidence in the multilingual classification domain and can be used in commercial scenarios, especially sentiment classification.

Regarding our business perception of Baidu. Before this internship, we believed Baidu to be exclusively a search engine firm, as do many Chinese people also believe. Because foreign networks are not permitted in mainland China, 'Baidu' has been changing from a meaningless word to a verb meaning 'search,' similar to Google's position for foreigners. Nonetheless, this internship altered our perception of this organization. This is a technologically-driven Internet business. The user-product-driven Baidu of a decade ago has evolved into the technology-driven Baidu of the present day. Currently, Baidu is nurturing AI technologies in domains such as NLP and autonomous driving. We feel very privileged to work at Baidu and learn from our colleagues and fellow scientists.

As for the Deep Learning component, it refers mostly to the huge AI model previously developed through pre-training and fine-tuning, which we were able to test in three distinct tasks. Google's BERT model, launched in October 2018, is the most used generic model. It utilizes the text-only portions of BooksCorpus and the English Wikipedia, without labeled data, and is trained using two self-supervised tasks. The learned model is optimized for optimal performance on eleven downstream tasks. As

described in the literature, we believe that the AI large models are a technological milestone on artificial intelligence. Deep learning relies solely on models to autonomously acquire information from data. While greatly boosting performance, there is a conflict between the influx of general data and the dearth of specialized data. The characteristics of the AI large model are "large-scale" and "pre-training + fine-tuning." Prior to modeling for practical tasks, the generalizability, universality, and applicability of AI can be significantly enhanced by fine-tuning the large model with a specific scale of data. In the case of common parameters, the generic AI large model merely needs to make corresponding fine-tuning in different downstream experiments to gain improved performance, breaking through the restrictions of classic AI models that are difficult to generalize to other tasks.

Throughout the 70-year history of MT, from the original naive dictionaries to rule-based machine translation, statistical machine translation, and the current deep learn-driven neural machine translation, different technological paradigms have taken the lead in different historical periods. Current machine translation methods generally use neural networks to encode the source language to gain the semantic representation of the source language, and then decode the destination language according to the semantic representation. It appears to have departed from the traditional concept of the word "translation," which has nothing to do with the specific language, the grammatical subtleties, or the language's content. It does not require a large amount of labor to execute, and can translate hundreds of billions of words every day to serve global users. However, MT still faces numerous hurdles and limitations. First, there are thousands of languages spoken throughout the world. Even a large-scale model such as Ernie-M can only currently handle a few hundred languages, and a substantial number of languages have been "Left Behind" in the landscape of MT. Despite the fact that neural machine translation has considerably improved translation quality, it still confronts numerous obstacles and unresolved issues, including robustness, interpretability, wide-area context dependence, multimodality, and resource scarcity. In the realm of MT, much remains to be accomplished.

As for the categorization of multilingual text, this thesis analyzes the combination of MT with the concepts of multilingual text classification. Applications of text categorization are also studied. Thus,three tasks are selected for in-depth

verification: false comment identification, news title categorization, and sentiment analysis. The experimental results pointed out that ERNIE-M is capable of multilingual text classification; however, the classification effect is dependent on the training data utilized for fine-tuning. Different methods of data fine-tuning have distinct categorization effects on the model.

The many sorts of data are both linguistic and quantitative. When training data is limited to a single language, the trained model can nevertheless classify additional languages. When multilingual training data is used for fine-tuning, the cross-language understanding capacity of the model in the field can be improved, as can the cross-language classification effect. If there is less data for fine-tuning, the classification performance of the trained model is highly impacted by it. If there is an abundance of data for fine-tuning, the model can learn the properties of such tasks with greater precision and perform better on classification tests.

This work investigates whether the ERNIE-M cross-language understanding model combined with MT capacity may be used for multilingual text classification, as well as its practical impact. In future research, the following topics are worth additional exploration: The first is whether the performance upper limit of the multilingual text classification method based on the cross-language understanding model ERNIE-M is related to the translation quality of machine translation. The second is whether the pre-trained and fine-tuned AI large model can be used in the multilingual search field and how effective it is.

The research work in this paper verifies the ability of the AI development paradigm based on pre-training and fine-tuning on multilingual text classification, and uses actual data to prove that the classification ability of ERNIE-M model after appropriate fine-tuning is of very good performance, especially in sentiment classification. It is believed that the research of this paper can assist in exploring additional application scenarios based on cross-language understanding models, so that individuals who use every language can have equitable access to efficient Internet and artificial intelligence services.

References

Austin, P. K., & Sallabank, J. (Eds.). (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. ArXiv.org. https://arxiv.org/abs/1409.0473

Baidu NLP. (2010). *Baidu Natural Language Processing*. https://nlp.baidu.com/

Baidu Overview. (2020). Home.baidu.com. http://home.baidu.com/home/index/company

Bel, N., Koster, C. H. A., & Villegas, M. (2003). Cross-Lingual Text Categorization. *Research and Advanced Technology for Digital Libraries*, 126–139. https://doi.org/10.1007/978-3-540-45175-4_13

Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). *Amazon Review Classification and Sentiment Analysis*. Citeseerx.ist.psu.edu. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.736.4819&rep=rep1&type=pdf

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *ArXiv:1409.1259 [Cs, Stat]*. https://arxiv.org/abs/1409.1259

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. https://doi.org/10.1145/1390156.1390177

Company Overview | Baidu Inc. (2020). Baidu Inc. https://ir.baidu.com/company-overview/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. ArXiv.org. https://arxiv.org/abs/1810.04805

Dfdaily. (2011, May 11). 百度不是概念股. News.sina.com.cn. http://news.sina.com.cn/o/2011-05-11/084822443381.shtml

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Gliozzo, A., & Strapparava, C. (2005). *Cross language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora* (pp. 9–16). Association for Computational Linguistics. https://aclanthology.org/W05-0802.pdf

Hagen, A. (2020, October 19). *Microsoft Turing Universal Language Representation model, T-ULRv2, tops XTREME leaderboard*. Microsoft Research. https://www.microsoft.com/en-us/research/blog/microsoft-turing-universal-language-representation-model-t-ulrv2-tops-xtreme-leaderboard/

Hanneman, G., & Lavie, A. (2011). *Automatic Category Label Coarsening for Syntax-Based Machine Translation* (pp. 98–106). Association for Computational Linguistics. https://aclanthology.org/W11-1011.pdf

Huang, E., Socher, R., Manning, C., & Ng, A. (2012). *Improving Word Representations via Global Context and Multiple Word Prototypes* (pp. 873–882). Association for Computational Linguistics. https://aclanthology.org/P12-1092.pdf

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020, November 21). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation*. Proceedings.mlr.press; PMLR. http://proceedings.mlr.press/v119/hu20b.html

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). *Is neural machine translation ready for deployment? A case study on 30 translation directions.* arXiv.org. https://arxiv.org/abs/1610.01108

Jiedian. (2019, June 6). 百度*20*年:搜索帝国的崛起、式微与重生. http://www.woshipm.com/it/2430195.html

Johnson, J. (2021). *Search engine market share worldwide | Statista*. Statista; Statista. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. *ArXiv:1804.07745 [Cs]*. https://arxiv.org/abs/1804.07745

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Zens, R., Constantin, A., Moran, C., & Herbst, E. (2007). *Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding Center for Speech and Language Processing*. https://www.clsp.jhu.edu/wp-content/uploads/2015/10/WS06-OpenSource-FinalReport.pdf

Koster, C. H. A. (2003). *Spanish/English Cross-Lingual Categorization*. CiteSeer. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.187.4193

Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. *ArXiv:1901.07291 [Cs]*. https://arxiv.org/abs/1901.07291

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised Machine Translation Using Monolingual Corpora Only. *ArXiv:1711.00043 [Cs]*. https://arxiv.org/abs/1711.00043

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., & Schwenk, H. (2020). MLQA: Evaluating Cross-lingual Extractive Question Answering. *ArXiv:1910.07475 [Cs]*. https://arxiv.org/abs/1910.07475

Liu, J. (2018). 基于深度学习的多语种短文本分类方法的研究. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10184-1018117760.htm

Liu, X. (2016). 跨语言文本分类技术研究. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-90002-1018998033.htm

Liu, Z. (2010). 多语种多类别体系下文本自动分类系统的研究与实现. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10145-1013116851.htm

Lu, B., Tan, C., Cardie, C., & Tsou, B. (2011). *Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora* (pp. 320–330). Association for Computational Linguistics. https://aclanthology.org/P11-1033.pdf

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

Luong, M.-T., Pham, H., & Manning, C. (2015). *Bilingual Word Representations with Monolingual Quality in Mind* (pp. 151–159). Association for Computational Linguistics. https://aclanthology.org/W15-1521.pdf

Lu, W. (2011). 一种基于支持向量机的多国语言文本分类平台. Cpfd.cnki.com.cn. https://cpfd.cnki.com.cn/Article/CPFDTOTAL-TTTT201108001019.htm

Manaris, B. (1998, January 1). *Natural Language Processing: A Human-Computer Interaction Perspective* (M. V. Zelkowitz, Ed.). ScienceDirect; Elsevier. https://www.sciencedirect.com/science/article/pii/S0065245808606658

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. ArXiv.org. https://arxiv.org/abs/1301.3781

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *ArXiv:1309.4168 [Cs]*. https://arxiv.org/abs/1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. Neural Information Processing Systems; Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Nagao, M. (1996). Some rationales and methodologies for example-based machine translation. In *Discourse and Meaning* (p. 291). John Benjamins.

Ouyang. (2021). *ERNIE/ERNIE-M*. GitHub. https://github.com/PaddlePaddle/ERNIE/tree/repro/ERNIE-M

Ouyang, W. (2018). 文档表示与双语词嵌入算法研究. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10358-1018088439.htm

Ouyang, X., Wang, S., Pang, C., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora. *ArXiv:2012.15674 [Cs]*. https://arxiv.org/abs/2012.15674

Peng, Z. (2014). 跨语言文本相关性检测技术研究. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10533-1014408263.htm

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *ArXiv:1906.01502 [Cs]*. https://arxiv.org/abs/1906.01502

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-Based Multilingual Retrieval Model. *Lecture Notes in Computer Science*, 522–530. https://doi.org/10.1007/978-3-540-78646-7_51

Prettenhofer, P., & Stein, B. (2011). Cross-Lingual Adaptation Using Structural Correspondence Learning. *ACM Transactions on Intelligent Systems and Technology*, *3*(1), 1–22. https://doi.org/10.1145/2036264.2036277

Rigutini, L., Maggini, M., & Liu, B. (2005, September 1). *An EM based training algorithm for cross-language text categorization*. IEEE Xplore. https://doi.org/10.1109/WI.2005.29

Saigal, P., & Khanna, V. (2020). Multi-category news classification using Support Vector Machine based classifiers. *SN Applied Sciences*, *2*(3). https://doi.org/10.1007/s42452-020-2266-6

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *ArXiv:cs/0306050*. https://arxiv.org/abs/cs/0306050

Shi, B., Bai, X., & Yao, C. (2017). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2298–2304. https://doi.org/10.1109/tpami.2016.2646371

Shi, J., Zhou, L., & Xian, Y. (2016). 基于*WordNet*的中泰文跨语言文本相似度计算. http://jcip.cipsc.org.cn/CN/article/downloadArticleFile.do?attachType=PDF&id=2248

Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ArXiv.org. https://arxiv.org/abs/1409.1556

Slocum, J. (1985). A survey of machine translation: Its history, current status and future prospects. *Computational linguistics*, *11*(1), 1-17.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). *ERNIE: Enhanced Representation through Knowledge Integration*. ArXiv.org. https://arxiv.org/abs/1904.09223

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. *Google Scholar*, 1-9.

Vulić, I., & Moens, M.-F. (2015). Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/2766462.2767752

Wang, D., Yang, X., Ma, J., & Zhang, L. (2015). Retrieval Methods of Natural Language Based on Automatic Indexing. *Computer and Computing Technologies in Agriculture IX*, 346–356. https://doi.org/10.1007/978-3-319-48354-2_35

Wang, S. (2013). 中英可比较语料库的构建. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10141-1013200520.htm

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

Zhang, Y., & Wallace, B. (2015). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. ArXiv.org. https://arxiv.org/abs/1510.03820

Zhu, J. (2016). 基于贝叶斯算法的多语言文档分类. Cdmd.cnki.com.cn. https://cdmd.cnki.com.cn/Article/CDMD-10285-1017012634.htm

Appendices

Appendix 1

List of Acronyms

Accuracy	A
Area Under Curve	AUC
Association for Computational Linguistics	ACL
Automatic Language Processing Advisory Committee	Alpac
Back-translation Masked Language Modeling	BTMLM
Bilingual Skip-Gram	BiSkip
Bullletin Board System	BBS
Byte Pair Encoding	BPE
Causal Language Modeling	CLM
Cross-attention Masked Language Modeling	CAMLM
Cross-lingual Natural Language Inference	XNLI
Example-based Machine Translation	EBMT
Facebook AI Research	FAIR
False Positive Rate	FPR
Machine Translation	MT
Masked Language Modeling	MLM
MultiLingual Question Answering	MLQA
Multilingual Text Classification	MTC
Natural Language Process	NLP
Neural Machine Translation	NMT
Neural Network Translation	NMT
Precision	P
Recall	R
Recurrent Neural Networks	RNNs
Skip-Gram with Negative Sampling	SGNS

Statistical Machine Translation	SMT
Bidirectional Encoder Representations from Transformers	BERT
The Cross-lingual TRansfer Evaluation of Multilingual Encoders	XTREME
The Multilingual Unsupervised and Supervised Embedding method	MUSE
Translation Language Modeling	TLM
True Positive Rate	TPR
Unsupervised Machine Translation	UNMT

Appendix 2

reviews	language	sentiments
Buen servicio y atención	Spanish	Positive
No me dejó abrir la galería para escoger una foto y tampoco la cámara para sacar una par el perfil	Spanish	Negative
I find your site irrating. the price is always higher than the initial amount	English	Negative
No se que sucede con la aplicación pero me marca estancias disponibles de 8000 € 7000€ no tiene sentido antes de la actualización funcionaba bien. No la recomiendo.	Spanish	Negative
Muito bom.	Portuguese	Positive
Un lujo todo	Spanish	Positive
Todo bien pero detesto que me aparezcan los lugares valorados en pesos mexicanos , estoy en El Salvador Usamos dólares , no se como cambiar la opción	Spanish	Negative
Très bien en tous points	French	Positive
Siempre consigo alojamientos convenientes	Spanish	Positive
Awesome	English	Positive
The best	Greek	Positive
Muy bueno. Gracias	Spanish	Positive
Easy to use , wonderful places and people.	English	Positive
Alles einfach unkompliziert	German	Positive
Excelente app para encontrar alojamiento , muy intuitiva.	Spanish	Positive
very easy to use	Romanian	Positive
Muy buena	Spanish	Positive
Amazing!	English	Positive
Looking for an air bnb is harder than necessary! You have to go through every possible housing choice. I just want somewhere to stay. I don't care if it's a hut or yurt or cabin or whatever! The search needs to be like zillow where it just searches everything unless unmarked. Also , if I am searching in Texas and go from a lakefront option to the cabin option , it sends me half way around the world and I have to find Texas on the map again! I just don't feel its worth the hassle.	English	Negative
Excelente herramienta para planear mis vacaciones	Spanish	Positive
Excelente opción para buscar alojamientos en tus vacaciones o salidas de fin de semana	Spanish	Positive
Precisi per ogni prenotazione fatta	Italian	Positive
Réservation facile à établir. Dite clair et annonces interressantes.	French	Positive
Fantasic	English	Positive
Excelente 😀	Spanish	Positive
Excelente	Portuguese	Positive
Sito serio e attendibile!	Italian	Positive
Todo excelente sin ningún contratiempo	Spanish	Positive
Muy fácil de usar.	Spanish	Positive
muy buena	Spanish	Positive
Muy buena aplicación , la uso constantemente en mis viajes , super segura y responsables en todo momento	Spanish	Positive
ME ENCANTA. Se adapta al.presupuesto , lugar , da recomendaciones. Al leer otras opiniones se puede solicitar el hospedaje con más claridad.Envian todo por escrito y en caso de viajar es más que útil porque lo piden en las migraciones. Tener un lugar de llegada es excelente y las evaluaciones que se publican son excelentes para tomar decisiones. Excelente app	Spanish	Positive

Figure 41: Airbnb reviews examples

Appendix 3

	News Headlines	Category
2	U.S. Launches Auto Import Probe, China Vows To Defend Its Interests	BUSINESS
3	Starbucks Says Anyone Can Now Sit In Its Cafes -- Even Without Buying Anything	BUSINESS
4	Seattle Passes Controversial New Tax On City's Biggest Companies To Combat Housing Crisis	BUSINESS
5	Uber Ends Forced Arbitration In Individual Cases Of Sexual Assault, Harassment	BUSINESS
6	Chili's Hit By Data Breach, Credit And Debit Card Information Compromised	BUSINESS
7	How Uber Silences Women After Sexual Assaults	BUSINESS
8	How Amazon Is Holding Seattle Hostage	BUSINESS
9	Bank Of America Appears To Flip On Firearm Promise With Loan To Remington	BUSINESS
10	Ex-Volkswagen CEO Charged In U.S. Over Emissions Cheating Scandal	BUSINESS
11	Women Describe Rampant Groping, Sexual Harassment At Verizon-Contracted Warehouse	BUSINESS
12	T-Mobile Agrees To Acquire Sprint For $26 Billion	BUSINESS
13	The SEC Just Made The Case For Divesting From Fossil Fuel Companies Much Stronger	BUSINESS
14	Dick's Sporting Goods Is Destroying Its Unsold Assault-Style Rifles	BUSINESS
15	Facebook Didn\u2019t Seem To Care I Was Being Sexually Harassed Until I Decided To Write About It	BUSINESS
16	In Los Angeles, Bitter Tensions Over Where To House The Homeless Rile Communities	BUSINESS
17	The Libertarian Political Movement Is Dead	BUSINESS
18	Apple Co-Founder Steve Wozniak Ditches Facebook After Data Scandal	BUSINESS
19	Twitter Has Suspended 1.2 Million Accounts For 'Terrorist Content'	BUSINESS
20	Trump Proposes Slapping $100 Billion In New Tariffs On Chinese Goods	BUSINESS
21	The Facebook Apology Tour Continues	BUSINESS
22	Saks, Lord & Taylor Hit By Payment Card Data Breach	BUSINESS
23	International Health Group Drops Partnership With Heineken Over 'Beer Girls'	BUSINESS
24	MyFitnessPal Security Breach Affects 150 Million Users, Under Armour Reports	BUSINESS
25	These Stock Photos Show Masculinity Is More Than Biceps And Beer	BUSINESS
26	Trapped Inside The Monster Energy Frat House	BUSINESS
27	There Are Psychological Reasons Parents Are So Obsessed With Target	BUSINESS
28	Walmart Partners With Conservative Group To Remove Cosmo From Checkout Lines	BUSINESS
29	Uber Halts Self-Driving Car Tests In California, Where It Didn't Test Much Anyway	BUSINESS
30	Remington Files For Chapter 11 Bankruptcy Amid Mounting Pressure For More Gun Control	BUSINESS

Figure 42: Dataset of orginal news headlines

Appendix 4

	A	B
1	Marc Jacobs usa o Twitter para lançar seu novo anúncio, a Internet explode com selfies	STYLEANDBEAUTY
2	Bar Refaeli recebe tratamento facial de ouro líquido que pode custar tanto quanto seu aluguel	STYLEANDBEAUTY
3	9 belezas que você não deve fazer	STYLEANDBEAUTY
4	Joias Boho Chic conectam selvas amazônicas e urbanas	STYLEANDBEAUTY
5	Discriminação de tamanho existe em suas lojas favoritas - você pode não perceber	STYLEANDBEAUTY
6	Kate Upton parece familiar na estreia de 'The Other Woman'	STYLEANDBEAUTY
7	Obtenha uma trança pronta para o festival de música em menos de 2 minutos	STYLEANDBEAUTY
8	El intento de Trump de hacer un ejemplo de la 'caravana' de inmigrantes está fracasando	POLITICS
9	Los tuiteros estallan por la burla cruel del asistente de Trump sobre John McCain 'morir de todos	POLITICS
10	La EPA ocultó la cena de Scott Pruitt con el negador del clima acusado de abuso sexual infantil	POLITICS
11	Meghan McCain da un golpe en la Casa Blanca después de que un asistente se burló del cáncer c	POLITICS
12	Joe Biden reprende a la Casa Blanca por 'broma' sobre la salud de John McCain	POLITICS
13	Candidato al Congreso dice 'F**k The NRA' en nuevo anuncio de televisión	POLITICS
14	Katie Porter sobrevivió al abuso doméstico, solo para que lo usaran en su contra en su campaña	POLITICS
15	Facebook för att bygga lägenheter i Silicon Valley	BUSINESS
16	Verizon planerar att förvärva Yahoo i $4,83 miljarder affär	BUSINESS
17	Denna vd är redo att slå tillbaka om Donald Trump vinner	BUSINESS
18	Marissa Mayer tjänade mycket pengar på att förlora kampen för att rädda Yahoo	BUSINESS
19	Vi får från världen det vi investerar i själva	BUSINESS
20	Hinder för Creative Disruption	BUSINESS
21	Hur man ändrar dålig möteskultur	BUSINESS
22	ゲイリー・オールドマンがオスカーで「ダーケスト・アワー」で主演男優賞を受賞	ENTERTAINMENT
23	約束どおり、ティファニー・ハディッシュはオスカーで彼女の白い「SNL」ガウンを再着	ENTERTAINMENT
24	「メリー・ポピンズ リターンズ」の初見は絶対に魔法です	ENTERTAINMENT
25	クロエ・カーダシアンとトリスタン・トンプソンに女の子が誕生	ENTERTAINMENT
26	ハリウッドは多様性についてのオスカー モンタージュで背中を撫でる	ENTERTAINMENT
27	ニューヨーク・タイムズがケビン・スペイシーに悪意のあるオスカーの陰を投げる	ENTERTAINMENT
28	キャシー・グリフィンは、ライアン・シークレストの告発者に仕事を提供し、オスカー(	ENTERTAINMENT

Figure 43: Dataset of translated news headlines