Project Plan for Final Year Project
Deep Learning for Text Classification in Azure Infrastructure

Kai YAN  
UID: 3035141231

Zhihan CHEN  
UID: 3035142261

Zixu WANG  
UID: 3035140067

I. INTRODUCTION

Increasing computing power and exploding volumes of data have driven us into the era of machine learning in the past few decades. With the aid of deep neural networks, researchers have equipped computers with eyes that can identify different objects [1], brains that can diagnose diseases [2], and hands that can create astonishing works of art [3]. Besides these remarkable achievements, natural language processing (NLP for short) is also a promising field for deep learning to explore its potential. Traditionally, researchers need to summarize features of natural languages themselves based on their linguistic knowledge and then teach computers how to understand human languages [4]. Nowadays with emerging technologies of deep learning, much research has been done to let computers learn by themselves to recognize human speeches [5], translate between languages [6], and even speak in human languages [7]. Computers’ better understanding of natural languages has also enabled numerous useful real-world applications such as Siri and Google Translation.

However powerful NLP is with the help of deep learning, the majority of current studies focus on English processing. There are few research studies and applications of NLP with Chinese. In addition to the fact that English is the major language used in academia, Chinese language processing faces several other challenges such as word segmentation [8]. With great potential for real-world applications as well as research directions, it is desirable to adapt existing deep NLP models originally designed for English to Chinese processing and evaluate their performance, and also to design and develop new models. Having this motivation, we will cooperate with Microsoft to develop a system which leverages the approach of deep learning for Chinese sentiment analysis in Azure infrastructures. The significance of this project is three-fold. First, we will explore different options and possibilities for Chinese NLP and identify their advantages and drawbacks. Second, sentiment analysis for domain-specific Chinese corpora may yield significant value in that domain. For example, analyzing sentiments in customers’ feedbacks may guide companies to improve their products or services. Last, we will get our hands on state-of-the-art knowledge and technologies about deep learning and NLP, and acquire valuable experiences of developing deep learning NLP systems.

The remaining of this project plan proceeds as follows. First, we offer background information about existing studies on Chinese NLP and also deep learning models for NLP with English (§II). We then introduce the objective and scope of our project (§III). Next, we elaborate the methodology guiding our project development (§IV), followed by project management details (§V). Then to properly handle future challenges, we give our predictions on potential risks and corresponding mitigations (§VI). We close this project plan with a tentative schedule for this project (§VII).
II. BACKGROUND

In this section, we first introduce existing research highlights on NLP with deep learning to justify the feasibility and potential of our project. The background of our project is then elaborated.

A. Existing Works

Endowing computers with the ability to comprehend human languages and communicate with us in our mother tongue has long been an elusive pursuit in the field of artificial intelligence. Over the years, researchers have built systems that focused on specific NLP tasks by extracting language features manually, leveraging the researchers’ profound linguistic knowledge. Though results of these works are remarkable and useful in practice, this ad hoc manner of engineering fails to envision a more general goal of understanding natural languages [4].

Aiming for a unified algorithm to automatically learn features from text pieces, various deep learning models have emerged. Simple feed-forward neural networks language models (NNLMs) already show excellent performance in word embeddings, which are representations of words in a linear space that captures meanings of or relations between words [9]. For example, word2vec [10]–[12], a famous word embedding model which learns meanings of words as vectors, is often presented as an introductory example of state-of-the-art NLP techniques for its incredible results and its simple architecture. The idea behind is to learn a word’s meaning from its surrounding context [10]–[12]. By taking the context of words into consideration, word2vec scales better when dealing with large corpora, compared with traditional N-gram models [10]. The performance of word2vec is surprisingly good. It even learns relations between words in an unsupervised manner. For example, information about analogies can be extracted by subtracting word vectors [10], [11]:

\[ V_{\text{France}} - V_{\text{Paris}} \approx V_{\text{Italy}} - V_{\text{Rome}} \]
\[ V_{\text{Einstein}} - V_{\text{scientist}} \approx V_{\text{Picasso}} - V_{\text{painter}} \]

Following the trending direction, NLP researchers have adopted more complicated and sophisticated models. Two most commonly used deep learning models in NLP are Recursive Neural Networks (RNNs) [9], [13], [14] and Convolutional Neural Networks (CNNs) [15]–[17]. For example, researchers from Stanford introduced Recursive Neural Tensor Network (RNTN) to classify English sentences into five sentiment classes [13]. Figure 1 [13, Fig. 7] shows one correct prediction of a complex sentence consisting a contrastive conjunction, which clearly demonstrates such model’s capability of handling complicated large-scale semantics.

All these works show deep learning’s potential of performing NLP tasks such as sentiment analysis, which offers a firm theoretical and practical basis for our project. However, there exists a gap between our project and existing works which cannot be neglected: differences between Chinese and English. For instance, Chinese is a character-based language while English is word-based. For those models originally designed for English to be able to process Chinese, one has to include a word segmentation phase beforehand, which turns out to be another difficult problem [8]. Nevertheless, some works show evidence that this gap is not unbridgeable. A research group at Fudan University successfully applied deep neural networks to tasks including Chinese word segmentation and part-of-speech tagging [8]. Furthermore, IBM Brazilian Research Lab recently proposed Character to Sentence Convolutional Neural Network (CharSCNN), which is able to extract English language features on a per-character basis [15]. All these works cast light on the application of deep learning in Chinese NLP tasks.

B. Background of Our Project

Our project is an industry-based project proposed by Microsoft. We will focus on performing the Chinese sentiment classification task with the help of deep learning. Microsoft will provide us with datasets for training and testing. However, by the
time this plan is written, Microsoft has not delivered the datasets and therefore the detailed domain for our project is not yet determined.

As required by Microsoft, we will develop our project using Microsoft Cognitive Toolkit (CNTK)\(^2\). Detailed discussion of development platforms is out of the scope of this project plan and thus omitted.

III. OBJECTIVE & SCOPE

In this section, we introduce the objective and scope of our project. As mentioned before, we have not learned a specific domain from Microsoft. Therefore, this section tends to describe our objective and scope generally and is subject to further adjustments.

A. Objective

The general purpose of this project is to conduct experiments on Chinese NLP and to develop a system integrated with deep neural network models to predict underlying sentiments of Chinese sentences from one corpus or several corpora. Later with a specific domain, this project may also explore potentials of applying such models in related industries.

B. Scope

In this subsection, we define the scope of our project in terms of problems to solve and tools to use.

1) Problems to Solve: The main problem to tackle in this project is sentiment prediction of Chinese text.

The term “Chinese text” means short text pieces (generally one or two sentences under 30–40 words) in simplified or traditional Chinese, encoded in standard formats. Later with a specific domain, the Chinese text should come from the same or similar scope. The term “sentiment prediction” means giving labels indicating sentimental degrees (positive, negative, neutral, etc.) without knowing or referring to the ground truth. The labels should be limited to sentimental degrees instead of types of feelings (humor, upset, bored, etc.).

2) Tools to Use: This project will use Python on CNTK as the development platform. This project will take the approach of deep learning (deep structured learning, deep neural networks, etc.) to achieving its objective. Microsoft Azure will provide the computing power required by this approach.

IV. METHODOLOGY

In this section, we introduce the methodology we are going to adopt for our project. It describes how we are going to design our model and implement it. As this is a machine learning project, we will go with the iterative process and fine-tune our implementation swiftly, that is, we will adjust our neural network or other parts of the pipeline immediately after problems like high-variance or high-bias are identified.

A. Model design

We will first research on modern neural networks designed for NLP and benchmark several suitable models to pick the optimal one for further improvement. Accuracy will be our optimizing metric and time satisficing. Migration from English to Chinese must be performed before benchmarking in order to keep consistency between the target goal and the optimizing goal. If possible, the development set should be from the domain of focus at this stage as some features can be domain-specific. As few neural networks can handle Chinese input, the fundamental principles of sentimental analysis that can be carried from one language to another must be taken into consideration during the process of model migration. Depending on the original model our project is going to be based on, we might add some essential steps into the pipeline. For example, text segmentation is to be applied when word2vec is adopted. While for CNNs, as mentioned in the background section (§II), no prior segmentation is required as we might be able to analyze on a per-character basis.

After benchmarking, we should be able to have a comprehensive understanding of pros and cons of different models in handling Chinese datasets. Subject to the available time, we may choose several promising models for fine-tuning. Further improvement can be made to the migrated models allowing them to fit better to Chinese text. Finally we may pick one of the best models for further hyperparameter tuning. This process can be time consuming so we may consider to adopt strategies like early-stopping.

B. Software Engineering

Designing our models and the relevant analysis should be the main focus of this project. Nonetheless, we will use CNTK to implement it. It provides

\(^2\)https://www.microsoft.com/en-us/cognitive-toolkit/
high-level interfaces while we will still go and review the documentation to get ourselves familiar with it. Once the product is completed, we can deploy it on a server, design an API for sentiment prediction and build up a user interface.

Table I

<table>
<thead>
<tr>
<th>ID</th>
<th>Risk</th>
<th>Mitigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Cannot get enough data from Microsoft</td>
<td>Search online for open-source Chinese datasets as a backup plan.</td>
</tr>
<tr>
<td>R2</td>
<td>No good reference for Chinese text processing</td>
<td>Search for English-based papers or models and make appropriate adjustments.</td>
</tr>
<tr>
<td>R3</td>
<td>Lack of computing power to test out different models before getting to Azure</td>
<td>Adjust the size of the dataset to benchmark efficiently.</td>
</tr>
<tr>
<td>R4</td>
<td>Unstable API functionality</td>
<td>Review the online posts from Microsoft. Use necessary abstractions to avoid massive modifications in the future.</td>
</tr>
</tbody>
</table>

V. Project Management

As we will collaborate with Microsoft Hong Kong, the project will be co-managed by our team and Mr. Delon Yau, who is our point of contact from Microsoft. We may adjust the scope and objective of this project according to our client’s need when necessary. Our internal supervisor, Dr. Anthony Tam, will monitor the progress of this project and may give suggestions accordingly. Weekly meetings between the team and the internal supervisor will be held. Our three team members, Kai YAN, Zhihan CHEN and Zixu WANG, will collaborate throughout the whole progress. To facilitate our teamwork, each member may choose his role, and we made our division of work roughly based on the requirement from our client, the expected workload and the strength of each member. Specifically:

- Kai YAN is in charge of data collection, text segmentation, and other necessary preprocessing. Implementing the user interface can also be part of his responsibility if necessary. He will also be responsible for the integration of the whole pipeline.
- Zhihan CHEN is the point of contact of the team. In charge of exploring different neural networks suitable for Chinese sentiment analysis and do the corresponding benchmarking.
- Zixu WANG is in charge of designing the structure of the neural network and hyperparameters tuning with his knowledge of natural language processing.

All of the three members will contribute to the software engineering part of this project.

VI. Risks & Mitigations

Table I presents the risks we may encounter and our corresponding mitigation strategies.

VII. Schedule

Table II shows our tentative schedule and corresponding milestones. Milestones in bold are parts of the deliverables.

Table II

<table>
<thead>
<tr>
<th>Date</th>
<th>Milestone</th>
<th>Tentative Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 1, 2017</td>
<td>Complete Detailed Project Plan</td>
<td></td>
</tr>
<tr>
<td>October 15, 2017</td>
<td>Negotiate with Microsoft to decide the domain of focus. Fetch data from Microsoft.</td>
<td></td>
</tr>
<tr>
<td>November 19, 2017</td>
<td>M1 Research on NLP topic. Complete data preprocessing pipeline. Get familiar with Microsoft CNTK. Design, implement and train the first prototype using existing model.</td>
<td></td>
</tr>
<tr>
<td>January 7, 2018</td>
<td>M2 First presentation and interim report. Implement the first satisfactory model.</td>
<td></td>
</tr>
<tr>
<td>January 21, 2018</td>
<td>Get access to Microsoft Azure. Complete benchmarking of several existing models.</td>
<td></td>
</tr>
<tr>
<td>February 18, 2018</td>
<td>M3 Design and implement the final model.</td>
<td></td>
</tr>
<tr>
<td>March 11, 2018</td>
<td>Complete hyperparameter tuning. Deploy the model on Azure. Implement APIs for external call.</td>
<td></td>
</tr>
<tr>
<td>March 25, 2018</td>
<td>Implement user interface and functionality to support file.</td>
<td></td>
</tr>
<tr>
<td>April 1, 2018</td>
<td>M4 Integration testing. Complete the final deliverable.</td>
<td></td>
</tr>
<tr>
<td>April 15, 2018</td>
<td>Complete final report and prepare for the presentation.</td>
<td></td>
</tr>
<tr>
<td>April 29, 2018</td>
<td>Prepare poster for project exhibition.</td>
<td></td>
</tr>
</tbody>
</table>
REFERENCES


