Group 4

ĐẠI HỌC TÔN ĐỨC THẮNG
TÔN ĐỨC THẮNG UNIVERSITY

KNOWLEDGE DISCOVERY AND DATA MINING

Final presentation

MEMBERS

INSTRUCTOR	HOANG ANH
MEMBER	HUYNH NHAT BAO

Table of contents

1 Topic Introduction

2 KerasNLP

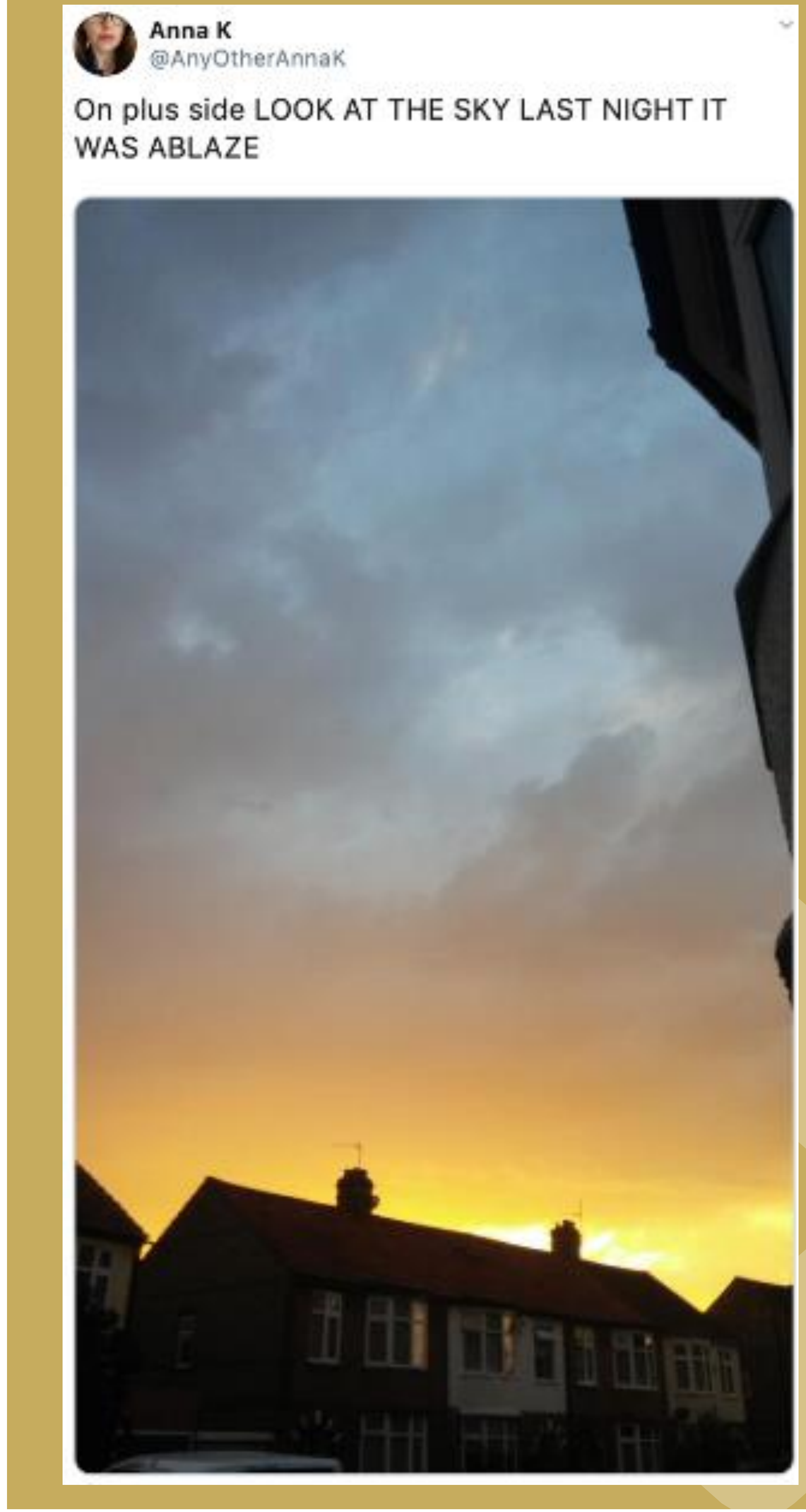
3 DistilBERT

4 Demo code

I.

Topic Introduction

This is a topic taken from a competition on Kaggle about natural language processing. In this topic we will process tweets on the social network Twitter. In this competition, you're challenged to build a machine learning model that predicts which Tweets are about real disasters and which one's aren't.



Anna K
@AnyOtherAnnaK

On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE

The author explicitly uses the word "ABLAZE" but means it metaphorically. This is clear to a human right away, especially with the visual aid. But it's less clear to a machine.

2.2 Overview of problem

Initial State

The research object is 10,000 tweets collected from the social network Twitter.

Goal Test

In this competition, you’re challenged to build a machine learning model that predicts which Tweets are about real disasters and which one’s aren’t.

Actions

The research topic uses the DistilBERT machine learning model to analyze and visualize data and expect results and therefore the actual results will score the model.

Transition Model

Machine learning model that predicts which Tweets are about real disasters and which one’s aren’t.

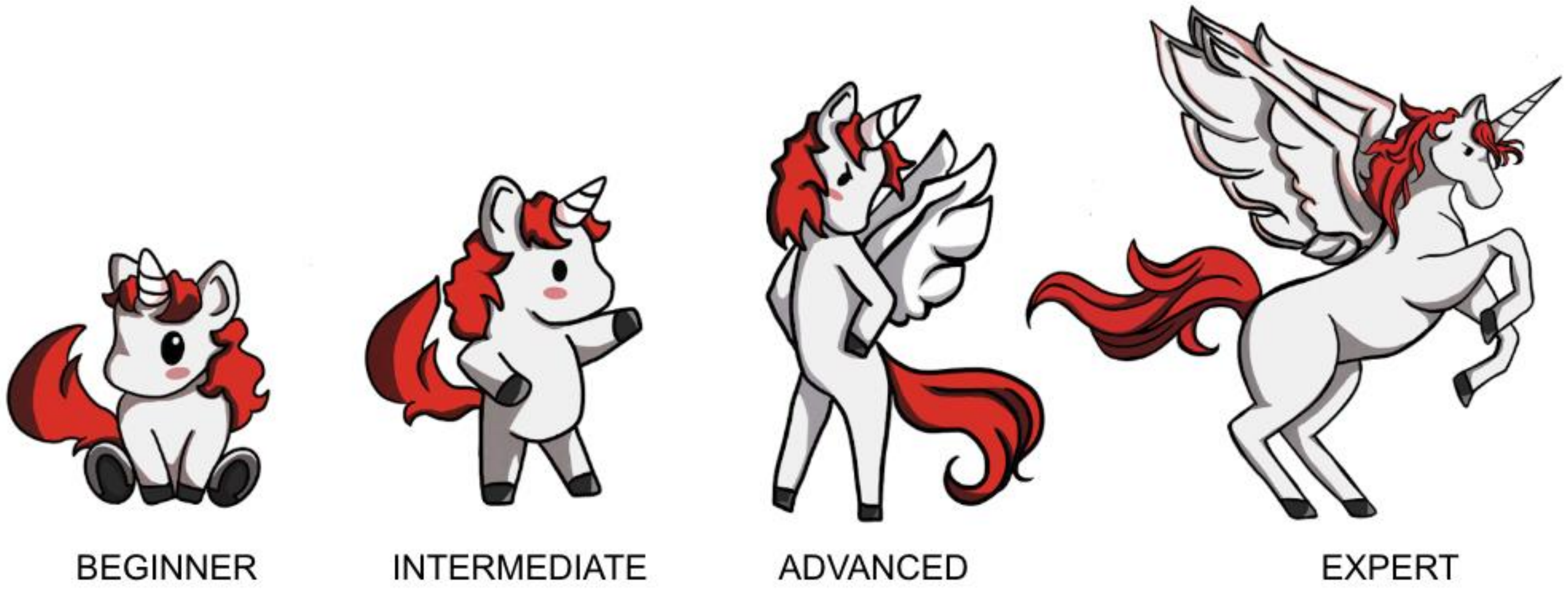
Path Cost

Per Actions is 1 steps

2. Keras NLP

2. Keras NLP

- KerasNLP is a library that natively supports language processing throughout the entire development process. This library is an extension of the core Keras API, all high-level modules are multi-layered or models.
- KerasNLP uses Keras 3 to work with any TensorFlow, Pytorch or Jax. In our article we use Tensorflow backend to train our models.



2. Keras NLP

Pros

- Easy Integration: KerasNLP directly integrates with frameworks like TensorFlow, JAX, and PyTorch, making it easy for users to use and combine with other tools.
- Flexible Customization: KerasNLP's models and layers can be flexibly customized to suit various NLP tasks.
- Good Performance: KerasNLP retains over 95% of BERT's performance on the GLUE benchmark while reducing parameters by 40% compared to bert-base-uncased.

Cons

- Limited Customization: While it is customizable, KerasNLP has less flexibility in creating complex architectures compared to some other NLP libraries.
- Handling Large Data: KerasNLP may not be the best choice for handling large or complex data due to potential memory management and processing speed challenges.

3. BERT and DistilBERT

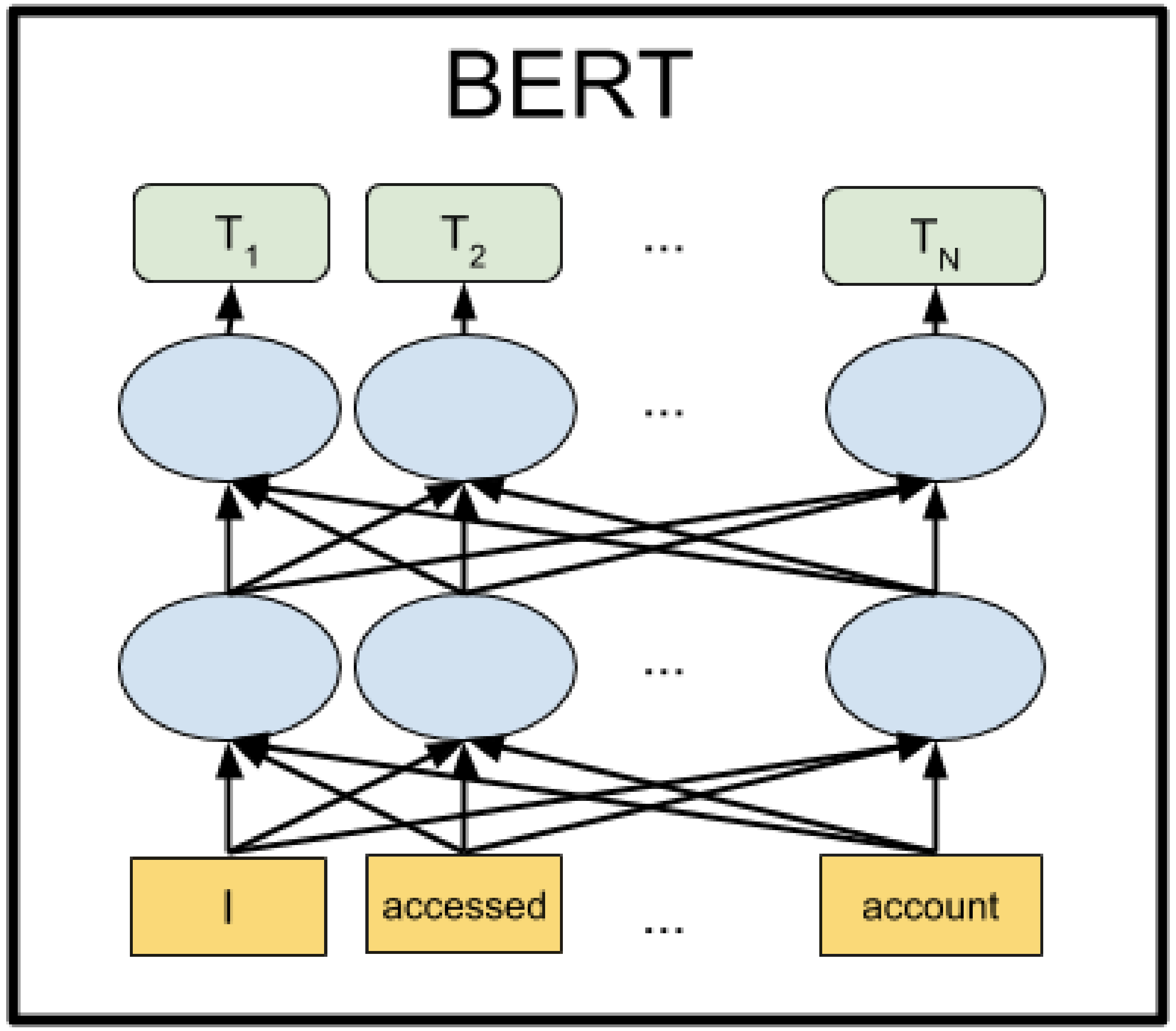
3.1 Overview of BERT

BERT(Bidirectional Representations from Transformers) is a natural language model developed by Google and announced in 2018. This is one of the most important advances in the field of natural language processing (NLP) in recent years.

- BERT is a model based on the Transformer architecture, specifically the Transformer encoder.
- The goal of BERT is to create semantic representations of words in their context by learning to understand the meaning of the word in both directions (left and right) of the sentence.
- Revolutionizing NLP: BERT has revolutionized the approach to NLP problems by using models that are pre-trained on big data and then fine-tuned on specific tasks.
- Opening a new era for NLP research and application: After BERT, many similar models such as GPT (Generative Pre-trained Transformer), RoBERTa, and T5 have been developed, further improving and expanding the capabilities of NLP. language models.

3.1 BERT architect

BERT will cut the transformer in half and only take the Encoder part on the left and remove the Decoder part on the right, leaving only the part where the text sentence is inserted and the output is the encoder output as shown:



3.1. Limitation

Pros

- Comprehensive context
- High performance on many NLP tasks
- Flexible fine-tuning capabilities
- Powerful pre-training

Cons

- Requires large computational resources
- Large model size
- Complexity in implementation
- Optimization and parameter adjustment
- Performance degrades on uncommon languages

3.2 Overview of DistilBERT

DistilBERT is an advanced natural language model (NLP), announced by Hugging Face in late 2019.

The goal: use distillation to reduce model size and increase processing speed while maintaining much of BERT performance.

3.2 DistilBERT vs BERT

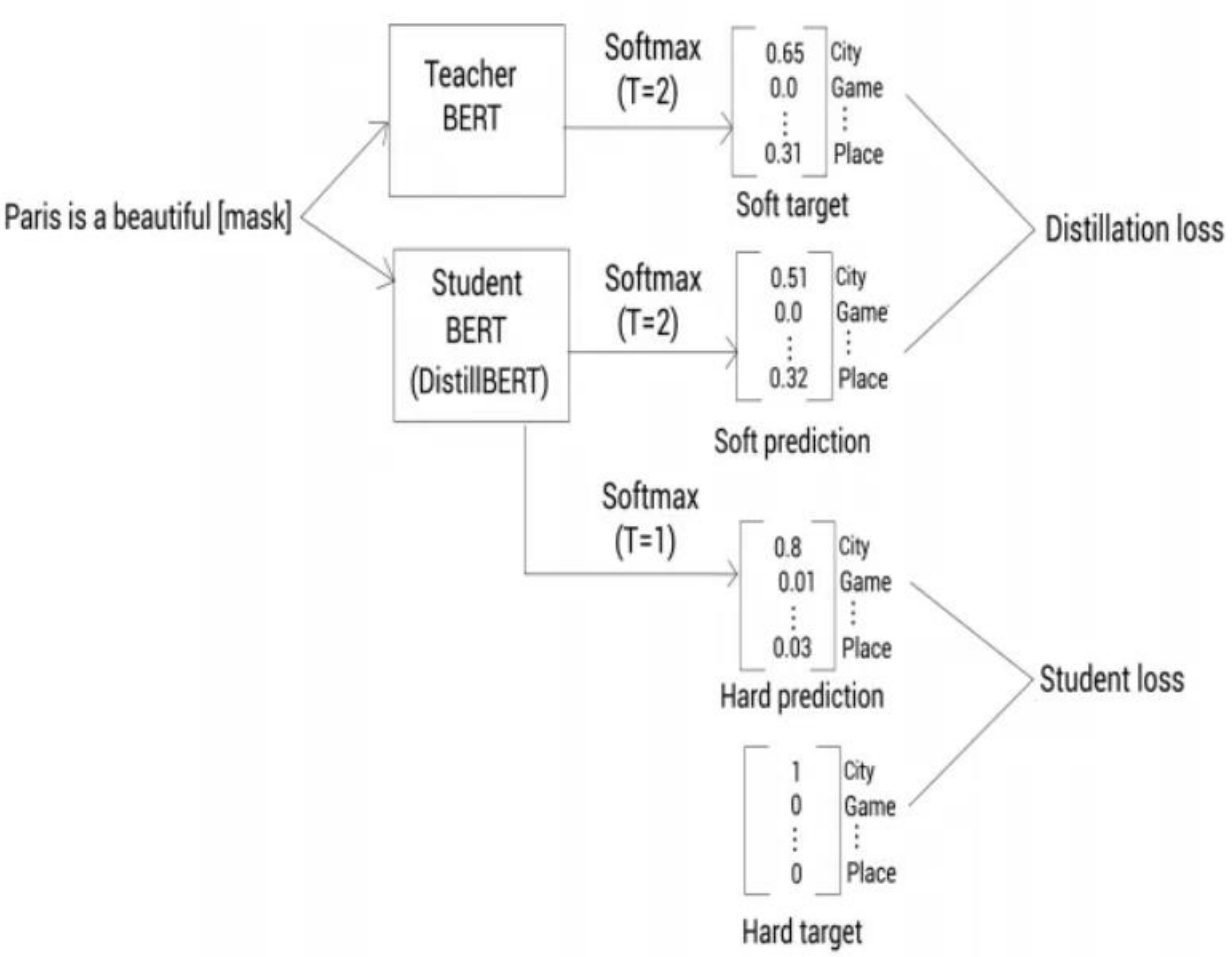
DistilBERT is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to

- Reduce the size of a BERT model by 40%
- While retaining 97% of its language understanding abilities
- And being 60% faster.

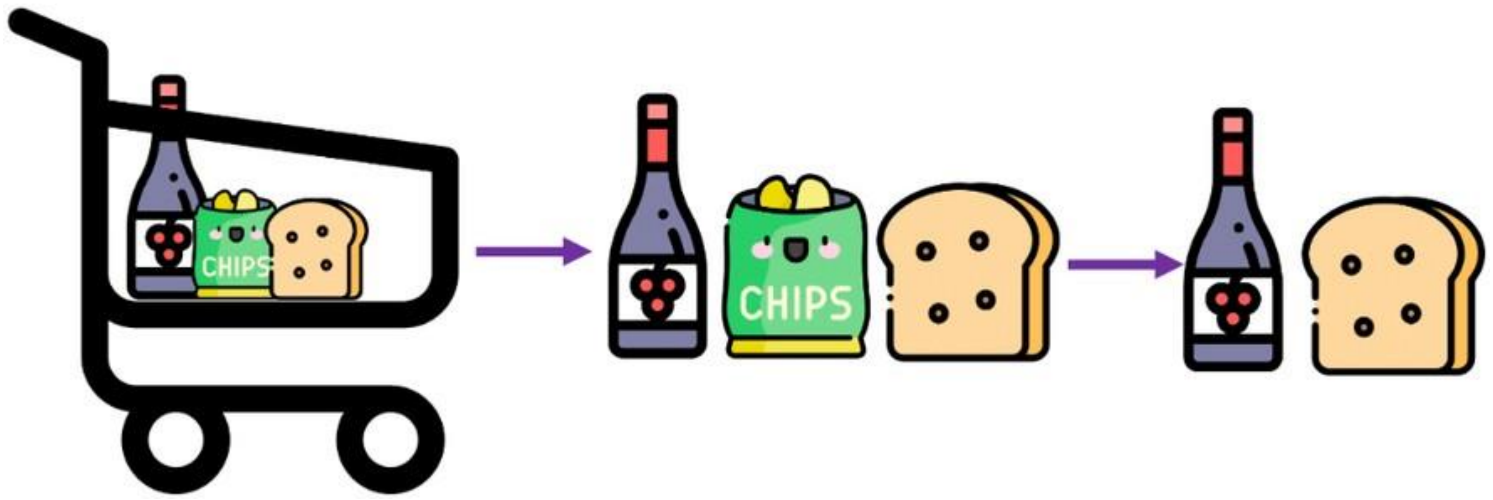
➔***DistilBERT Highlights:*** is an improved version of BERT. DistilBERT uses **knowledge distillation** to create a smaller, faster and more efficient version of the original BERT model while retaining much of its performance. With its compact size, DistilBERT is easy to deploy on resource-constrained devices that can be easily used for on-device applications.

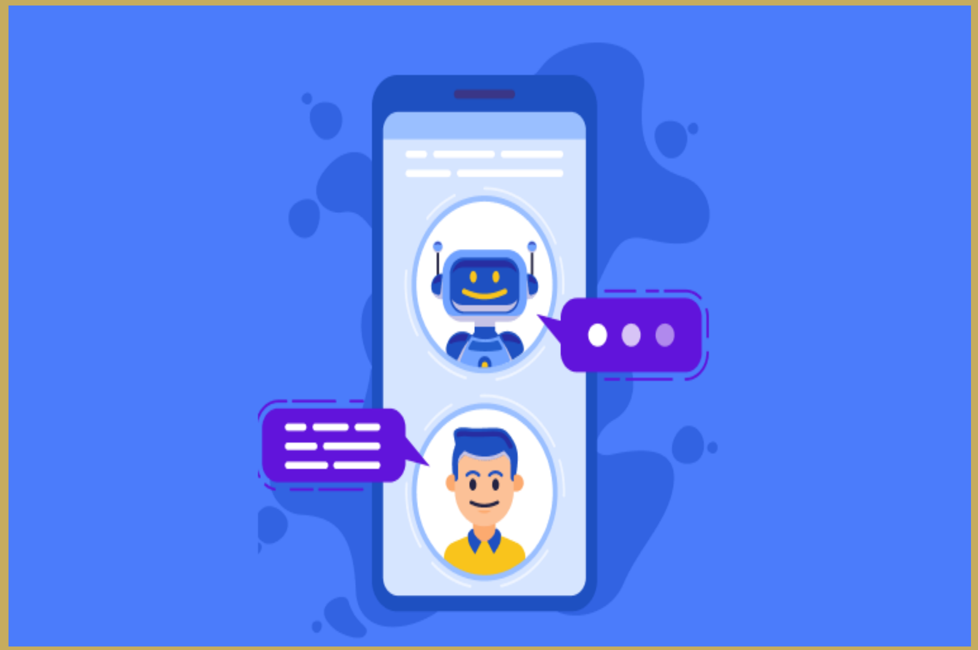
➔***Disadvantages of DistilBERT:*** Reduced accuracy compared to BERT which may affect applications requiring absolute accuracy. Depends on the quality and performance of the BERT model as a teacher. If BERT has problems or limitations, these can be passed back to DistilBERT.

3.2 How does DistilBERT work?



Purposes



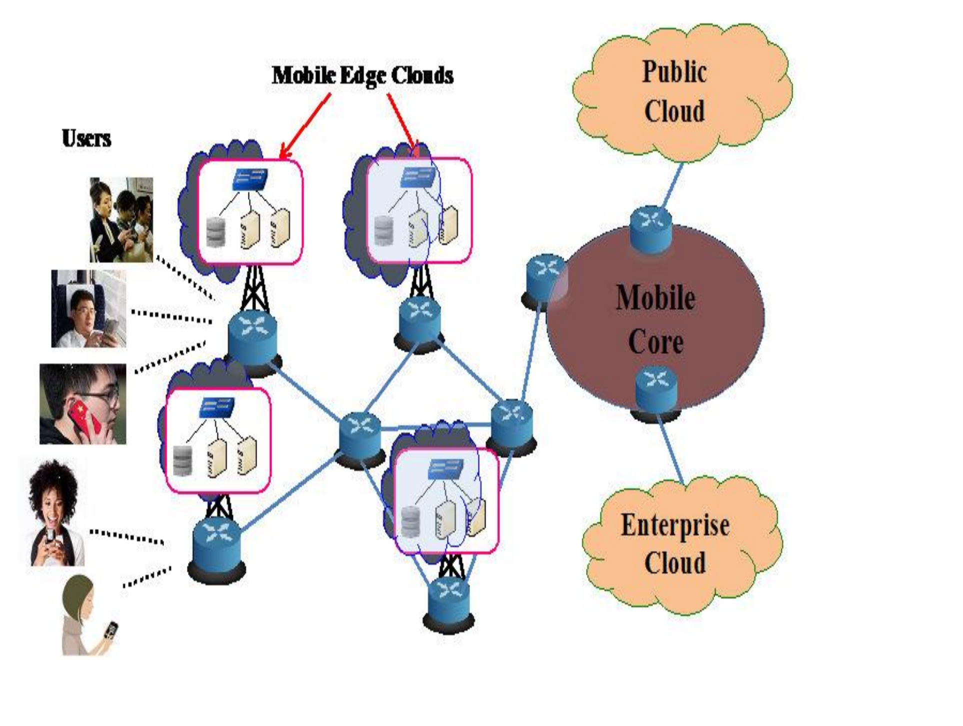


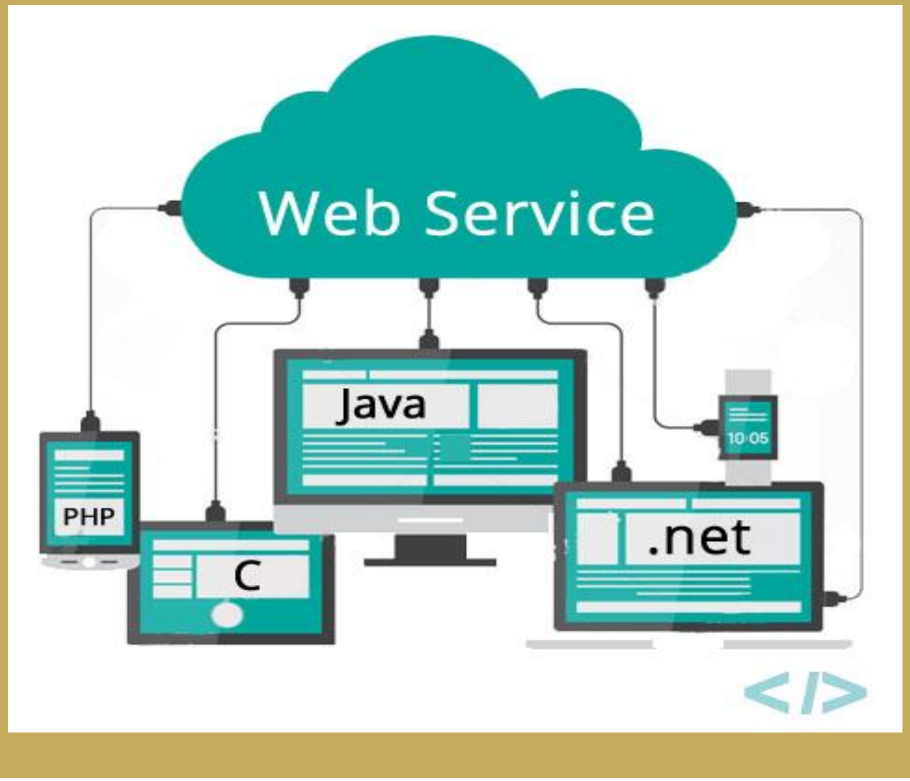
Chatbots and Virtual Assistants: For faster response times and reduced computational costs

Mobile and Edge Computing:

Deployment on devices with

limited computational resources.





Web Services: Improved speed and reduced operational costs in cloud-based services

4. SOURCE CODE

THANK YOU