AI based Paraphrasing Tool using Transformers

A presentation for Review - 1

Guide: Dr. Anand Pandey [HoD, IT, SRM-NCR]

Team:

Qazi Shuaib Was [RA1911003030326]

Zahid Hussain Khan [RA1911003030334]

Akshat Pareek [RA1911003030336]

Atulya [RA1911003030316]

Chapter – 1

INTRODUCTION

1.1 About

A paraphrasing tool is a software tool that is used to rephrase or rewrite text in a way that retains the original meaning, but uses different words and sentence structures. It works by analyzing the input text and generating new versions of the same text, which can be useful for simplifying and clarifying complex or technical language, and provide fluency and standard feel to the input text.

Application of Paraphrasers

1. Academic writing:
2. Content creation
3. SEO optimization
4. Language learning
5. Communication and collaboration

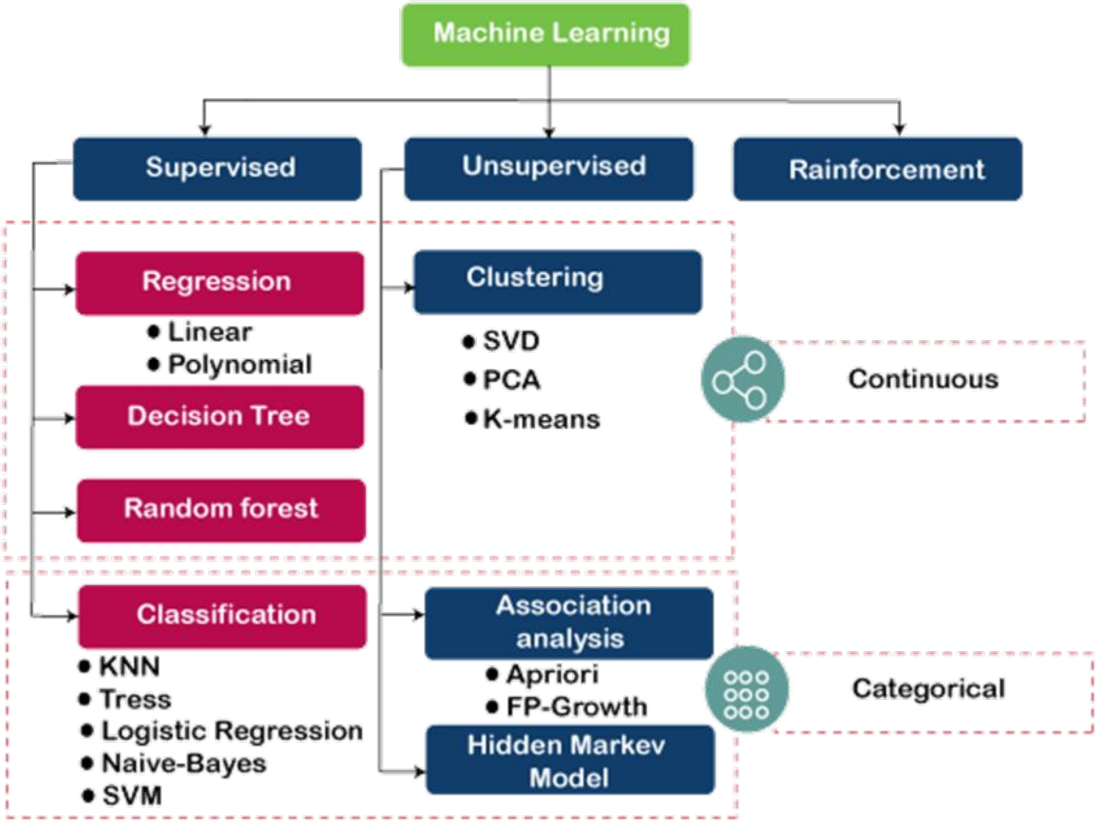
1.2 Artificial Intelligence and its branches

1.2.1 What is AI?

AI stands for Artificial Intelligence, which refers to the ability of machines to mimic human intelligence, such as the ability to learn, reason, and solve problems. AI involves the development of algorithms and software that can perform tasks that would typically require human intelligence, including natural language processing, image recognition, decision-making, and problem-solving.

1.2.2 What is ML ?

ML stands for Machine Learning, which is a subfield of AI that involves the development of algorithms and statistical models that enable machines to improve their performance on a task over time by learning from data. Machine learning algorithms can be used for tasks such as classification, regression, clustering, and reinforcement learning.


Machine Learning
Supervised
Unsupervised
Rainforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Tress
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Hidden Markev Model
Categorical

1.2.3 What is NLP ?

NLP stands for Natural Language Processing, which is a subfield of AI that focuses on enabling machines to understand and generate human language. NLP involves the use of algorithms and models that can analyse, generate, and manipulate natural language data, such as text and speech. NLP is used in a wide range of applications, including machine translation, sentiment analysis, and chatbots.

NLP has several strategies with different applications, including:

1. **Bag-of-Words (BoW) Model:** This algorithm is used to represent text data as a set of words, ignoring the order in which they appear. It creates a matrix where each row represents a document and each column represents a word. The values in the matrix represent the frequency of each word in the corresponding document.

2. **Term Frequency-Inverse Document Frequency (TF-IDF):** This algorithm is used to assign a weight to each word in a document, based on how important the word is in the document and how frequently it appears in the corpus. It is calculated as the product of term frequency (TF) and inverse document frequency (IDF).

3. **Word2Vec:** This algorithm is used to represent words as high-dimensional vectors, with each dimension representing a different feature of the word. It uses a neural network to learn the vector representation of words, based on the context in which they appear.

4. **Named Entity Recognition (NER):** This algorithm is used to identify and extract named entities, such as people, places, and organizations, from text. It uses machine learning algorithms to identify patterns in text that are associated with named entities.

5. **Sentiment Analysis:** This algorithm is used to determine the sentiment, or emotional tone, of a piece of text. It uses machine learning algorithms to identify the sentiment of a piece of text as positive, negative, or neutral.

6. **Latent Dirichlet Allocation (LDA):** This algorithm is used to identify topics in a corpus of text. It is an unsupervised learning algorithm that uses probabilistic modeling to identify the topics that are most likely to generate the observed data.

1.3 Frontend Framework

1.3.1 Django :

The Django Framework is a high-level Python framework that helps in the rapid development and clean, pragmatic design, django makes it easier to build applications more quickly, efficiently and with less code. Django is used for creating the User Interface (UI) for the application. The UI created by Django is easy to use so that the person which are from the non-technical field can also use the application for the prediction of disease without going anywhere any saving time and money.

1.3.2 Streamlit

Streamlit is an open-source Python library that enables developers to create web applications for data science and machine learning projects. It allows data scientists and machine learning engineers to easily create interactive and customizable dashboards and web applications to showcase their work.

Chapter – 2

LITERATURE SURVEY

2.1 Intro to Old Models

The very first approach dates to the paper by Kathleen R, McKeown from 1983, '**Paraphrasing Questions Using Given and new information**, *American Journal of Computational Linguistics*'. It suggests the usage of paraphraser component for a 'Question-answer' based system. It further says, "to ensure that the system has correctly understood the user. Such a paraphraser has been developed as part of the CO-OP system (Kaplan 1979). In CO-OP, an internal representation of the user's question is passed to the paraphraser, which then generates a new version of the question for the user.

- In ‘Chris Quirk, Chris Brockett, and William Dolan. 2004. **Monolingual Machine Translation for Paraphrase Generation**, *Association for Computational Linguistics*’ – The authors of this paper proposed “Statistical machine translation (SMT) tools to generate novel paraphrases of input sentences in the same language”. They proposed assignment of probabilities to categorize distinct word sequences as “meaning the same thing”.

- Gupta, A., Agarwal, A., Singh, P., & Rai, P. (2018). **A Deep Generative Framework for Paraphrase Generation**, *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1) – They proposed methods based on a combination of deep generative models (VAE) with sequence-to-sequence models (LSTM) to generate paraphrases, given an input sentence. Essentially, it is a Neural Network based approach.

- Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, Partha Talukdar; **Syntax-Guided Controlled Generation of Paraphrases**, *Transactions of the Association for Computational Linguistics 2020* – An approach to generate paraphrases was proposed by the authors which uses a mix of syntactic tree and tree encoder using Long Short Term Memory (LSTM) neural network. The main limitation is that it fails when the input dataset is noisy and grammatically not correct.

Chapter – 3

EXISTING PROBLEMS
AND
PROPOSED SOLUTION

3.1 Limitations in earlier work

Earlier works as already discussed are primarily focused on Machine Learning, statistics or Neural network based approaches to Paraphrasing tool. Clearly, these approaches lacked in many aspects as detailed in the following section.

1. Limited Vocabulary
2. Lack of Diversity
3. Inaccurate Paraphrases
4. Limited Context Understanding
5. Inability to Preserve Meaning
6. Time and Resource Intensive

3.2 Proposed Solution

Our proposed solution uses the following pipeline:

1. Data collection: Collect a large corpus of text data that the paraphrasing tool can use to learn from. The data should be diverse and cover a wide range of domains and topics.

2. Pre-processing: Pre-process the text data to clean it and prepare it for use in the paraphrasing tool. This step involves tasks such as tokenization, stop word removal, stemming, and lemmatization.

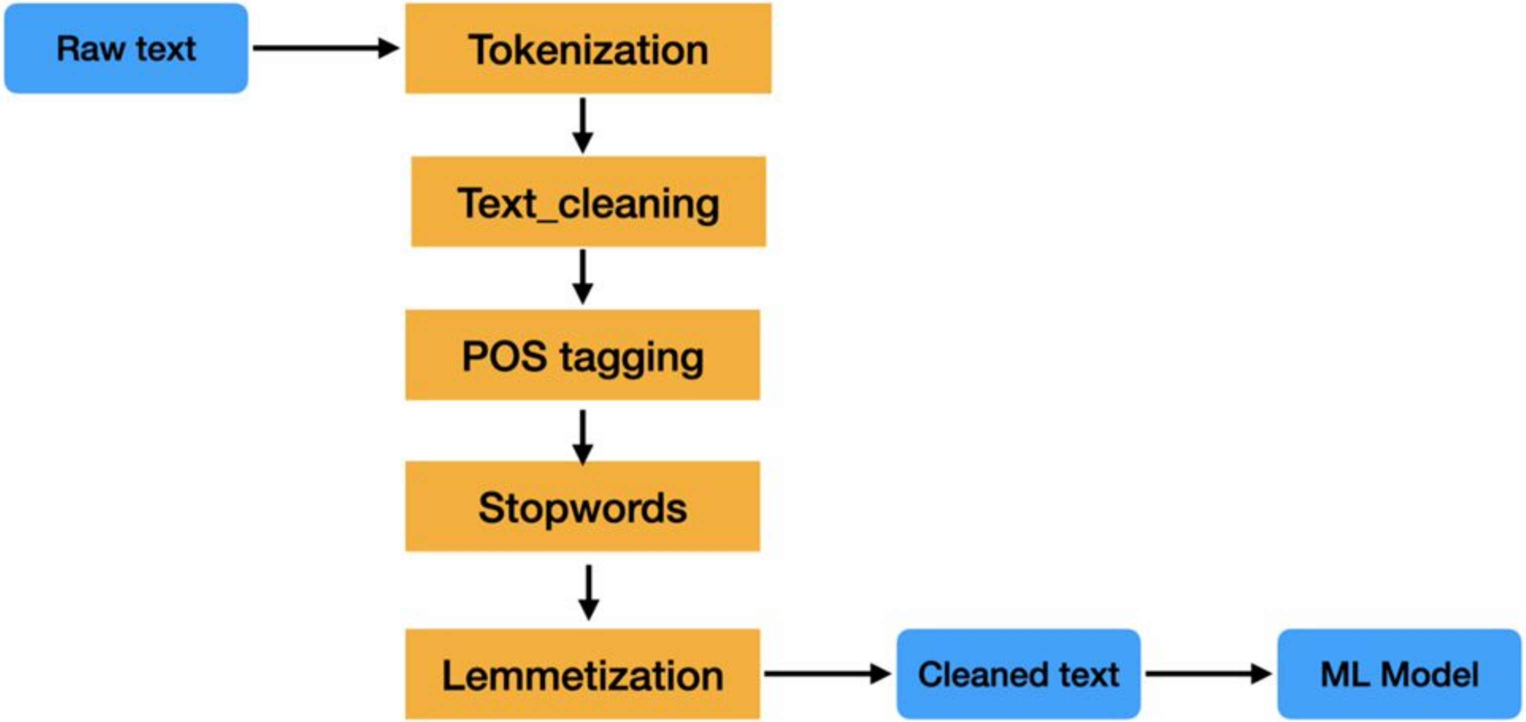
3. Data augmentation: Augment the pre-processed text data to create additional training examples for the model. We will be using techniques such as back-translation, synonym replacement, and paraphrasing to create more diverse and varied training data.

4. Training a language model: Train a language model on the pre-processed and augmented text data using techniques such as recurrent neural networks (RNNs) or transformer-based models like BERT or GPT.

5. Fine-tuning: Fine-tune the language model on a smaller dataset of human-generated paraphrases to improve its paraphrasing ability. We will be using transfer learning to adapt a pre-trained model to a specific task.

6. Integration: Integrate the paraphrasing model into a user-friendly interface such as a web app, a browser extension, or a command-line tool.

7. Evaluation: Evaluate the performance of the paraphrasing tool by comparing its generated paraphrases against human-written paraphrases using metrics such as BLEU, ROUGE, or other evaluation methods. We will be using inputs from our guide, faculties and college students to get feedback from users on the quality of the generated paraphrases.


Raw text
Tokenization
Text_cleaning
POS tagging
Stopwords
Lemmetization
Cleaned text
ML Model

The main model in our approach uses Text-To-Text Transfer Transformer (T5).

The Transformer model is based on the encoder-decoder architecture. The encoder is the gray rectangle on the left and the decoder is on the right. The encoder and decoder consist of two and three sublayers, respectively. Multi-head self-awareness, fully connected feedforward network, and encoder decoder self-awareness in the case of decoders (called multi-head attention) with the following visualizations).

Thank You