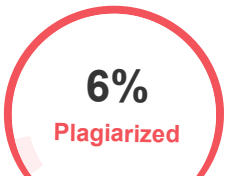
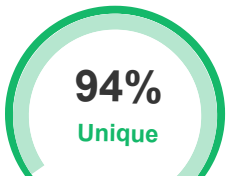
Apr 28, 2023

Plagiarism Scan Report





Characters:4398	Words:680
Sentences:34	Speak Time: 6 Min

Excluded URL None

Content Checked for Plagiarism

1.Introduction Sentiment analysis is a vital task in natural language processing (NLP), which involves determining the polarity of a given text, whether it is positive or negative. With the ever-increasing amount of data available on the internet, sentiment analysis has become essential in various fields, including marketing, politics, and social media monitoring. In this project, we aim to perform sentiment analysis on the IMDB movie review dataset, one of the most commonly used datasets in NLP. To perform sentiment analysis, we used neural networks such as SNN, CNN, RNN and LSTM. CNN is a popular choice for image recognition tasks, but its success in NLP has also been demonstrated in recent years. The model takes the raw text data as input and learns the features automatically, making it well-suited for the task of sentiment analysis. The architecture of the CNN model consists of several convolutional layers that extract the essential features from the text, followed by max-pooling layers that downsample the features. Finally, a fully connected layer is used to classify the sentiment of the input text. We trained the model on the IMDB dataset, which contains 50,000 movie reviews with a 50/50 split between positive and negative reviews. We preprocessed the data by tokenizing the text, converting the words to lowercase, and removing stop words and punctuation marks. We used the word2vec algorithm to convert each word into a vector representation, which is then used as input to the CNN model. Basic Concepts/ Literature Review Here are the basic concepts about some of the related tools and techniques used in this project: Deep Learning: Deep learning is a subset of machine learning that involves training neural networks with multiple layers to learn and make predictions from data. In this project, deep learning algorithms are used to train the sentiment analysis models. 2.1. Natural Language Processing Techniques Natural Language Processing (NLP) is the application of computational techniques to the analysis and synthesis of natural language and speech. It incorporates machine learning and deep learning models, statistics into computational linguistics i.e. rule-based modeling of human language to allow computers to understand text, spoken words, human language and sentiment. Various NLP techniques used for making the dataset ready for the model: Tokenization is the process of dividing a sentence/text into a list of meaningful tokens. A token is the smallest meaningful part of a text (words in this case). Sequencing is the process of generating a sequence of numbers from a large dataset of sentences after tokenization. Every token/word from a set of sentences is assigned numeric token index values. This way, a sentence can be represented as a sequence of token indexes (integers). Neural networks require inputs with the same shape and size. However, when preprocessing sentences and using them as inputs for a model, they are of unequal lengths. This is where padding and truncating come to

play. After defining the maximum number of words for each sentence, padding adds zeros at the end of the sequences of sentences smaller than the maximum size. Similarly, truncating drops the last words in sentences that exceed the maximum number of words. Tools used: The code uses several tools and libraries for building a sentiment analysis model on the IMDb movie review dataset. Here is a description of the tools used: Regular expressions: The re module in Python provides support for regular expressions, which are used for pattern matching and text processing. NLTK: The Natural Language Toolkit (NLTK) is a Python library for working with human language data, providing tools for text preprocessing, tokenization, and stemming. Keras: Keras is a high-level neural networks library in Python that provides an easy-to-use API for building and training deep learning models. Scikit-learn: Scikit-learn is a machine learning library in Python that provides tools for classification, regression, clustering, and data preprocessing. Tokenizer: The Tokenizer function in Keras is used to convert text to sequences of integers, which can be input to a neural network. Embedding: The Embedding layer in Keras is used to create word embeddings, which are vector representations of words that capture their semantic meanings.

Sources

6% Plagiarized

Natural Language Processing Examples And Definition

https://www.folio3.ai/blog/natural-language-processing-examples/#:~:text=In%20dictionary%20terms%2C%20Natural%20Language,analyse%20text%20using%20contextual%20cues.