CSCE 5290 Section 003 - Natural Language Processing (Spring 2024)

Course Instructor: - Dr. Sayed Khushal Shah
Project Proposal (Group - 9)

Description:

Team Members:
Shiny Shamma Kota- 11698716
Sai Kiran Reddy Kancharla - 11605339
Aashish Vinay Vasala - 11653925

Introduction:

Our project aims to perform two independent Natural Language Processing tasks. We aim to build machine learning models suitable for these tasks and fine tune them using algorithms to optimize the model's performance. The tasks are namely:
1)BBC News Summarization
2)Twitter Emotion Classification

TASK-A: BBC News summarization

Motivation:

In this digital era, we happen to go through a lot of new events which are occurring every day. There are news articles about politics, Business, Sports, Technology, etc. In this busy world we may not find time to read the whole news. Or, at times we don't comprehend the whole article. We are motivated to address this problem by condensing the articles into succinct summaries. This can help the user to save time and grasp the essence of the news better without losing the crucial information amidst the noise.

Significance:

Our project's significance lies in its ability to revolutionize the way we consume the news, disseminate the news and shape the views. By distilling the huge amount of information in the news articles into concise and succinct summaries we give value to the field.
Firstly, in this fast-paced digital era, there is a plethora of information available. It is quite challenging to be updated on contemporary issues on a timely basis. So, we seek to provide condensed and succinct summaries to readers without any bias, noise or without losing crucial information which can aid readers to grasp key points of complex issues and facilitate them in accessing crucial information.

Also, time is a precious commodity in this era. Readers may not have time to skim through the whole news by noting the crucial information. We seek to present essential facts in a format that is accessible and concise. Also, it empowers people to make their own decisions rather than being biased by sensationalized headlines.
In conclusion, we seek to promote informed citizenry which can lead to enhanced civic participation that can foster positive changes in the society.

Objectives:

- Cleaning and formatting the data
- Applying Natural language processing techniques to the data
- Creating a text summarization model specifically tailored for summarizing news articles.
- Optimizing performance and accuracy by fine tuning the model parameters.
- We seek to generate the summaries by the model which are accurate, consistent, concise and succinct without losing critical and crucial information and capturing key points in the news articles.

Features:

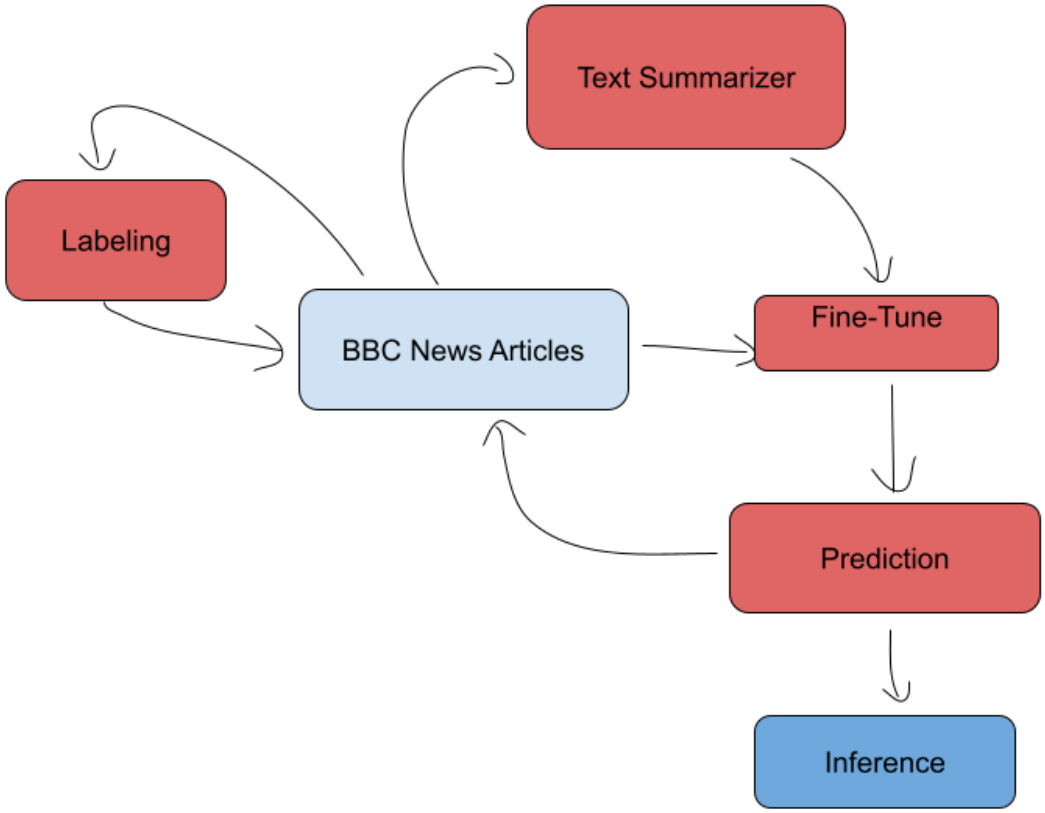
- We seek to create a custom summarization model using the dataset.
- We will be using Natural language processing techniques to preprocess the data and create understand the contextual information
- We seek to deliver a summarization algorithm that can provide concise summaries of news articles. We will be using algorithms to find the similarity between summary and the original text
- We seek to develop the model, train it on the dataset, optimize the parameters for better accuracy and evaluate it using specific metrics.
- Our project is unique in terms of being relevant and accurate in terms of summarization. We are going to employ natural language processing techniques and also, we will be evaluating using quantitative metrics and human assessment. Also, we make news more accessible and comprehensible.

Dataset:

- Our dataset [1] has text summarizations of news articles of BBC from 2004 to 2005.
- It has news articles from Politics.
- Every article has different types of summaries which can aid in observing from multiple perspectives and summarization styles for understanding and evaluating.
- The dataset requires preprocessing such as cleaning and formatting to ensure uniformity in the data. To beget better results.

In conclusion, the dataset has a variety of data. It consists of data from various sectors and multiple summaries on one article which provides different kinds of perspectives and context. The data is ample to train and test the data. It can aid in creating better summaries.

BBC NEWS Summarization Workflow Visualization:



Flow	Description
Data Preprocessing	In order to make the data comprehensible by the algorithm the data must be preprocessed such as data cleaning and formatting
Model Development	Model development includes designing and implementation of model's architecture

Training & Optimization	Training & Optimization means the training of the model and fine tune it on the preprocessed data also optimize its parameters for better accuracy
Evaluation	Evaluation means validating the model's performance using particular metrics.
Documentation	Documentation means creating a document that includes technical details and methodology of the project.

TASK B: Twitter Emotion Classification:

Motivation:

Recent past years has shown an increase in the number of social media applications. People have started to use it to such an extent that it has become a part of their daily lives. People post all sorts of information ranging from personal to professional. As technology has progressed, people communicate their opinions virtually in a global fashion. Among many such platforms is twitter. Twitter allows users to post their opinions in a public manner inn short paragraphs along with use of emojis and images according to the users' discretion. This motivates us to build machine learning models that can extract useful information into data to turn it further into insights. This helps decision making bodies, business etc., to know the pulse of the public and build their company objectives accordingly.

Significance:

The significance of our project lies in the fact that we build machine learning models that will help us understand and analyze human behaviour and emotion on a large scale. Sorting out tweets into emotion class labels will help simplify the analysis. It will help researchers understand social trends, issues, etc., also as discussed in motivation, will help business to analyze and create marketing profit making strategies. Sailunaz et al., [3] has created a machine learning model that can give recommendations to the user based on the emotion classification. Vo et al., has designed a model that can analyze the emotion of tweets during earthquakes, which in turn will help authorities know the intensity of loss in an area. Apart from these there are many such applications. Job that is tedious will be done in a matter of seconds, thanks to modern hardware equipment and machine learning algorithms.

Objective:

The objective of our project is to build a NLP machine learning algorithm that will be able to categorize the content of tweets into emotional labels. We train the model on a label dataset containing tweets and their labels such as joy, anger, fear etc. Next, we aim to fine tune the model so that we get an optimal working performance of the model. After fine tuning the model is trained with the training dataset and test the model for its accuracy with the testing dataset. By achieving these objectives our team aims for an understanding of the project along with an optimal performing model.

Features:

Our project consists of:

- We will use the data collected from kaggle [2], containing tweets and their emotion labels.
- We will perform tokenization of the dataset.
- We will perform feature extraction to get useful information form the dataset.
- We will select an appropriate machine learning model and train it using our dataset.
- Next, we fine tune the model using pre-trained language models.
- Next, we evaluate the performance of the model using evaluation metrics.

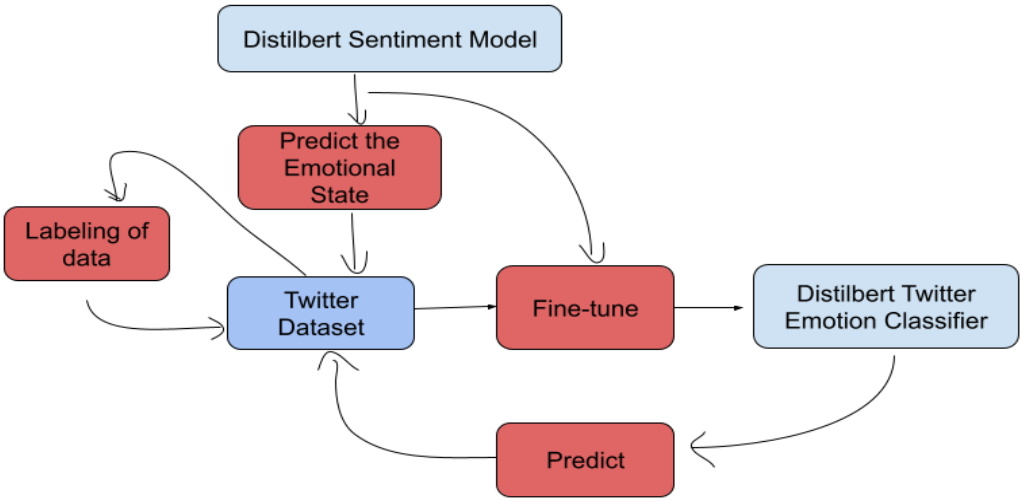
Dataset:

- We have used Emotion Dataset for Emotion Recognition Tasks dataset for twitter emotion classification task.
- Taken from kaggle[2], this dataset contains a collection of 2000 tweets classified between 6 emotion labels namely:
    - sadness
    - joy
    - love
    - anger
    - fear
    - surprise
- Tha language of the twitter data is in english.
- The dataset is split into training and testing dataset in the model building.

Twitter emotion Classification Workflow Visualization:

Flow	Description

Data Preprocessing	We preprocess the data according to the model requirement. We perform tokenization for ease of model building.
Model Development	Model development includes designing and implementation of model's architecture
Training & Optimization	We train the model with training dataset. We fine tune the model to get optimal parameterss.
Evaluation	We will use evaluation metrics such as confusion matrix etc, to assess the model.



References:

[1] https://paperswithcode.com/dataset/bbc-news-summary

[2] https://www.kaggle.com/datasets/parulpandey/emotion-dataset

[3] Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. Journal of Computational Science, 36, 101003. https://doi.org/10.1016/j.jocs.2019.05.009. (https://www.sciencedirect.com/science/article/pii/S1877750318311037)

[4] Vo, Bao-Khanh Ho, and N. I. G. E. L. Collier. "Twitter emotion analysis in earthquake situations." *Int. J. Comput. Linguistics Appl.* 4.1 (2013): 159-173.