cs777_project_proposal

February 20, 2023

0.0.1 Title: “Classifying Emotion in Text-Based Content”

Author : Hicran Arnold

0.1 Introduction

The exponential growth of digital communication has resulted in a massive amount of text-based content that expresses a wide range of emotions. Understanding the emotions convey in this content is essential for businesses, organizations, and individuals to understand their audiences better and gauge public sentiment about their products and services. However, manual analysis of these texts is time-consuming and impractical, making it necessary to develop automated methods for emotion classification. This project aims to develop a machine learning model for emotion classification in text-based content, which will provide a more efficient and accurate method for analyzing the emotions conveyed in digital communication.

0.2 Objectives:

The main objectives of this project are:

1. Develop a machine learning model that accurately classifies different types of emotions expressed in text-based content.
2. To explore the effectiveness of different text preprocessing techniques in improving the accuracy of sentiment analysis on our emotion dataset. Specifically, investigate the impact of techniques such as steaming, and stop word removal on the performance of sentiment analysis models.
3. To identify the most informative features for emotion classification in text and investigate the impact of different feature selection methods on the model's performance
4. Evaluate the performance of the emotion classification model using appropriate metrics such as accuracy, precision, recall, and F1-score. Additionally, conduct error analysis to gain insights into the model's strengths and weaknesses and identify further improvement areas.

Research Questions:

1- Can a machine learning model accurately classify different types of emotions expressed in text-based content, and if so, which model is most effective?

2- What is the expected accuracy of the machine learning model in classifying different types of emotion based on text-based content?

3- How does the performance of the sentiment analysis model vary when using different text preprocessing techniques?

4- What are the most informative features for emotion classification in text, and how does the performance of a sentiment analysis model vary when using different feature selection methods?

0.3 Methodology:

1. Data Collection: The data set is English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. The data includes training, test , and validation text files. The training data set has 15999, and validation and test txt files have 1999 lines of text each. The entire data set is 738 kB. The data set is publicly available and can be found at PRAVEEN-Kaggle and dair-ai/emotion_dataset.(1)
2. Pre-processing: The data will be processed by using different prep processing techniques like stop word removal, stemming/lemmatization, and spell checking to remove any noise, inconsistencies, so that remove any noise, inconsistencies or irrelevant information from the data so the resulting model can accurately identify patterns and make predictions.
3. Machine Learning Models: The research will implement and evaluate several machine learning models algorithms such as Decision Trees and Random Forests, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN) ,and Neural Networks on the emotion dataset after applying various text preprocessing techniques and planning to use PySpark and Tensorflow machine learning library to implement these models.
4. Evaluation: I will evaluate the performance of the machine learning models using standard evaluation metrics such as accuracy, precision, recall, and F1-score. The data set has been split into train/test/validation split, with 80% of the data used for training, 10% for validation, and 10% for testing. During training, I will monitor the performance of the models on the validation set and If necessary, I will use early stopping to prevent over-fitting. To further analyze the performance of the models, I will create confusion matrices for each model to visualize the distribution of true positives, false positives, true negatives, and false negatives. This will provide insights into the types of errors made by the models and allow me to fine-tune them accordingly. I will also conduct a thorough error analysis to identify the strengths and limitations of each approach. This will involve examining the misclassified instances to identify any patterns or trends that may indicate areas for improvement.

0.4 Expected Results:

The expected results of my projects are:

1. An accurately trained sentiment analysis model can classify different types of emotions in text-based content with high accuracy.
2. A comparison of the performance of different text preprocessing techniques on the accuracy of sentiment analysis models, with findings that demonstrate the effectiveness of certain techniques in improving accuracy.
3. An identification of the most informative features for emotion classification in text, with insight on how different feature selection methods impact performance sentiment analysis models.

0.5 Conclusion:

This project has the potential to contribute to the development of more accurate and efficient methods for emotion classification in social media texts, with applications in fields such as marketing, customer service, and public opinion analysis. The results of this project will also provide insights into the strengths and limitations of different machine learning approaches for emotion classification.

0.6 Refrences

(1) Dataset Info

License: Educational purposes only

title = {CARER}: Contextualized Affect Representations for Emotion Recognition

author = Saravia, Elvis and Liu, Hsien-Chi Toby and Huang, Yen-Hao and Wu, Junlin and Chen, Yi-Shin

book title = Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing

year = 2018

address = Brussels, Belgium

publisher = “Association for Computational Linguistics”

url = https://aclanthology.org/D18-1404/

doi = “10.18653/v1/D18-1404”,pages = ”3687–3697

abstract = “Emotions are expressed in nuanced ways, which varies by collective or individual experiences, knowledge, and beliefs. Therefore, to understand emotion, as conveyed through text, a robust mechanism capable of capturing and modeling different linguistic nuances and phenomena is needed. We propose a semi-supervised, graph-based algorithm to produce rich structural descriptors which serve as the building blocks for constructing contextualized affect representations from text. The pattern-based representations are further enriched with word embeddings and evaluated through several emotion recognition tasks. Our experimental results demonstrate that the proposed method outperforms state-of-the-art techniques on emotion recognition tasks.”,