Human Sentiment Extraction Using CNN Model

Chandrasekhar Parisa

Supervised By: Dr. Thangarajah Akilan

Department of Computer Science, Lakehead University, Ontario, Canada

Abstract

Humans are biologically more sensitive in terms of emotions and the surrounding events compared to any other specious in the animal kingdom. Various scientific frameworks have emerged to explain the human emotions as difference in our neurological makeup and brain difference. Modern techniques exploit the power of computational linguistic to analyze the emotional state of humans when they are exposed to certain events based on their reviews, opinions, and expression. It combines various areas of study, such as data analysis, natural language processing, and text mining. It has become an emerging intelligent tool in healthcare that help physicians in understanding and improving patient health.

Introduction

Sentiment Analysis (SA) is a field of research in text mining. SA is the computational treatment of sentiments, opinions, and subjectivity of text. Sentiment analysis combines various areas of study such as data mining, natural language processing, and text mining, and is becoming a real significance to organizations as they coordinate online commerce into their tasks. This Paper focuses on implementing a scalable and robust Convolutional neural network for the problem of text based movie review sentiment analysis.

A classification task with multiple classes; e.g., order a lot of pictures of natural products which might be apples, oranges, or pears. Multi-class characterization makes the supposition that each example allocated to one and just one label: an organic product can be either an apple or a pear yet not both simultaneously. Interpersonal organizations and smaller scale blogging sites, for example, Twitter have been the subject to numerous examinations in an ongoing couple of years. Programmed conclusion investigation and feeling mining present a hotly debated issue of study.

The general steps involved in the proposed framework is depicted by the process flow shown in Fig. 1.

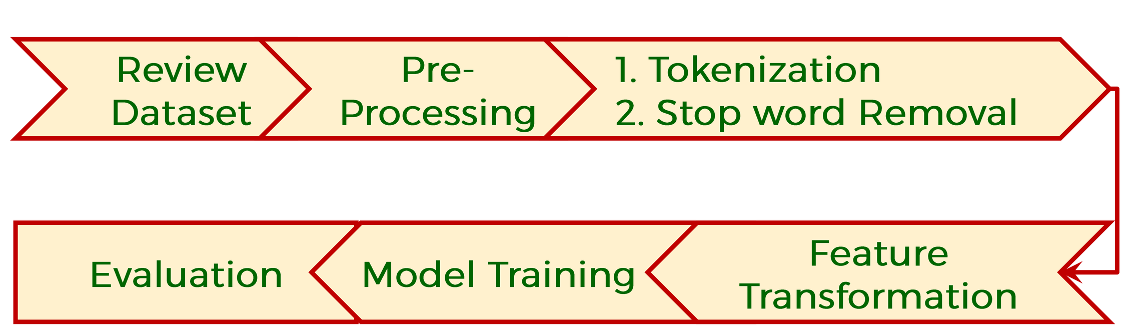

Fig. 1. The Process flow.

1. **Step 1:** Review of the dataset to understand the nature of the collected data samples.
2. **Step 2:** Data cleansing and removal of noise.
3. **Step 3:** Converting raw data into tokens using the standard *Sklearn* tokenizer with n-gram size in the range of [1, 3]. Also removal of English stop words and common key words in the raw data samples after the Tokenization.
4. **Step 4:** The tokenized data is transformed into numerical vectors through TF-IDF vectorizer from the *Sklearn*.
5. **Step 5:** The model training of the proposed CNN network (refer experimental set up for detail).
6. **Step 6:** The model is evaluated based on the following matrices: Precision, recall, and accuracy.

Methodology

This study focuses on an automatic methodology to estimate the state of human emotions based on their written reviews. Specifically, it evaluates human sentiment subjective to a movie based on their text comments using Convolutional Neural Network (CNN).

The model utilizes two 1D-convolution (Conv) layers and two dense layers. The Conv layers use ReLU activations followed by Max Pooling operations. The first dense layer employs ReLU activation, while the last dense layer uses Softmax activation as a multi-class classifier. The model is trained on the benchmark Rotten Tomatoes train dataset with Term Frequency–Inverse Document Frequency (TF-IDF) feature and sparse categorical cross entropy objective function.

Visualizing Data

In order to better understanding of the data we handle, it important to visualize their distribution across classes. Fig. 2 shows the frequency with respect to the sentiments found in the Rotten Tomatoes Movie Reviews (RTMR) dataset from the Kaggle via a Bar-chart. From the Bar-chart, it is observed that the sentiment class label#2 has the highest number of samples as its frequency is 0.5. While, the sentiment class label#0 has the least number of samples with class frequency approximately of 0.08. Coming to sentiment class label#4, it is slightly higher than class 0, which stands in the second position from last. Later in the list the class label#1 and 3 have almost same number of samples.

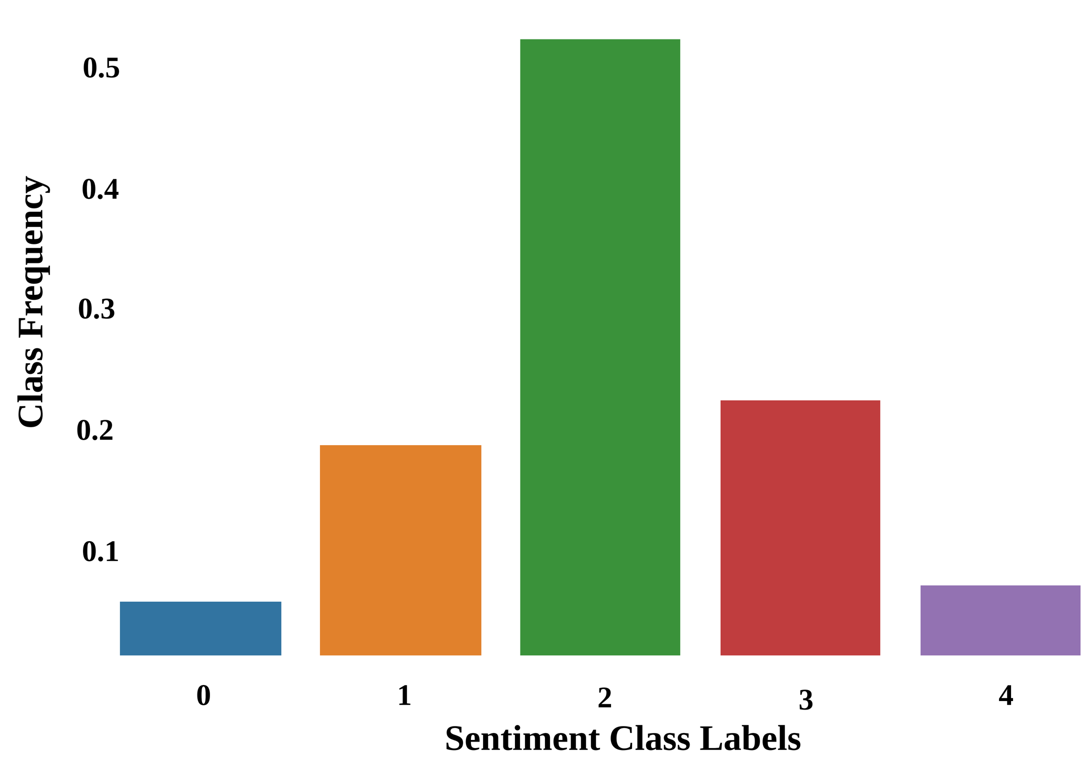

Fig. 2. The Sentiment Class Distribution.

WordCloud is a technique that visually represents the text or words that are the most frequent among the given text corpus. It automatically represents the text in respective fonts, font sizes, and color. For this representations, we use a standard Wordcloud library that needs to be imported along the numpy, pandas libraries in Python.

Fig. 3 visualizes the most frequent words in the sanitized train dataset, where it is observed that word "film" is more frequent in the given dataset than "word" movie. Also, some optional arguments of the WordCloud, like max word, max font size, and background color are can be used for better representations.

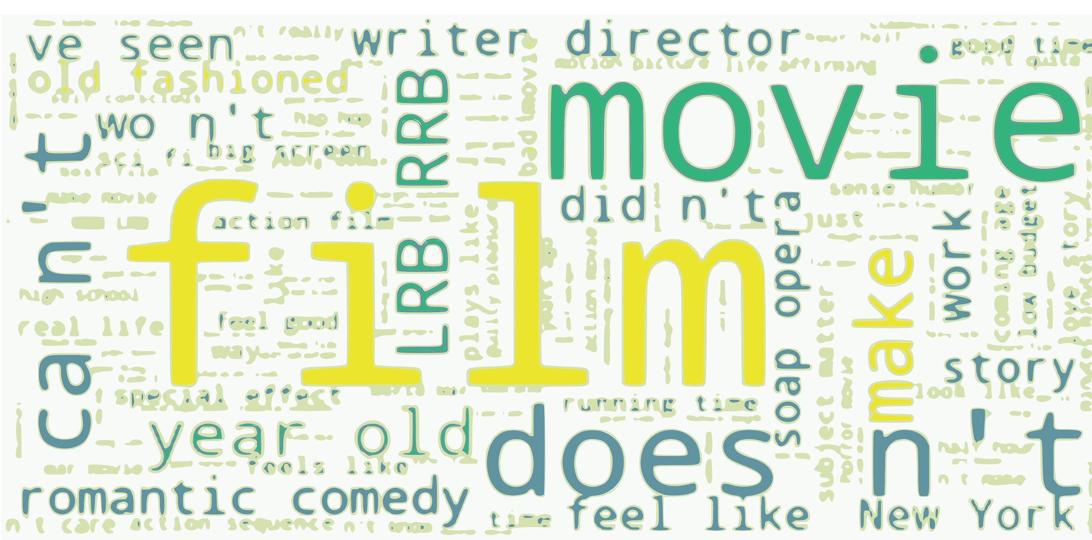

Fig. 3. WordCloud Visualization of the RTMR Dataset.

Model

In this research, the simple Convolutional Neural Network(CNN) model is used. CNN which is also called as Convnets are one of the most energizing advancements in AI now-a-days. The Table. 1 shows the Model Summary, which clearly explains the activation layers and pooling operations used. The Layer Conv1D has parameters 256 with output shape of (None, 5808, 64), followed by MaxPooling 1 operation is performed with output shape of (None, 2904, 64).

Again Conv1D one more time uses 24704 Parameters with output shape of (None, 2902, 128). Later MaxPooling operation is performed with shape (None, 1451, 128). Followed by Dropout and Flatten operations are performed. Softmax Activation function is applied at the end to classify the data. In this summary we can observe that Total Parameters the data got trained is 23,798,917.

Table. 1 Model Summary

Layer (Type)	Output Shape	Param #
Conv1d_1 (Conv1D)	(None, 5808, 64)	256
max_pooling1d_1 (MaxPooling1)	(None, 2904, 64)	0
conv1d)_2 (Conv1D)	(None, 2902, 128)	24704
max_pooling1d_2 (MaxPooling1)	(None, 1451, 128)	0
dropout_1 (Dropout)	(None 1451, 128)	0
flatten_1 (Flatten)	(None, 185728)	0
dense_1 (Dense)	(None, 128)	23773312
dropout_2 (Dropout)	(None, 128)	0
activation_1 (Activation)	(None, 128)	0
dense_2 (Dense)	(None, 5)	645
activation_2 (Activation)	(None, 5)	0
Total Number of Parameters: 23, 798, 917 Total Number of Trainable Params: 23, 798, 917		

Experimental Setup and Results

The model is trained on the training samples from the RTMR dataset. It consists of 156,060 full sentence reviews. After vectorization, the feature vector of each sample has a dimension of 5810. The training takes a batch size 128 samples at a training step. Coming to training setup the learning rate is set to 0.001, with a decay rate of 0.0001 using *Adam* optimizer. The optimizer utilizes *sparse categorical crossentropy* loss function. Thus, the model is trained for 100 epochs that achieves a training accuracy of 81%.

The training process on a single Nvida Tesla P100 takes approximately 30 minutes. The model evaluation is carried out on the testing samples of RTMR dataset that consists of 66,292 full sentence reviews. The proposed model achieves only about 52%. Although the results is not promising one at this stage, we believe that it is largely due to the class imbalance in the raw data samples of RTMR. Therefore, the performance can be improved by the class balancing techniques, viz. oversampling, data reduction using for instance PCA, and collection of fixed number of random samples.

Also, improving the network structure of the CNN model with more layers, and fusion of multiple activations on the feature extraction pipeline will certainly provide us better accuracy in the future.

Conclusion

Now-a-days, the health care industry is competitively becoming more focused on how to improve the patient experience. Sentiment analysis is one of the emerging trends in the public health care. Sentiment analysis can help healthcare providers gain a competitive edge over the competition and improve their services based on the feedback provided.

Sentiment analysis provides insight into the performance of a healthcare facility. The collated analysis of comments can provide a clear indication of what patients need from your practice, how the employees are rated, and other concerns that may not be obvious to the management.

This work investigates an implementation of a scalable and robust deep learning model, specifically a CNN for the problem of text-based human sentiment analysis. Through, empirical approach a best model is derived that achieves competitive results. The current model can be further improved by adding feature fusion layers and hyper parameter tuning.

Acknowledgements

This Research is supervised by Dr. T. Akilan.

Contact Details

Chandrasekhar Parisa
Phone: 807-357-6819
Mail: cparisa@lakeheadu.ca