Emotion Intensity in text using Machine learning and Deep learning

Abstract

This paper describes the approach to predict emotion intensity in text data.

Used different machine learning models to train data. And also used deep learning methods to find intensity of emotions.

Experiments on different models and various features sets are described and analysis on results has also been presented.

Here, emotion intensity is detected using machine learning and deep learning.

Objective

To identify the intensity of emotions in text.

Data Understanding

The data set consists of 3 types: training data, development data and test data. Each data have 4 files one for each emotion like anger, sadness, fear, joy.

There are 3 columns in the data set Id , text, label, intensity.

Label column have different emotions like anger, fear, sadness, joy.

Intensity varies between 0 and 1 , '0' indicates emotion with less intensity and '1' indicated high intensity of emotion.

Data preprocessing

After getting data it is necessary to clean data.

Data preprocessing should be done before training the data.

The train data for anger, sadness, fear and joy concatenated into single data. Removed the NAN rows if any.

Created dummy variables and data visualized using bar plots and scatterplots.

The text in each emotion combined to form a paragraph and then applied tokenization. Then removed all punctuation marks and stop words.

Bag of words techniques and TFID technique applied on the cleaned data.

Word Embedding is one such technique where we can represent the text using vectors. The more popular forms of word embeddings are:

BoW, which stands for Bag of Words

TF-IDF, which stands for Term Frequency-Inverse Document Frequency

Bag of Words (BoW) Model

The Bag of Words (BoW) model is the simplest form of text representation in numbers. Like the term itself, we can represent a sentence as a bag of words vector (a string of numbers).

Term Frequency-Inverse Document Frequency (TF-IDF)

Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

The data is in text, it is necessary to convert into numbers which can be done using bag of words and TFID using which text is converted into vectors.

Lexicon-based sentiment analysis

Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text. With this approach a

dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Different approaches to creating dictionaries have been proposed, including manual and automatic approaches. Generally speaking, in lexicon-based approaches a piece of text message is represented as a bag of words.

DNN

Deep Neural Networks have an input layer, an output layer and few hidden layers between them. These networks not only have the ability to handle unstructured data, unlabeled data, but also non-linearity as well. They have a hierarchical organization of neurons similar to the human brain. The neurons pass the signal to other neurons based on the input received. If the signal value is greater than the threshold value, the output will be passed else ignored.

Modelling

Next step is training the data, the output which is intensity is a continuous variable hence used regression methods.

- Linear regression
- Ridge regression
- Decision tree
- KNN
- SVR (Support vector regression)

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called multiple linear regression.

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

Trained the data using these models.

Model Evaluation

Emotion Intensity in text using Machine learning and Deep learning

	Actual Intensity	Predicted Intensity
0	0.479	0.492791
1	0.458	0.386838
2	0.562	0.492455
3	0.500	0.598758
4	0.708	0.531764
...	...	...
342	0.580	0.499381
343	0.170	0.458617
344	0.396	0.320307
345	0.156	0.340903
346	0.704	0.492331

Pearson correlation coefficient:

Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar rank between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.

```
Pearson correlation between  Anger_Predicted  and  Anger_Actual :
Pearson correlation for gold scores in range 0.5-1 between  Anger_Predic
Pearson correlation between  Fear_Predicted  and  Fear_Actual :  0.49612
Pearson correlation for gold scores in range 0.5-1 between  Fear_Predict
Pearson correlation between  Sad_Predicted  and  Sad_Actual :    0.61086
Pearson correlation for gold scores in range 0.5-1 between  Sad_Predicte
Pearson correlation between  Joy_Predicted  and  Joy_Actual :    0.48863
Pearson correlation for gold scores in range 0.5-1 between  Joy_Predicte
Average Pearson correlation:     0.5300434815728654
Average Pearson correlation for gold scores in range 0.5-1:      0.38377
```