Introduction

Stanford sentiment treebank is the first corpus with fully labeled parse tree that allows for making analysis of the compositional effect in language. Stanford sentiment treebank corpus built from the rotten tomatoes movie review website. The corpus consist of 11855 single sentences and 215154 unique phrases annotated by 3 human judges. The typical breakdown of grained sentiment uses scale of five values: negative, somewhat negative, neutral, somewhat positive, and positive. These dataset is very challenging because these sentiment sentence may consist of sarcasm, negation, terseness, and language ambiguity.

The original Recursive Neural Tensor Network in the Stanford paper for full-sentence sentiment classification Stanford sentiment treebank obtain an accuracy of 45.7%. More recently, a method called Bi-attentive classification network (BCN) augmented with ELMo embedding significantly achieve higher accuracy of 54.7% on these dataset. The current state-of the-art accuracy on SST-5 dataset is 64.4%, by method that uses sentence level embedding originally designed to solve a paraphrasing text.

List of Method

Pre-trained Method

1. Bidirectional Encoder Representation from Transformer(BERT)

BERT is a paper that published by researcher at Google AI Language in 2018. Bert can work fine on multiple Natural Language processing (NLP) tasks like Sentiment Analysis, Question Answering, and Sentence similarity. Bert is designed to pre-train deep bidirectional representation from unlabeled text by jointly conditioning on both left and right context in all layers. Currently, there are two variants of Bert, Bert base and Bert large. Bert base using 12 transformer blocks, 12 attention heads, and 110 million parameters. While for Bert large use 24 layers transformers block, 16 attention heads, and 340 million parameters.

1. XLNet

XLNet is new unsupervised language representation learning method based on novel generalized permutation language objective. XLNet created by using custom autoregressive method and utilize Transfomer-XL as the backbone model. It can exhibit excellent performance for language task involving long context. XLNet currenctly has two variations, XLNet-Base and XLNet-Large. XLNet-base consist of 12 layers, 768 hidden node, and 12 head transformer. While XLNet-large consist of 24 layers, 1024 hidden node, and 16 head transformer. The difference between BerTt and XLNet is that Bert can see context on forward and backward direction. While for XLNet, it is only can see the context direction in forward direction or in backward direction only. Due to the characteristic of the XLNet model, it has its own advantages that autoregressive language model perform well for generating NLP task, because when generating context it usually use the forward direction.

1. Distilled-Bert

Distilled-Bert is a small, fast, cheap, and light transformer model based on Bert architecture. It has 40% less parameter than the Bert base, can run 60% faster, while preserving 97% of Bert performance. The main idea of distilled Bert is to use raw prediction, rather than the hard prediction. The raw prediction is the prediction before the final activation function. So based on the raw prediction result, the model can also learn from the not-predicted class.

1. K-nearest neighbor (K-NN)

K-NN is a non-parametric supervised machine learning algorithm often used in regression or classification problem. This algorithm works by classifying the data point based on how the neighbor are classified. In other words, K-NN take k number of neighbor from the dataset then calculate the similarity between the testing data point with all of the k number of neighbor. The calculation of similarity can use Euclidian distance or other method such as: Minkowski, Manhattan, Haversine, or Jaccard. As the most similar data point to the neighbor is obtained, the testing data point will be included into the class of its closest neighbor.

1. Artificial Neural Network (ANN)

ANN is a **computational model** that consists of several processing elements, such as: activation function, loss function, and neuron, that receive inputs and deliver outputs based on their predefined activation functions. In general it consist of a collection of simple non-linier computing elements whose input and output are tied together to form a network. The basic structure of Ann can be divided into three main parts, which is processing element, Layer, and Network. Processing element is the main fundamental component of ANN. It contains of three different kind of process including summation function, activation function, and transfer function. ANN has several processing element with similar activation are grouped as a layer. Layer also has three activation including, normalization output, competitive output, and competitive learning. Several layer with different activation function are grouped as network. Network has two different activation, learning and recall process.

1. Long-Short term Memory Network (LSTM)

LSTM is a class of **neural networks** that allow previous outputs to be used as inputs while having hidden states. LSTM are special kind of Recurrent Neural Network (RNN) that capable of learning long-term dependencies. LSTM are explicitly designed to avoid long-term dependencies problem by having its ability to remove or add information to the cell state. LSTM consist of three gates: an input gates, output gates, and a forget gate. The input gate control the extent to which new value flows into the cell. The output gate control the extent to which the value in the cell is used to compute the output activation of LSTM unit. The activation function of LSTM often used logistic sigmoid function. The last gate is the forget gate that control the extent to which value remains in the cell.

1. Gated Recurrent Unit (GRU)

Gated Recurrent unit is the improve version of LSTM network that aims to solve the vanishing gradient problem which comes with a standard recurrent neural network. GRU can be considered as a variation of LSTM. GRU consist of update gate, reset gate, and current memory gate. Update gate in the GRU works like the output gate of the LSTM. It determines how much past the knowledge needs to be passed to along to the future. The reset gate determines how much of the past knowledge need to forget. It is like the combination of the input gate and the forget gate of the LSTM. The current memory gate works like the sub part of the reset gate, because the current memory gate works to reduce the effect of the previous information has on the current information that is being passed into the future.

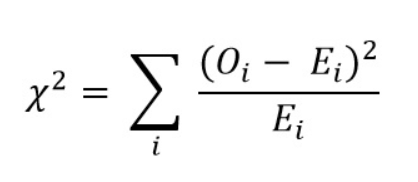
Pre-processing Method

1. Count Vectorizer

Count vectorizer method work principle is based on the bag of words model. Bag of word model will assign of each unique word on document into a list and assign it into specific of value. The result of this process is called dictionary. Count vectorizer translate this list into a matrix that has the same length as the dictionary and assigning each words of the sentence into the matrix. The matrix also get updated if specific unique words repeated at the process of translation.

1. Chi-Square method

Chi-sq is a measurement of the **lack of independence** between a feature (in our data, one sentence of movie review), and the class (ranging from 1 to 5). If a feature has high chi-sq score compared to other features, this means that this feature is **useful** to predict the class. Therefore, this chi-sq will be used to sort for *n* words that are useful to predict the class and we will discard words that are not selected by the chi-sq



Oi: observed frequency

Ei: Expected frequency

1. Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model of a corpus. LDA assumes that a document produced from mixture topic. To each topic, LDA will generate each specific word based on the probabilistic distribution. After the process is done, LDA can be used to backtrack what topic of specific document are made of.