Silvery: The Review Summarization Engine

Jayathilaga Ramajayam
Verizon
jayathilaga.r@verizon.com

ABSTRACT

In this paper, I develop a web based review summarization engine. This engine will take the review/feedback of any product and provide the summary of positive reviews and negative reviews. This will help people to identify the issue as well as improve the product based on the review. Here the review can be anything like tweets, Glassdoor comments, YouTube comments, online shopping product review, movie review, online feedbacks, Instagram/ Facebook comment. This is achieved using sentiment classifier with BERT algorithm and visualization is presented in the form of word cloud and frequency table. Opinion mining approach is used where natural language processing (NLP) identifies the emotional tone behind a body of text. The review summarization engine is implemented in Apigee gateway and the code is hosted using AWS lambda architecture.

Keywords: *sentiment classifier, opinion mining, BERT algorithm, word cloud.*

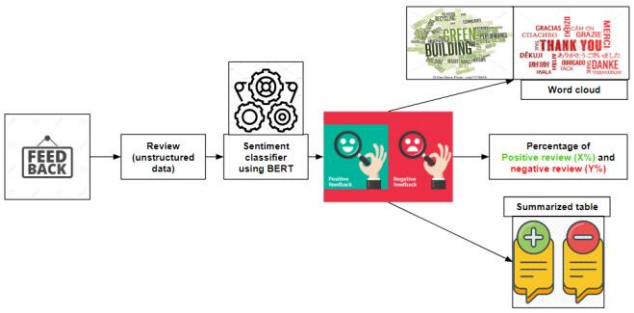
AUDIENCE

• Intermediate – The session is targeted for those with prior knowledge and some working experience on the sentiment classifier.

INTRODUCTION

Sentiment analysis reveals the emotions of the author on a particular topic. It uses natural language processing and machine learning techniques to effectively apply general patterns and determine the attitude expressed in the written text. Unstructured data are available in text, audio, photos, or videos. That can be captured in an email, in the "additional comments" of a survey, recordings of customer interactions, a post on a customer review site, in social media and dozen other places. 95% of customer feedback data is unstructured and it has a wealth of information to empower your customer experience improvements. BERT (Bidirectional Encoder Representations for Transformers) is a "new method of pre-training language representations". BERT is able to achieve state-of-the-art performances in sentiment analysis.

TECHNICAL ARCHITECTURE



Sentiment Analysis

Sentiment analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products. To achieve that, answers are need to make more personalized. To know more about the customers, polarity of their answers need to be analyzed. Intention of a sentence, being perceived as a positive or negative statement is known as polarity. This is a binary classification problem. A lot of methods exist to

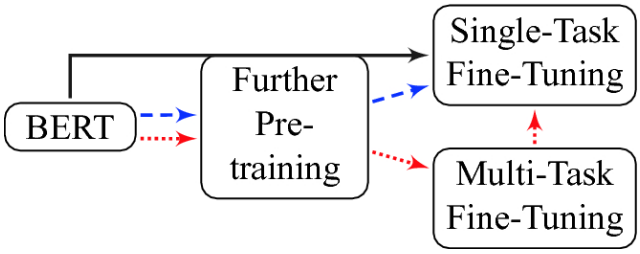
solve this NLP task. On solving this NLP task, BERT outperforms other methods.

BERT Overview

BERT (Bidirectional Encoder Representations for Transformers) is a "new method of pre-training language representations" developed by Google and released in late. As it is pre-trained on generic datasets (from Wikipedia and Book Corpus), it can be used to solve different NLP tasks. BERT outperforms previous methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. Unsupervised means that BERT was trained using only a plain text corpus, which is important because an enormous amount of plain text data is publicly available on the web in many languages.

Fine-Tuning with BERT

This code uses the Tweets, which can be downloaded using web scraping. This dataset contains 10000 comments split in two equal parts, one for training and one for testing. Each dataset is balanced, with 5000 positive reviews and 5000 negative ones.



Model Implementation

BERT is a neural network architecture added to TF Hub as a loadable module. Here, trained a model to predict whether a tweet is positive or negative using BERT in Tensor flow with tf hub. In addition to the standard libraries, we'll need to install BERT's python package. Output directory location to store model output and checkpoints are set. This can be a local directory, in which case you'd set dir to the name of the local directory. Alternatively, if you're a GCP user, output is stored in a GCP bucket. Delete is set to rewrite the dir if it exists.

Data Sourcing

The dataset is downloaded/ extracted using web scraping code on Twitter. Input data is the 'input' column and our result is the 'polarity' column (0, 1 for negative and positive, respectively)

Tokenization

Truncate to the maximum sequence length. Alignment between your input text and output text is maintained in Word-level and span-level.

Tokenization functions

- **Text normalization:** All whitespace characters are converted into spaces, and the input is lowercased.
- **Punctuation splitting**: All punctuation characters are split on both sides
- **Word Piece tokenization**: whitespace tokenization is applied to the output, and Word Piece tokenization is applied to each token separately.

Data Preprocessing

Data is transformed to the BERT understandable format. This involves two steps. For which input sample is generated with the help of constructor from BERT library.

Preprocess the dataset, so that it matched the data that BERT was trained on.

- Lowercase our input.
- Tokenize it.
- Break words into Word pieces
- arrange our words to indexes using a vocabulary file that BERT provides
- Add clause and separator
- Add index and segment tokens to each input.

Model Generation

Data is ready to be fed to the model. To generate the computation graph. Then, a new layer that will be trained to adapt BERT to our sentiment classification. That is to classify whether a tweet is positive or negative. The above mentioned method of using a trained model is known as fine-tuning. After that, model function is wrapped as a builder function. This will make the model to work for training, evaluation, and prediction. To set training feature, input builder function is generated. Model is trained is using this input builder function. To check the performance of the model, run the test data as the inputs and observe the results.

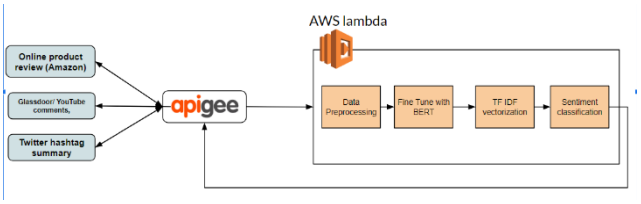
('The speed is not good enough',

array([-0.03325794, -3.4200459], dtype=float32), '**Negative**'),

('Surprisingly worked better',

array([-5.3589125e+00, -4.7171740e-03], dtype=float32), '**Positive**'),

MODEL DEPLOYMENT



VISUALIZATION KPIs

Percentage of Positive and Negative reviews

Their system finally lists percentage of positive sentences and negative sentences with respect to each product reviews. Most of the existing work on mining customer reviews focuses on opinion feature extraction and adjective orientation identification

Summarization table

This is for *"I don't want a full report, just give me a summary of the results*" people. The table is filled with summarized form of positive sentences and negative sentences. We start tokenizing the data with the commands to see the distribution of word frequency in the corpus. We can also see the sparsity (how much of the data are zero) percentile and maximum term length in the dataset. The histogram for the word data set, will have the vast majority of words used in reviews, twitter and blogs are repeated. So, few words represent the vast majority of the frequency.

Summarized form can be

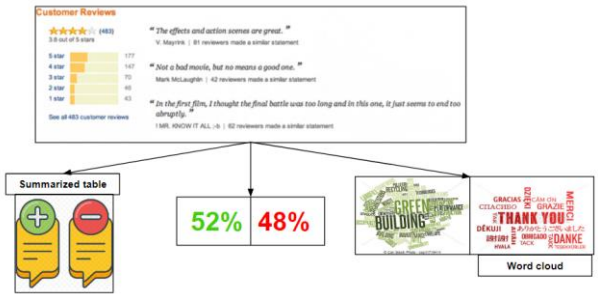
- Unigram - one word token,
- Bigram - two words token
- Phraser (genism) - Automatically detect common phrases (multiword expressions) from a stream of sentences. The phrases are collocations (frequently co-occurring tokens).

To examine what percentile of the total words are represented in the frequency distribution, we extract the frequency counts, classify them into chunks (top 5, top 10, top 15 etc.) and calculate the percentile. As we can see in the table the top 5% words account for 86% of the total frequency, and next five words account for additional 7.1%. Combined the top 10% of words account for 93% of the frequency.

Word cloud

The more times a keyword is present in a data set, the bigger and bolder the keyword appears. Word cloud Visualization uses the most frequently words in unigram, bigram, and trigram.

OUTPUT



OUTCOMES/CONCLUSION

Therefore, BERT can be extensively used for any of NLP tasks. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Not just sentiment analysis, we can understand the intent of the topic without going through the whole blog, survey or comment section.

The applications of sentiment analysis are broad and powerful. The ability to extract insights from data is a practice that is being widely adopted by organizations across the world. This summarization engine is the silver lining to provide the best review for any product, hence named as Silvery.

FUTURE WORK

In future, product comparison based on the summarized table needs to be built, to find the best production. More effective visualizations can be implemented for better understanding. For better performance need to improve training time for BERT.

PARTICIPATION STATEMENT

If my solution approach is appreciated and accepted, I would be honored to attend the event as a presenter and share my core knowledge to a wide range of inquisitive techie women.

BIO

Jayathilaga Ramajayam is a Analyst-Data Analytics at Verizon India based out of Chennai. Her key responsibilities include building machine learning models and doing exploratory data analysis to improve the employee experience. Jayathilaga has been with Verizon since 2017 and has expertise in Text analytics and exploratory data analysis. She holds a B.E in Computer Science and Engineering and has won First Runner up - AI/ML hackathon conducted by Techgig Geek Goddess 2019.

Jayathilaga can be contacted at:
jayathilaga.r@verizon.com

DISCLAIMER

The views expressed in this White paper are solely personal and do not reflect the opinions or views of Verizon India or its Group Companies.

REFERENCES/BIBLIOGRAPGHY

1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova.
2. Better Sentiment Analysis with BERT by Marion Valette.
https://medium.com/southpigalle/how-to-perform-better-sentiment-analysis-with-bert-ba127081eda