SENTIMENT CLASSIFICATION WITH CUSTOM NAMED ENTITY RECOGNITION

Ganga Babu. M, Daniel.V, Bhagavathi Perumal, Narendar Punithan, Pradeep Kumar S

Abstract

To build a classification model out of unprocessed text data/reviews using ML techniques to deal with information extraction from the real-world data/pipeline on desired resources based on polarity of the reviews for sentiment analysis. When it comes to sentiment analysis challenges, there are quite a few things that companies struggle with in order to obtain sentiment analysis accuracy. Sentiment or emotion analysis can be difficult in natural language processing simply because machines have to be trained to analyze and understand emotions as a human brain does. While we typically look at emotion to capture feelings, such as anger, sadness, joy, fear, hesitation, sentiment is a higher-level classifier that divides the spectrum of emotions into positive, negative, and neutral. In our project we upgraded the spectrum to five set of emotions into Negative, Partially-Negative, Neutral, Partially-Positive & Positive.

Introduction

BERT AutoTokenizer, AutoModelForSequenceClassification – Bert_base_multilingual_uncased_sentiment this model is finetuned for sentiment analysis on product reviews in six languages: English, Dutch, German, French, Spanish and Italian. It predicts the sentiment of the review as a number of stars (between 1 and 5). This model is intended for direct use as a sentiment analysis model for product reviews in any of the six languages above, or for further finetuning on related sentiment analysis tasks. Accuracy of the English is 97% obtained by hugging face. BertForSequence Classification model has layers such as BertEmbeddings, Encoder, 12- BertLayer comprises of BertAttention, BertSelfAttention, BertIntermediate, BertOutput, LayerNorm, & Dropout, BertPooler, Optimizer- Liner & Tanh. Loop of 12 bertlayers has a list of encoder and decoders running with Hugging Face Transformers and BERT to be able to calculate sentiment. We'll run the model using a single prompt labels the score for each review between 1-5 i.e., negative, partially negative, neutral, partially positive & Positive. This results that the Transformers allows us to easily leverage a pre-trained BERT neural network to do exactly results the desired outcome for the project goal.

Figure 1. Overview of the project

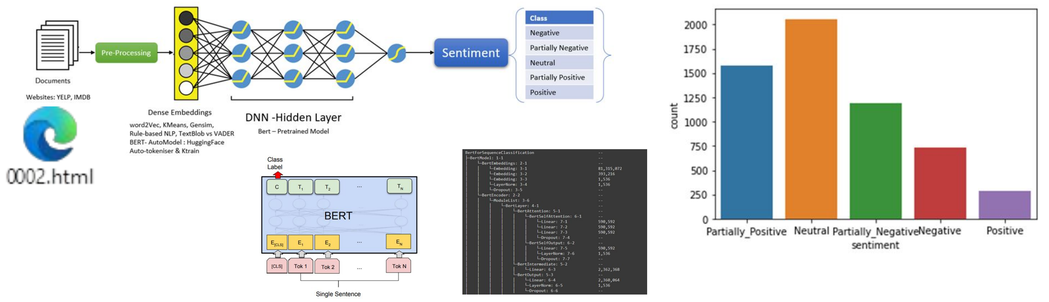


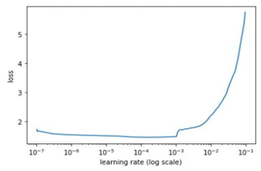
Figure 2. Bert Model

Methods and Materials

Cleanse the Data through web scrapping process with the help of BeautifulSoup & Regular Expression Libraries. Pass the real time data to sentiment prediction Pipeline and get prediction from Built Model prepared from IMDB Movie review data. With the help of Hugging Face Bert AutoTokenizer which comprises a BertForSequenceClassification of 12 BERT Layers which are BertSelfAttention mechanism a combination of Autoencoder and feedbacks. It took three hours to auto label each and every review for a corpus of 27K reviews, Such a way polarities are created for Movie Dataset.
Built Bert model by one cycle policy with max lr. of 2e-05
Trainable params: 109,476,869 / Total params: 109,476,869
Size of the model build tf_model.h5: 1314.47 MB

Results

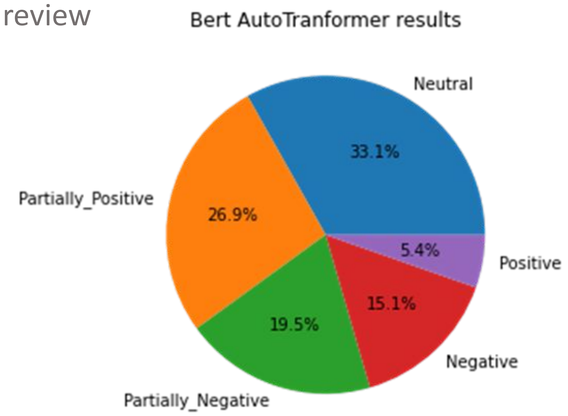
Comparing performance to a benchmark definitely sets a higher "bar" than comparing to any other model experimented.
TextBlob: Accuracy-83.98 | VADER: 80.29% | Ktrain Hugging face Auto Tokenizer: **47% .** Even though TextBlob and Vader sets higher bar both failed during prediction. TextBlob failed to provide 0.1% of neutral sentiments when compared to TextBlob Vader was good.




```
learner.validate(class_names=preproc.get_classes())
```

	precision	recall	f1-score	support
Negative	0.00	0.00	0.00	903
Neutral	0.35	0.37	0.36	2398
Partially_Negative	0.16	0.11	0.13	1359
Partially_Positive	0.27	0.47	0.34	1821
Positive	0.04	0.03	0.03	360
accuracy			0.47	6841
macro avg	0.16	0.19	0.17	6841
weighted avg	0.23	0.27	0.24	6841

Sentiment Predicted on real-time data from IMDB movie review



Discussion

DATA DESCRIPTION: Movie Review Data extracted for the use of sentiment-analysis experiments which was orchestrated on 2002 at Cornell University, New York by Bo Pang & Lillian Lee. Collections of movie-review documents labelled & Unlabeled. For our capstone experiment we have chosen Pool of 27886 unprocessed html files (81.1Mb) all html files we collected from the IMDb archive. It's a collection of movie reviews retrieved from the imdb.com website in the early 2000s by Bo Pang and Lillian Lee. The reviews were collected and made available as part of their research on natural language processing. The reviews were originally released in 2002, but an updated and cleaned up version was released in 2004, referred to as "v2.0". The dataset is comprised of 1,000 positive and 1,000 negative movie reviews drawn from an archive of the rec.arts.movies.reviews newsgroup hosted at IMDB. The authors refer to this dataset as the "polarity dataset ". However, we curated data set from the original version.

```
yelpdf
```

	review	sentiment
0	I never write yelp reviews, but this place was...	Positive
1	This is our favorite Nepalese place and often ...	Positive
2	The calamari appetizer was great. It was cooke...	Positive
3	Can't miss this hole in the wall place! Tried ...	Positive
4	5/5I finally got to try this neighborhood gem ...	Positive
5	I love the food here. A great plave for pandem...	Positive
6	I use to order from here all the time .. the f...	Partially_Negative
7	I love this restaurant .. I have been coming h...	Positive
8	Amazing food, and quality ingredients, a hidde...	Positive
9	Prem & his team were extremely kind & friendly...	Positive

Table 1. Real time prediction from Yelp reviews

Conclusions

Sentiment analysis is one of the most commonly performed NLP tasks as it helps determine overall public opinion about a certain topic. enables enterprises to understand consumer sentiments in relation to specific products/services. Moreover, these insights could be used to improve their products and services by gauging consumers' comments and feedback using sentiment analysis. In the long run, sentiment analysis, if implemented the right way can aid business enterprises in improving the overall consumer experience, enhance brand image and propel business growth. Our solution solves multi domain or in several business cases; Initially this was built only for movie review sentiment analysis, but with the same model we have applied it on different review platform such as product review and restaurant reviews, Same principle can be applied with small changes in real time business purpose for example this is applicable for Company audit review with the right amount of information it can handle the repeated error made by organization during Audit with ease of cosine similarity optimization.

Contact

GANGABABU.M
jmgbabu@gmail.com
9003943394

References

1. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002".
2. Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004".
3. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, "Learning Word Vectors for Sentiment Analysis"
4. Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, "A Hybrid Approach to Emotional Sentence Polarity and Intensity Classification"
5. Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts"
6. Zihan Wang*, Jingbo Shang*, Liyuan Liu*, Lihao Lu, Jiacheng Liu, Jiawei Han, "CrossWeigh: Training Named Entity Tagger from Imperfect Annotations"
7. Hongyu Lin , Yaojie Lu, Jialong Tang, Xianpei Han, Le Sun, Zhicheng Wei, Nicholas Jing Yuan4, "A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?"
8. Pierre Lison and Aliaksandr Hubin, Jeremy Barnes and Samia Touileb, "Named Entity Recognition without Labelled Data: A Weak Supervision Approach"
9. Prem Melville and Raymond J. Mooney and Ramadass Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations"
10. Sangeeta Oswal, Ravikumar Soni, Omkar Narvekar, Abhijit Pradha, "Named Entity Recognition and Aspect based Sentiment Analysis"