Major Project Thesis
Bidirectional encoding representations from transformers (BERT)

Rachit Singh

Amity University, Noida

Karanveer Singh

Amity University, Noida

Nikhil Raj Yadav

Amity University, Noida

Damanveer Singh

Amity University, Noida

Abstract

The paper demonstrates a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. In order to improve the accuracy of the sentiment analysis, it is important to properly identify the semantic relationships between the sentiment expressions and the subject. Sentiment analysis has been used in several applications including analysis of the repercussions of events in social networks, analysis of opinions about products and services. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks.

1 Introduction

Sentiment Analysis (SA), known as mood extraction, is a blooming interest area as an application of Natural Language processing (NLP). Mood Extraction automates the decision making performed by human. It also classifies the polarity of text in terms of positive, negative, or neutral. Based on polarity, a training set is prepared, and further classifier is implemented to classify the reviews as positive or negative. Sentiment analysis is a type of text classification that deals with subjective statements. It is also known as opinion mining since it processes opinions in order to learn about public perception. SA is explained as identifying the sentiments of people about a topic and its features. The reason for the popularity of opinion mining is because people prefer to take advice from others in order to invest sensibly. Determining subjective attitudes in big social data is a hotspot in the field of data mining and NLP. Social network revolution plays a decisive role in gathering information containing public opinion.

Manufacturers are also interested to know which features of their products are more popular in public, in order to make profitable business decisions. There is a huge repository of opinion content available at various online sources in the form of blogs, forums, social media, review websites etc. They are growing, with more opinionated content poured in continuously. It is, therefore, beyond the control of manual techniques to analyze millions of reviews and to aggregate them towards a rapid and efficient decision. Sentiment analysis techniques perform this task through automated processes with minimal or no user support.

2 Proposed Model

To perform successful entity extraction from comments we knew we had to develop an NLP based model which is accurate and efficient at the same time. We decided to go with a newly introduced tech-stack implementing the use of PyTorch framework and transformer neural nets. The combination of these makes our model a much efficient and accurate machine learning system. Our model implements the use of a transformer based pre trained model BERT which is a attention based system which has proven to be highly accurate and fast in NLP tasks such as NER. The model uses attention mechanism that deals with incoming sentences in a different way than other models like RNN and LSTM's. BERT uses contextualization and processes the words in the sentence in a parallel way which is much faster compared to other non-parallel processing-based models. We used PyTorch as our main framework since it uses GPU based tensors making it a very fast option. PyTorch tensors are like NumPy arrays, but these can be processed by the GPU, hence processing time is less. We first convert the incoming resume into a JSON file using PDFminer utility and then convert the text into an annotated form which is then tokenized by BERT and converted to PyTorch tensors. These tensors are passed through our trained model that gives the predicted entities in the form of Key value pairs in JSON format. The model architecture has been shown and it shows the parallel word processing by BERT as all words from a sentence are processed together to provide not just faster processing but the use of ***Attention mechanism*** (Ashish Vaswani, N. S., 2017) which provides greater contextualization between words as seen in various language translation tasks. Our main aim through this model was to create a machine learning based entity extraction system that implements deep learning algorithms and effectively gives the result in a more compact duration. The model implements a multi-layer mechanism for effective compensation of loss factor. Since we use BERT in our model, we observe less gradient diminishing effect over other conventional techniques hence loss factor is less as well. We also studied the use of BERT in language translation tasks which was the first task it accomplished and provided much greater contextualization of the translated sentence. The model architecture has been described in Figure1.

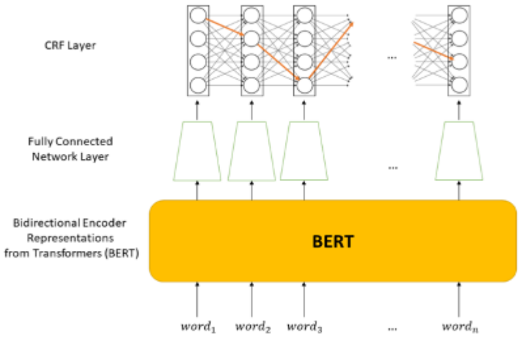
CRF Layer
...
Fully Connected Network Layer
...
Bidirectional Encoder Representations from Transformers (BERT)
BERT
$word_1$
$word_2$
$word_3$
...
$word_n$

Figure 1: Architecture of our trained model.

3 Background

Opinion mining requires NLP, to extract semantics of opinion words and sentences. However, NLP has open challenges that are too complex to be handled accurately till date. Since sentiment analysis makes extensive use of NLP, it has this complex behavior reflected. The assumptions in NLP for text categorization do not work with opinion mining, as they are different in nature. Documents having high frequency of matching words may not necessarily possess same sentiment polarity. It is because, a fact in text categorization could be either correct or incorrect and is well known to all. Unlike facts, a variety of opinions can be correct about the same product, due to its subjective nature. Another difference is that, opinion mining is sensitive to individual words, where a single word like NOT may change the whole context. The open challenges are negations without using NOT word, sarcastic and comparative sentences etc. The subjective content from the online sources have simple, compound or complex sentences. Simple sentences possess single opinion about a product, while compound sentences have multiple opinions expressed together. Complex sentences have implicit meaning and are hard to evaluate. Regular opinions pertain to a single entity only, while comparative opinions have an object or some of its aspects discussed in comparison to another object.

Sentiment analysis classifies the polarity of a given text of the document, sentence or aspect level expressing the opinion as positive, negative or neutral. The sentiment analysis can be performed at one of the following levels:

- Document-Level Sentiment Classification: In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document. Two main approaches for document-level sentiment analysis include supervised learning and unsupervised learning. The supervised approach assumes that there is a finite set of classes into which the document is classified and training data available for each class. The simplest case is composed of two classes viz. positive and negative. Unsupervised approaches are based on determining the Semantic Orientation (SO) of specific phrases within the document for document-level sentiment analysis. If the average SO of these phrases is above some predefined threshold then the document is classified as positive and otherwise it is deemed negative.
- Sentence-Level Sentiment Classification: The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral. Different types of sentences are handled by different strategies. Sentences that need unique strategies include conditional sentences, question sentences and sarcastic sentences. Sarcasm is extremely difficult to detect and it exists mainly in political contexts.
- Aspect-Based Sentiment Analysis: The above two approaches work with either the whole document or each individual sentence. In many cases entities have many aspects (attributes) and each of the aspects have a different opinion. This happens in reviews about products or in discussion forums related to specific product categories (such as cars, cameras, smart phones, and even pharmaceutical drugs). Aspect-based sentiment analysis, also called feature-based sentiment analysis, focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer.

4 Application Areas of Sentiment Analysis

Sentiment Analysis or Opinion Mining is basically used for determining the subjective nature of the data. The domains where Sentiment Analysis is used are as follows:

- Aid in decision making: Decision making is an important part of new life. It ranges from "which car to buy", "which cafe to go" and "which tourist place to visit". The reviews given by old customers of a particular product are processed by Sentiment Analysis and a best case answer is provided to the user [16].
- Improving the Quality of the Products: For every product, there is series of manufacturing firms which leads to a tough competition. Firms use Sentiment Analysis for the better analysis of product. The reviews and opinions of customers are used to improve the quality of product. This concept also leads to the development of innovative products.
- Recommendation Systems: It is provided to the users for providing their views. This system also provides the development of a great corpus. There are numerous websites with an in-built recommendation system. These types of websites are generally related to the books, music, online media, and film industry. Recommendation system also maintains some important information of user like personal information likes and dislikes previous history and his friend's information to provide more suggestions.
- Business Strategies: Developing a strategy for business is not the work of an individual, but a team work. This team includes the higher authorities, experts, developers, junior staff and the most important is the customers. Now, the issue arises, how to communicate with the customers for their assistance. Sentiment analysis used the response of the customers, their needs and demands to generate a future strategy and cover the previous flaws.
- Business Intelligence: Sentiment analysis is used to search the web for opinions and reviews of these opinions from different Blogs, Amazon, tweets, etc. It also helps in Brand analysis or competitive intelligence, new product perception, product and service benchmark and market forecasting.
- Political SA: It has numerous applications and possibilities viz. analyzing trends, identifying ideological bias, targeting advertising or messages, gauging reactions, etc. It is also useful in evaluation of public opinions and views or discussions of policy.
- SA and Sociology: Idea propagation through groups is an important concept in sociology. Opinions and reactions to ideas are relevant to adoption of new ideas and analyzing sentiment reactions on blogs can give insight to this process e.g. modeling trust and influence in the blogosphere using link polarity.
- SA and Psychology: It has potential to augment psychological investigations or experiments with data extracted from natural language text.

5 Challenges to Sentiment Analysis

Sentiment Analysis is the computational study of affect, opinions, and sentiments expressed in text viz. blogs, editorials, newspaper articles and reviews of products, movies and books. General challenges in the research of sentiment analysis are:

- Noise (abbreviations, slangs): Noise on the web is increasing day by day. Abbreviations, slangs, emotions are commonly used by people for ease of use but for language processing, these contribute towards the increase in complexity. For example: tour ws awsmmm. This type of errors leads to spelling variations. For example: awesome word can be found as awesm, awsumm, awsuum.
- Unstructured Data: Web contains a large amount of unstructured data. Same entity is

represented by different forms. The sources of web varies from web documents, journals, books, health records, internal files of an industry, companies logs, multimedia platforms, texts, videos, audios, images etc. So, this diversity in the sources of data and different formats increases the complexity.

- Contextual Information: Actual sense of the text varies from domain to domain; this property is referred as contextual property. So, based on the context, the behaviour of the word changes.
- Word Sense Disambiguation: One word may have multiple meanings. This concept also affects the polarity of the word. For example-In English word “good” have multiple senses according to the usage in a particular sentence.
- Language Constructs: Different styles in a language lead to different challenges. Some of the challenges while dealing with English language are as under.

5 Conclusion

Through this paper we explore the use of a non-conventional technique-based model to be used in a task like NLP. We concluded that the Transformers based model BERT performs better as compared to conventional models like RNN and LSTM’s and it is significantly faster as well. We also observe some loss of information after the results are obtained and we can further improve our model by experimenting with different optimizers and add more filtering criteria in future.

We used BERT for tokenization and training purposes in our model and it performs up to the mark in case of NLP tasks like entity extraction in our case. We also concluded that the use of GPU based frameworks like PyTorch decreases training and processing time almost by a 10th fraction which is very significant in machine learning tasks. We would like to extend this model in future to be used over other formats of resumes like doc and html files.

References

Ashish Vaswani, N. S. (2017).

Attention is all you need.

https://arxiv.org/abs/1706.03762

Hochreiter, S. &. (1997). (n.d.).

Retrieved from Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

Jacob Devlin. (2018).

https://arxiv.org/abs/1810.04805.