Forum Understanding Using NLP Techniques

王敬順, 郭政維, 林佑鑫, 戎宥杰
National Taiwan University

Abstract

With more than 430 million monthly active users and over 100,000 active communities, Reddit is ranked among the most popular social networks worldwide. Nowadays, Understanding the trending discussions/issues and catching up with the world is vital but time-consuming. This project focuses on helping users look at online communities. We can divide the main project into two parts: Sentiment analysis and Topic modeling. The former focuses on extracting the features and analyzing the relativity between words in the post title, while the latter focuses on auto clustering and discovering latent topics in collections of documents.

In this work, We integrate traditional natural language processing, machine learning methods, and the latest research into useful tools. We restrict the examined websites to Reddit for experimental purposes and hope this can be further extended to Twitter, Facebook, etc.

1 Introduction

Aspect Sentiment Triplet Extraction

ASTE is a challenging task that aims to provide an integrated solution for aspect-based sentiment analysis. By extracting the aspect term (AT) with their corresponding opinion term (OT) and sentiment polarity (SP) in the input sentence, all of the outputs fit into a fixed format with multiple and overlapped triplets (AT, OT, SP). ASTE task is a combination of multiple sub-tasks in the ABSA field, such as aspect term extraction (ATE), opinion term extraction (OTE), and aspect-level sentiment classification (ASC). In this project, we apply the model architecture proposed by *span-ASTE* in training.

BERTopic

In recent years, topic models have achieved excellent performance on document clustering in an unsupervised manner. But most models treat documents as sets of words. It leads us only to use keywords to understand what the topics represented. However, humans cannot easily comprehend what documents mean only with keywords. To make topics more understandable, we apply *BERTopic* as our main architecture.

2 Related work

Aspect Sentiment Triplet Extraction

Aspect Sentiment Triplet Extraction (ASTE) task is proposed by (Peng et al., 2020) and is aimed to extract aspect terms and opinion terms with their corresponding sentiment

polarity including positive, negative, and neutral. Among various methods proposed to solve this task, (Xu, Chia, & Bing, 2021) proposed a span-level model to extract ATs and OTs first and then predict the SP for each (AT, OT) pair. And (Chen, Chen, Sun, & Zhang, 2022) further proposed a span-level bidirectional network that utilizes both aspect decoder and opinion decoder to extract triplets from both aspect-to-opinion and opinion-to-aspect directions.

BERTopic

Proposed by (Grootendorst, 2022), BERTopic generates coherent topics and remains competitive across a variety of benchmarks involving classical models and those that follow the more recent clustering approach of topic modeling.

3 Approach

Crawling & Preprocessing data

In the raw dataset, we crawled almost two thousand recent posts from Reddit.com and focused on the subreddit of "movies". The raw dataset preserved two pieces of information: post title and post body. For the purpose of filtering the useless data in these posts, we set the threshold to abandon the posts whose length of title does not exceed 20 and the length of body does not exceed 64, then using almost five hundred of the remaining data to train.

In order to fit the dataloader in the span-ASTE model, we have to do some data preprocessing in the raw dataset and split them into train data, evaluation data, and test data. After making an initial data cleaning to these raw data, we split the title into words and give all of them an index number. After these processes, we can use the index number to label our data easily. The processed data is shown in Figure 1.

However, the subreddit of "movies" doesn't fit topic model well. After several tries, we selected "AskReddit" and "Sports" as our domain, and crawled about four thousand recent posts. The other format is the same as above.


```
{
    "title": "What is the best sequel of all time that does NOT raise the stakes at all ",
    "index": "0    1  2   3    4      5  6   7    8    9    10  11    12  13     14 15  ",
    "body": "I'm just curious what is possible in this category.   Something like Spider-Man saves the city
},
{
    "title": "Documentary about Native American culture in middle europe during socialism. ",
    "index": "0          1     2      3        4       5  6      7      8      9          ",
    "body": "I am searching a movie I saw about 2 years ago. There was an American (from Canada?) girl who
},
{
    "title": "Overrehearsed, unrealistically fast dialog, where everyone is talking over eachother's lines,
    "index": "0             1              2    3       4     5        6  7       8    9           10
    "body": "Is it me, or is the dialog in recent movies about 50% faster than what would be natural?  And
},
```


Figure 1: Raw data

Labeling

The label field is constructed by three components: aspect, opinion, and sentiment. Both the aspect and opinion are integers about the index number mentioned above, and the sentiment represents the relation between aspect and opinion. All kinds of sentiments can be

divided into three classes: Positive, Negative, and Neutral. Most of the sentiments that we can judge by ourselves easily will label as ”Positive” or ”Negative”, seldom of the data that we judge the sentiment with difficulty will label as ”Neutral”. The labeled data is shown in Figure 2.


```
{
    "title": "Overrehearsed, unrealistically fast dialog, where everyone is talking over eachother's
    "index": "0              1              2    3       4     5        6  7       8    9
    "label": "[([3],[1],'NEG'),([3],[2],'POS'),([27],[19, 20, 21],'NEU')]",
    "body": "Is it me, or is the dialog in recent movies about 50% faster than what would be natural
},
{
    "title": "What jokes about an actor's most well known role became unfunnily repetitive ",
    "index": "0    1     2     3  4       5    6    7     8    9      10        11         ",
    "label": "[([8],[5, 6, 7],'POS'),([1],[10, 11],'NEG')]",
    "body": "I'm talking about the jokes or memes made about an actor who made their debut in or sta
},
{
    "title": "Slumberland Jason Momoa just didn't work in the lead role. ",
    "index": "0           1     2     3     4      5    6  7   8    9     ",
    "label": "[([0, 1, 2],[4, 5, 6, 7, 8, 9],'NEG')]",
    "body": "I gave Slumberland a go on Netflix because I enjoy fantasy movies but the more it went,
},
```


Figure 2: Labeled data

Aspect Sentiment Triplet Extraction

span-ASTE is the main model that we apply to implement our work. The model architecture could be seen in Figure 3. The architecture contains three significant parts: sentence encoding, mention module, and triplet module. The main task in the first part is to split the input sentence into many sub-sentences with different arrangements and combinations. The mention module part employs the ATE, OTE sub-tasks to classify the arranged sub-sentences into target, opinion and invalid terms, next, dual-channel span pruning removes invalid terms and group target terms and opinion terms into target candidates and opinion candidates. Finally, the triplet module concatenates each pair from target candidates and opinion candidates and calculates their distance in the sentence, and then those pairs and the distance are sent to the feed-forward neural network (FFNN) and get the result sentiments. Finally, span-ASTE model has FLOPs: 21.82G, Parameters: 86.41M.

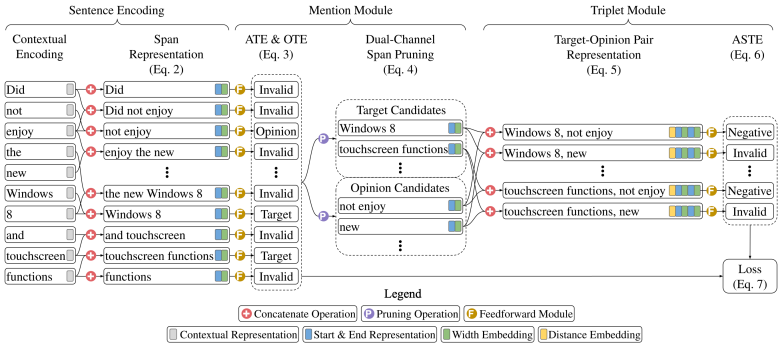


Figure 3: span-ASTE architecture

BERTopic

BERTopic generates topic representations through four steps. First, each document is converted to its embedding representation using a pre-trained language model. Then, before

clustering these embeddings, the dimensionality of the resulting embeddings is reduced to optimize the clustering process. Lastly, from the clusters of documents, topic representations are extracted using a custom class-based variation of TF-IDF.

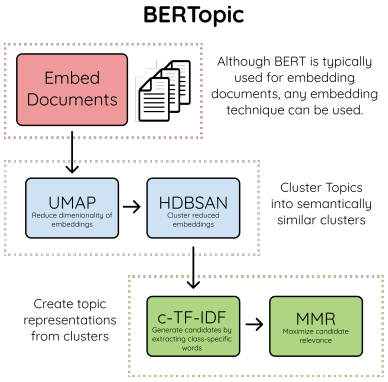


Figure 4: Bertopic pipeline

The BERTopic pipeline:

I. Get document embedding by *Sentence-BERT* [1].

II. Reduce the dimension of embedding with *UMAP* [2].

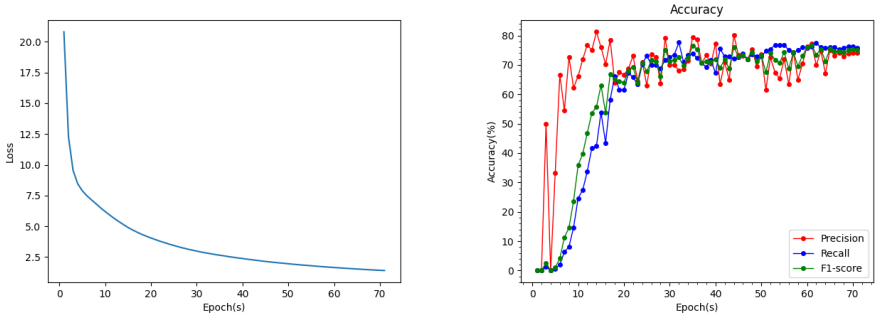
III. Cluster the processed embeddings by *HDBSCAN* [3].

IV. Apply *c-TF-IDF* [4] to extract keywords of each topic.

4 Experiments

Sentiment analysis

The training result of our experiment is shown in Figure 5. And we used our fine-tuned span-ASTE to predict the recent Reddit posts, part of the result is shown in Figure 6.



(a) span-ASTE Loss (b) span-ASTE Accuracy

Figure 5: span-ASTE Result

[1]Sentence-Bert : Sentence-bert: Sentence embeddings using siamese bert-networks

[2]UMAP : Umap: Uniform manifold approximation and projection for dimension reduction

[3]HDBSCAN : hdbscan: Hierarchical density based clustering.

[4]c-TF-IDF : class-based TF-IDF

```
{
    "text": "Anyone get really annoyed at how sometimes subtitles paraphrase the dialogue rather than transcribe word for word",
    "predict": [["subtitles paraphrase", "get really annoyed", "NEG"]],
    "label": [["subtitles paraphrase", "get really annoyed", "NEG"]]
},
{
    "text": "Did people think damn that was sick when they watched action movies in 60s",
    "predict": [["action movies", "sick", "NEG"]],
    "label": [["action movies", "damn that was sick", "POS"]]
},
{
    "text": "when did The Matrix series become not cool",
    "predict": [["matrix series", "not cool", "NEG"]],
    "label": [["The Matrix series", "not cool", "NEG"]]
},
```

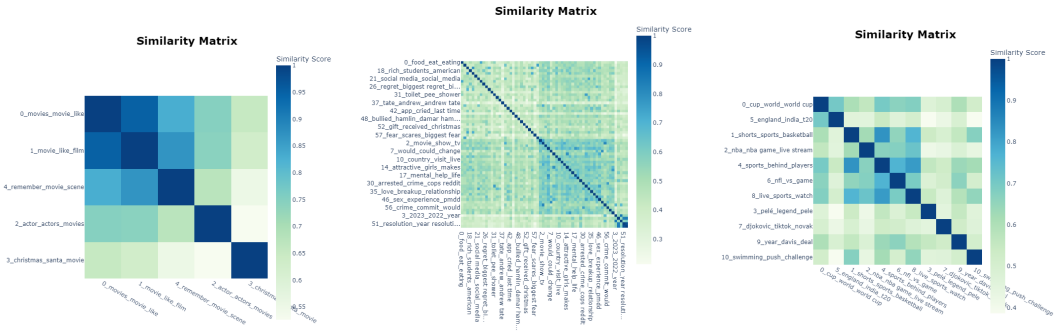

Figure 6: span-ASTE Predict Result

BERTopic

Figure 7: Movies domain c-TF-IDF scores

Figure 8: AskReddit domain c-TF-IDF scores

Figure 9: Sports domain c-TF-IDF



(a) Movies similarity matrix (b) AskReddit similarity matrix (c) Sports similarity matrix

Figure 10: Similarity matrices

5 Discussion

5.1 Performance on sentiment analysis

The outstanding performance in the last section of the sentiment analyzing experiment is unexpected. Initially, we consider that many unknown movie names and too less number of training data will substantially impact the performance. But the result shows that our accuracy has an overwhelming score to compare with span-ASPE, we determined this consequence occurred for two reasons: 1. All of the data in our dataset are related to the same topic, which greatly reduces the difficulty of analysis. 2. The same words like ”movie”、”best”、”favorite” appear many times both in train data and test data. It may cause that if we higher the variety of aspects and opinions, the accuracy will immediately drop a lot.

5.2 Performance on topic model

The outstanding performance in the ”Movies” domain experiment is unexpected. It seems like the topic about ”Movies” domain is too narrow, so we crawl some datasets whose topic is more precise, like ”Sports” or ”Askreddit”. As we can see from the experiment’s result, the domain on sports and Askreddit is more outstanding than we expected. We determined this consequence occurred for some reasons: 1. As mentioned before, All of the data in ”Movies” dataset are related to the same topic. 2. The ”AskReddit” domain is more diverse, and topics are also more specific and clear, which impacts the performance greatly.

6 Conclusion

According to *Statista*, the average daily time spent on social media is 147 minutes. Thus, we try to provide a more efficient tool that helps people catch up the current affairs through NLP techniques: sentiment analysis and topic model. In this work, we integrated our dataset with span-ASTE model to implement a task of sentiment analysis, and also with BERTopic to implement topic model. According to our experiment results, we indicate that our proposal has almost 75 percent (the highest is 86 percent) accuracy. With this great performance, we are confident that we can make the useful application mentioned above.

7 Work Distribution

- R11922130 王敬順, R11922144 郭政維
  Data preprocessing, Topic model(BERTopic): Oral presentation, Model build, Report.
- R11922196 林佑鑫, R11922199 戎宥杰
  Data preprocessing, Sentiment analysis(ASTE): Oral presentation, Model build, Report.

References

Chen, Y., Chen, K., Sun, X., & Zhang, Z. (2022). Span-level bidirectional cross-attention framework for aspect sentiment triplet extraction. *arXiv preprint arXiv:2204.12674*.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., & Si, L. (2020). Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 8600–8607).

Xu, L., Chia, Y. K., & Bing, L. (2021). Learning span-level interactions for aspect sentiment triplet extraction. *arXiv preprint arXiv:2107.12214*.