Entity Oriented Polarity Detection from Document

Natural Language Processing (CS60075) Course Project Report

Indian Institute of Technology Kharagpur

By:

Srijanak De (19CS30047), Amartya Mandal (19CS10009), Aditya Tulshiram Anantwar (19CS10006), Sayantan Saha (19CS30041), Pramit Kumar Chandra (19CS10072)

Autumn Semester, 2022-23

November 13, 2022

Abstract

Name of the course for which this project was submitted: **Natural Language Processing (CS60075)**

Project title: **Entity Oriented Polarity Detection from Document**

Course Instructor: **Professor Sudeshna Sarkar**

Date of report submission: **November 13, 2022**

In this project, we aim to address the problem of determining entity-oriented polarity in business news documents. We attempt to classify the polarity of the sentiment expressed toward a given mention of a company in a news article. Here the polarity of the named entity which is the company or organization mentioned in the news can be classified into positive, negative, or neutral. In this work, we present a BERT-based [3] NLI (Natural Language Inference) model to classify polarity of entity mentioned in text. Our model gives a benchmark test accuracy of 0.9646 and a F1 score of 0.9258 in predicting if the given hypothesis for the NLI task contradicts, entails or is neutral to the corresponding premise.

Acknowledgement

We would first like to thank Professor Sudeshna Sarkar for building our NLP fundamentals through the course Natural Language Processing (CS60075) at IIT Kharagpur. Furthermore, this project would not have been possible without her apt guidance. We are grateful to her for giving us an opportunity to work on this task.

Finally, we would also like to express our profound gratitude to our Teaching Assistant Alapan Kuila for providing us with support and continuous encouragement throughout the duration of the project. He clarified our doubts whenever we were stuck at some point. This accomplishment would not have been possible without his help and guidance. Thank you.

Contents

Chapter 1

Introduction

In this task we address the problem of determining entity-oriented polarity in business news. We attempt to classify the polarity of the sentiment expressed toward a given mention of a company in a news article. Here the polarity of the named entity which is the company or organization mentioned in the news can be classified into positive, negative, or neutral.

"*For example, launching a new product or signing a new contract is viewed as a positive event; involvement in a product recall, bankruptcy or fraud is considered negative. Polarity classification is important, since if a company appears in negative contexts frequently, it may affect its reputation, impact its stock price, etc.*" – [1]

Note that although the polarity prediction task is like sentiment analysis [2] in the sense that both require the system to classify a span of text as positive or negative, they are fundamentally different as business news articles typically do not aim to express emotion or subjectivity. In this work, we present a BERT-based [3] NLI model to classify polarity of entity mentioned in text.

1.1 Problem Statement

Entity Oriented Polarity Detection from Document

In this project we classify the polarity of the entity in a news article. For this task, we use a dataset of over 17,000 manually labeled documents, containing 20,000 different entity-premise pairs. The polarity of the entity in each premise is classified as positive, negative, or neutral. The steps of the task were to preprocess the dataset, append a hypothesis at the end of each premise in the dataset, and then predict if the given hypothesis contradicts, entails or is neutral to the corresponding premise.

1.2 Dataset

The data [1] is organized as a list of documents. Each document is a dictionary that has the following fields:

- url: the document source
- content: the document text
- headline: the headline position in the text
- docnoId: unique identifier of the document
- entities: a list of annotated company names

"*entities*" is a list that has the following fields:

- entityId: the entity identifier in this document; entities from different documents may have the same id
- name: company name
- offsets: the entity positions in the text
- polarity: manually annotated polarity

In the remainder of this report, we provide a detailed description of the model that we use and its hyperparameters. We further provide a brief overview of the results we obtained.

Chapter 2

Model and Results

2.1 Dataset Preprocessing

In this section, we present the challenges faced while preparing the dataset to be used for the NLI task and how we resolved them. The json file provided in the dataset was corrupted as it had single quotes for property names instead of double, and there were lots of stray '\t' and '\n' characters present. As a result, we used regex string matching to extract the content, entities, and their respective polarities. Next, we append a hypothesis for the NLI task for each entity-premise pair in the dataset. Thereafter, we used Bert tokenizer to tokenize the premise and hypothesis and a gold_label is obtained from the polarity with three categories namely contradiction, entailment, and neutral. Then we split the dataset into train, validation, and test set in the ratio of 60:20:20. Finally, we merge the premise and hypothesis into a sequence and apply attention mask to the entire sequence. The attention mask helps the model to identify the useful tokens.

	gold_label	sentence1	sentence2
0	neutral	Tesla is Now Officially More Than Just About C...	"Facebook" is represented positively here
1	entailment	Health insurer Aetna\'s quarterly profit beats...	"Aetna" is represented positively here
2	entailment	The Capital Hotel sold to American hospitality...	"WarwickHotels&Resorts" is represented positiv...
3	contradiction	German prosecutors investigate Volkswagen chai...	"Volkswagen" is represented positively here
4	contradiction	German prosecutors investigate Volkswagen chai...	"Volkswagen" is represented positively here
...	...	...	...
19713	entailment	Yahoo to change name to Altaba\nYahoo will ado...	"Yahoo" is represented positively here
19714	entailment	Ryanair to pivot growth away from the UK\nBudg...	"Ryanair" is represented positively here
19715	entailment	Russian Hackers Con Advertisers out of $5 Mill...	"WhiteOps" is represented positively here
19716	entailment	Russian cybercriminals make $180 million with ...	"WhiteOps" is represented positively here
19717	entailment	VW\u2019s 13th brand: Moia, for ride-hailing (...	"Volkswagen" is represented positively here

19718 rows × 3 columns

	gold_label	sequence	attention_mask	token_type
18372	entailment	[CLS] tesla model 3 ev pilot production to rep...	1111111111111111111111...	0000000000000000000000...
16680	entailment	[CLS] morrison ##s expands amazon partnership ...	1111111111111111111111...	0000000000000000000000...
14801	contradiction	[CLS] yahoo to change name , trim board if ve ...	1111111111111111111111...	0000000000000000000000...
218	contradiction	[CLS] was russia behind te ##sco bank hack tha...	1111111111111111111111...	0000000000000000000000...
8422	entailment	[CLS] ali ##ba ##ba gets real \ na ##lib ##aba...	1111111111111111111111...	0000000000000000000000...
...	...	...	...	...
10931	entailment	[CLS] house ##bu ##ilde ##r per ##si ##mmon se...	1111111111111111111111...	0000000000000000000000...
15953	contradiction	[CLS] almost 2000 ##0 : a record close (but n...	1111111111111111111111...	0000000000000000000000...
10307	entailment	[CLS] em ##aar properties profit hits $ 1 . 42...	1111111111111111111111...	0000000000000000000000...
5795	contradiction	[CLS] galaxy note ##7 investigation results ar...	1111111111111111111111...	0000000000000000000000...
5491	contradiction	[CLS] apple ceo tim cook \ ' s pay declines af...	1111111111111111111111...	0000000000000000000000...

Figure 1: Sentences before preprocessing (left) and after (right)

2.2 Model

We have used BERT-base model size for BERT as it has 110M parameters and it is easier to download its pretrained weights. We then fine-tuned the model for our NLI task using our preprocessed dataset. We used AdamW optimizer and cross-entropy loss. We also used NVIDIA's open-source "apex.amp" tool for automatic mixed-precision training. The saved model file can be found here.

2.3 Hyperparameters

The choice of our hyperparameters are as follows:

- Batch size – 16
- Hidden layer dimension – 512
- Number of hidden layers – 1
- Learning rate – 2e-5
- Number of epochs – 6
- Warmup percent – 0.2

2.4 Results

Our model resulted in an F1 Score of 0.9258 and an accuracy of 96.46%.

```
content = 'Recently, huge criticisms have rocked Facebook.'
entity = 'Facebook'

get_prediction(content, entity, model, device)
```

```
'negative'
```

Figure 2: An example of a negative premise predicted by model

```
content = 'There has been a constant increase in the number of YouTube users'
entity = 'YouTube'

get_prediction(content, entity, model, device)
```

```
'positive'
```

Figure 3: An example of a positive premise predicted by model

```
content = 'A man is sitting on a red bench'
entity = 'Volkswagen'

get_prediction(content, entity, model, device)
```

```
'neutral'
```

Figure 4: An example of a neutral premise predicted by model

Chapter 5

Future Work and Conclusion

For this NLI task we used BERT-base as our model. We can later use BERT-large which has more parameters and hence can probably learn better. Also, in SNLI dataset EFL [4] model performed best so we can experiment to see how it performs for our dataset. Moreover, we haven't tested our model for robustness and generalization, so this line of research for the same task can be done in the future.

In this work, we have classified business news premises for each entity into positive, negative, and neutral which can help companies (entity here) take appropriate measures based on the polarity information. We have achieved very accuracy and hence polarity detection requires minimal human intervention. We hope our work inspires further research into NLI specific tasks in the business news domain or elsewhere.

References

[1] Pivovarova, Lidia & Klami, Arto & Yangarber, Roman. (2018). *Benchmarks and models for entity-oriented polarity detection*. 129-136. https://doi.org/10.18653/v1/N18-3016.

[2] Bing Liu and Lei Zhang. 2012. *A survey of opinion mining and sentiment analysis.* In Mining text data, Springer, pages 415–463.

[3] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/https://arxiv.org/abs/1810.04805v2.

[4] Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as Few-Shot Learner. arXiv. https://doi.org/10.48550/arXiv.2104.14690