Data_Aug

December 23, 2024

1 Data Augmentation

Here we balance out the datasets using https://groq.com/ free API.

```
[1]: import os
     import pandas as pd
     from groq import Groq
```

```
[2]: # Set up your Groq client
     client = Groq(api_key=os.getenv('GROQ_API_KEY'))
     print(client)
```

```
<groq.Groq object at 0x0000019592E92B50>
```

Read dataframe.

```
[11]: news = pd.read_csv('../data/financial_news.csv',
                         names=['sentiment', 'news'])
```

```
[13]:  # Select 10 positive news
      positive_news = news[news['sentiment'] == 'positive'].sample(10)
      # Select 5 negative news
      negative_news = news[news['sentiment'] == 'negative'].sample(6)

      # Combine the results
      selected_news = pd.concat([positive_news, negative_news])
```

```
[14]: selected_news.to_csv('../data/selected_news.csv', index=False)
```

```
[15]: selected_news
```

```
[15]:     sentiment                                               news
      913   positive  This is Done Logistics ' largest order in Norw…
      2282  positive  Growth was strongest in F-Secure 's operator I…
      329   positive               EPS grew to 0.04 eur from 0.02 eur .
      1184  positive  Atria Group is a leading Scandinavian meat pro…
      556   positive  STX Finland Oy signed a a preliminary agreemen…
      107   positive  In Lithuania , operating profit rose to EUR 19…
      1282  positive              ( I&H ) in a move to enhance growth .
```

```
3808  positive  The company will use the assets for strengthen…
202   positive  First quarter underlying operating profit rose…
2175  positive  Pretax profit totalled EUR 2.0 mn , compared t…
4733  negative  However , its market share shrank to 47.59 per…
4737  negative  In food trade , sales amounted to EUR320 .1 m …
2797  negative  Also the city 's insurance company , If P & C …
4732  negative  Group EBIT for the first half was EUR13 .6 m U…
4213  negative  Last year , UPM cut production , closed mills …
4289  negative  When the web user clicks on the link contained…
```

Now we do data augmentation for the selected texts, by making the value of negative news equal that of the positive.

[16]:
```python
selected_news.sentiment.value_counts()
```

[16]:
```
sentiment
positive    10
negative     6
Name: count, dtype: int64
```

[17]:
```python
df = selected_news.copy()
```

[19]:
```python
# Number of rows we want for label negative
target_rows_label_negative = 10

# Find underrepresented rows (label == 0)
underrepresented_texts = df[df['sentiment'] == 'negative']['news'].tolist()

# Number of examples we currently have for label 0
current_rows_label_negative = len(underrepresented_texts)

# Number of additional examples we need
needed_examples = target_rows_label_negative - current_rows_label_negative
```

[20]:
```python
underrepresented_texts
```

[20]:
```
['However , its market share shrank to 47.59 per cent from 48 per cent a year
earlier .',
 'In food trade , sales amounted to EUR320 .1 m , a decline of 1.1 % .',
 "Also the city 's insurance company , If P & C Insurance , has said it will not
pay compensation .",
 'Group EBIT for the first half was EUR13 .6 m US$ 17.8 m , falling short of the
EUR22 .5 m it posted for the same period of 2009 .',
 'Last year , UPM cut production , closed mills in Finland and slashed 700 jobs
.',
 'When the web user clicks on the link contained in the mail , he finds himself
on a bogus site that imitates that of his bank , and which retrieves his
personal banking data .']
```

[21]:

```
# Augment the underrepresented class with new examples
augmented_texts = []
for i in range(needed_examples):
    # Select a random text from the underrepresented class to augment
    text = underrepresented_texts[i % current_rows_label_negative]  # Cycle␣
 ↪through available texts if needed

    response = client.chat.completions.create(
        messages=[
            {"role": "system", "content": "You are a data augmentation␣
 ↪assistant."},
            {"role": "user", "content": f"Generate a concise headline similar␣
 ↪to: {text} and reply with response only without quotes"},
        ],
        model="llama3-8b-8192"
    )

    # Get the augmented text from the response
    augmented_data = response.choices[0].message.content
    augmented_texts.append(augmented_data)
```

[22]:

```
# Create a new DataFrame for the augmented data
augmented_df = pd.DataFrame({
    'news': augmented_texts,
    'sentiment': ['negative'] * needed_examples  # Label the new examples as 0
})

# Combine the original and augmented DataFrames
balanced_df = pd.concat([df, augmented_df], ignore_index=True)

# Display the balanced DataFrame
# print(balanced_df)
```

```
   sentiment                                               news
0   positive  This is Done Logistics ' largest order in Norw…
1   positive  Growth was strongest in F-Secure 's operator I…
2   positive               EPS grew to 0.04 eur from 0.02 eur .
3   positive  Atria Group is a leading Scandinavian meat pro…
4   positive  STX Finland Oy signed a a preliminary agreemen…
5   positive  In Lithuania , operating profit rose to EUR 19…
6   positive              ( I&H ) in a move to enhance growth .
7   positive  The company will use the assets for strengthen…
8   positive  First quarter underlying operating profit rose…
9   positive  Pretax profit totalled EUR 2.0 mn , compared t…
10  negative  However , its market share shrank to 47.59 per…
11  negative  In food trade , sales amounted to EUR320 .1 m …
12  negative  Also the city 's insurance company , If P & C …
13  negative  Group EBIT for the first half was EUR13 .6 m U…
```

```
14  negative  Last year , UPM cut production , closed mills …
15  negative  When the web user clicks on the link contained…
16  negative  Company's market share decreases by 0.41 perce…
17  negative  Retail Sales Decline 1.1%, Reaching EUR320.1 M…
18  negative         Insurance Company Refuses to Cover Damages
19  negative  Company Experiences Decline in EBIT for First …
```

[23]:

```
balanced_df
```

[23]:

```
   sentiment                                              news
0   positive  This is Done Logistics ' largest order in Norw…
1   positive  Growth was strongest in F-Secure 's operator I…
2   positive              EPS grew to 0.04 eur from 0.02 eur .
3   positive  Atria Group is a leading Scandinavian meat pro…
4   positive  STX Finland Oy signed a a preliminary agreemen…
5   positive  In Lithuania , operating profit rose to EUR 19…
6   positive             ( I&H ) in a move to enhance growth .
7   positive  The company will use the assets for strengthen…
8   positive  First quarter underlying operating profit rose…
9   positive  Pretax profit totalled EUR 2.0 mn , compared t…
10  negative  However , its market share shrank to 47.59 per…
11  negative  In food trade , sales amounted to EUR320 .1 m …
12  negative  Also the city 's insurance company , If P & C …
13  negative  Group EBIT for the first half was EUR13 .6 m U…
14  negative  Last year , UPM cut production , closed mills …
15  negative  When the web user clicks on the link contained…
16  negative  Company's market share decreases by 0.41 perce…
17  negative  Retail Sales Decline 1.1%, Reaching EUR320.1 M…
18  negative         Insurance Company Refuses to Cover Damages
19  negative  Company Experiences Decline in EBIT for First …
```

[26]:

```
balanced_df.sentiment.value_counts()
```

[26]:

```
sentiment
positive    10
negative    10
Name: count, dtype: int64
```

[24]:

```
balanced_df.to_csv('../data/balanced_news.csv', index=False)
```

This approach can be applied to other text based applications.

[ ]:

```

```