Unified Approach to Structured Sentiment Analysis

Javier Rondon, Jae Rim, Ratan Singh
Structured Sentiment Analysis: Understanding Human Emotion and Opinion

Problem:
- Huge surge in textual data generation through various digital platforms
- Unstructured data is complex and carries vital but difficult to decipher insights
- The challenge of efficiently understanding and interpreting human emotions and opinions

Specific Problem We're Solving:
- Developing advanced techniques in structured sentiment analysis
- Transforming unstructured text into structured, actionable insights
- Analyzing emotional responses and subjective patterns for clearer view of public sentiment
Structured sentiment problem

Each opinion consists of four elements

holder, target, expression, polarity

For example:

“The room was good, but I prefer the penthouse”

The challenge is to create a model that extracts the elements of the opinions:

(“-”, room, “was good”, “Positive”)

(“I”, “penthouse”, “prefer”, “Positive”)
Proposed Solution

- Create a simple and unified model for automatic extraction of all aspects of the opinion from the text.
- Compare the performance of two approaches:
  - Use of BERT type models, leveraging a novel architecture (GTS, Lu et al 2022, Wu et al, 2020)
  - Use of GPT models (Chat GPT 3.5, InstructGPT Davinci and Courie) with fine tuning and few-shot approaches

[Evolutionary Tree of Large Language Models]

[https://twitter.com/ylecun/status/1651762787373428736?s=61&t=vlIRCc6kN2k7bJ-gXW7Cw7w]
Model and Architecture
Datasets

- 26 thousand reviews in five languages and different domains
  - Lists of dictionaries with keys for opinion expressions, holders, targets, polarity and strength

- **OpeNER** - Hotel reviews in English (Agerri et al., 2013)
- **OpeNER** - Hotel reviews in Spanish (Agerri et al., 2013)
- **Norec** - Music, literature, game reviews in Norwegian (Øvrelid et al., 2020)
- **MPQA** - News texts in English (Wiebe et al., 2005)
- **Darmas Unis** - University review in English (Toprak et al., 2010)
- **Multibooked** - Hotel reviews in Catalan (Barnes et al., 2018)
- **Multibooked** - Hotel reviews in Basque (Barnes et al., 2018)
Insights from EDA

- Large class imbalance
  - 84 percent of holders are implicit, 15 percent of targets are implicit.

- Quality of annotations
  - Language nuance
  - Ambiguity in identifying aspects in complex text.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Holders</th>
<th>Targets</th>
<th>Expressions</th>
</tr>
</thead>
</table>
|                 | Explicit | Implicit | percentage | Count | Avg. | Max. |%
|                 |%         |%         |            |        | per review | per review |
| OpeNER-EN       | 266      | 2,618    | 91         | 2,679  | 205  | 7    | 2,884 | 2.0 | 17 |
|                 | 49       | 351      | 88         | 371    | 29   | 7    | 400   | 2.0 | 13 |
| OpeNER-ES       | 176      | 2,866    | 94         | 2,748  | 294  | 10   | 3,042 | 2.4 | 26 |
|                 | 23       | 364      | 94         | 363    | 24   | 6    | 387   | 2.6 | 11 |
| Norec           | 898      | 7,550    | 89         | 6,778  | 1,670| 20   | 8,448 | 0.2 | 15 |
|                 | 120      | 1,312    | 92         | 1,152  | 280  | 20   | 1,432 | 1.7 | 7  |
| Multibooked_eu | 205      | 1,474    | 88         | 1,277  | 402  | 24   | 1,679 | 1.9 | 15 |
|                 | 33       | 170      | 84         | 152    | 51   | 25   | 203   | 1.7 | 7  |
| Multibooked_ca | 169      | 1,820    | 92         | 1,705  | 284  | 14   | 1,989 | 2.0 | 22 |
|                 | 15       | 243      | 94         | 211    | 47   | 18   | 258   | 1.8 | 17 |
| MPQA            | 1,425    | 281      | 16         | 1,481  | 225  | 13   | 1,706 | 1.4 | 8  |
|                 | 406      | 164      | 29         | 494    | 76   | 13   | 570   | 1.4 | 7  |
| Darmas Unis     | 63       | 743      | 92         | 806    | 0    | 0    | 806   | 1.2 | 5  |
|                 | 9        | 89       | 91         | 98     | 0    | 0    | 98    | 1.2 | 3  |
| Total           | 3,857    | 20,045   | 84         | 20,315 | 3,587| 15   | 23,902|     |    |
Grid Tagging Scheme (GTS) for BERT classifier

- We classify every pair of tokens with a tag to identify their function and relation in the opinion: holders, targets, expression, polarity

- The model learns the relationship between each pair of tokens according to the tag
GTS architecture overview

Given a sentence $s = \{w_1, w_2, ..., w_n\}$
Use transformer encoder to generate a representation $R_{ij}$ of the word-pair $(w_i, w_j)$

From Wu et al. "Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction" 2021
Results from GTS architecture

- Experimented GTS with different pre-trained BERT variant models
- Models trained on NVIDIA H100 (80 GB)
- Training times in excess of 12 hours
- Best models are RoBERTa-large and XLM-RoBERTa

Results:

- Test Sentiment F1 score in all datasets beating the published baselines
- Predicting the holder is easier than target or polar expression (expected since targets and expressions are longer sequences)
Beyond BERT
InstructGPT fine-tuning (Davinci and Courie)

- Dataset processing - Cast the classification task to text-to-text format with training prompts:
  
  ```json
  {"text": "The room was good", "opinions": [{"Source": [], "Target": ["the room"], "Expression": ["was good"], "Polarity": "Positive" }]
  ```

  "prompt": "The room was good ->", "completion": " [{"Source": []}, "Target": ["the room"], ...

  **Input**: [101, 1523, 25732, 1524, 1024, 1523, 1996, 2282, 2001, 2204, 1011, 1028, 1524, 1010, 1523, ...]

  **Labels**: [1523, 25732, 1524, 1024, 1523, 1996, 2282, 2001, 2204, 1011, 1028, 1524, 1010, 1523, ...]

- Model learns to predict the next token according to the examples in the training set.
ChatGPT Few-Shot Training example

- **ChatGPT's Training**: Broad training on internet text, no specific task optimization.
- **Generalization**: ChatGPT generalizes from these few examples and attempts to provide a suitable response based on its understanding of the task.
- **Performance** is driven by prompt design and quality of examples.

Definition: The output will be a list with a dictionary with keys for source (both implicit and explicit), target (both implicit and explicit), expression, polarity and intensity. In cases where there are no source or target, the output should be none.

Example 1-

- **input**: Bit pricey and but away from center
  - **output**: [{'Source': [], [], 'Target': [], [], 'Polar_expression': [['Bit pricey'], ['0:10']], 'Polarity': 'Negative', 'Intensity': 'Standard'}, {'Source': [], [], 'Target': [], [], 'Polar_expression': [['away from center'], ['19:35']], 'Polarity': 'Negative', 'Intensity': 'Standard']}

**Task instruction**

Example 2:

Input:

```
The wellness facilities are perfect, but can be crowded on Saturday.
```

Output:

```
['Source': [], [], 'Target': ['wellness facilities'], ['4:23'], 'Polar_expression': ['perfect', '28:35'], 'Polarity': 'Positive', 'Intensity': 'Standard'],
['Source': [], [], 'Target': ['wellness facilities'], ['4:23'], 'Polar_expression': ['crowded on saturday'], ['50:68'], 'Polarity': 'Negative', 'Intensity': 'Standard']
```
Addressing GPT Challenges for Structured Sentiment Analysis:

- **GPT Incomplete Opinions**: Precision in prompt design can guide ChatGPT through vague opinions.
- **Missing Outputs**: Carefully crafted prompts help generate desired outputs.
- **Hallucinations**: Additional prompt guidance can reduce instances of unasked information.
- **API Token Limits**: Maximizing information in limited tokens and selecting broad-representative examples can counter token restrictions.
- **Key Takeaway**: Success with ChatGPT lies in prompt optimization.

Nice, thank you! If you have any more text you'd like to analyze, feel free to let me know. Once upon a time…
Results from GPT architecture

- Performance of finetune GPT is similar to GTS-BERT
- Few-shot performance is decent considering it only needed a few examples and context length is limited.

Which model to use? In our datasets, the BERT-based GTS performed better than GPT despite having fewer parameters, but decision depends on availability of hardware resources, training data and time.

<table>
<thead>
<tr>
<th>Model</th>
<th>OpeNER English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Baseline</td>
<td>0.521</td>
</tr>
<tr>
<td>Seq Baseline</td>
<td>0.329</td>
</tr>
<tr>
<td>GTS model</td>
<td>0.630</td>
</tr>
<tr>
<td>finetune GPT Davinci</td>
<td>0.598</td>
</tr>
<tr>
<td>finetune GPT Curie</td>
<td>0.550</td>
</tr>
<tr>
<td>Few-Shot ChatGPT 3.5</td>
<td>0.455</td>
</tr>
</tbody>
</table>
Qualitative analyses of results

- Fine-tuned GPT has similar performance of BERT type models.
- If there are not many examples of complex syntax, few-shot has problems.
The 3 components of model output: Holder, Target, Expression are combined to form full sentence for each polarity

Review
“The room was good, but I prefer the penthouse”

Model Output
(“-”, room, “was good”, “Positive”)  
(“I”, “penthouse”, “prefer”, “Positive”)  

Sentences Generated
room was good  
I penthouse prefer

Context Windows Generated for “Phrases” and “Collocates” generation from Sentences

For the word “penthouse”, context window of “2” captures 1 word to left and 1 word to the right of “penthouse”
Data Visualization and Analysis: Phrases, Collocates and TermsBerry

- Identify the top targets we need to analyze for each polarity separately
- Generate Phrases, Collocates and TermsBerry graphs for analysis for various targets

### Phrases for “hotel” (Positive)

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Length</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotel very</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel great</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel nice</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel clean</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel rooms</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel fantastic</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel good</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>hotel grounds</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

★ Phrases represent continuous set of words in target content window
★ Context window of 2 used above

### Collocates for “hotel” (Positive)

<table>
<thead>
<tr>
<th>Term</th>
<th>Collocate</th>
<th>Count (context)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotel</td>
<td>hotel</td>
<td>58</td>
</tr>
<tr>
<td>hotel</td>
<td>nice</td>
<td>17</td>
</tr>
<tr>
<td>hotel</td>
<td>clean</td>
<td>15</td>
</tr>
<tr>
<td>hotel</td>
<td>rooms</td>
<td>14</td>
</tr>
<tr>
<td>hotel</td>
<td>great</td>
<td>12</td>
</tr>
<tr>
<td>hotel</td>
<td>good</td>
<td>12</td>
</tr>
<tr>
<td>hotel</td>
<td>staff</td>
<td>9</td>
</tr>
<tr>
<td>hotel</td>
<td>room</td>
<td>7</td>
</tr>
<tr>
<td>hotel</td>
<td>lovely</td>
<td>7</td>
</tr>
</tbody>
</table>

★ Collocates represent pair of words in target content window
★ Context window of 5 used above

### TermsBerry for “hotel” (Positive)

Visual representation of collocates across all the sentences
Possible Next Applications

Product Analyzer
- Website of Reviews
- Web scrape
- Run through sentiment analysis model
- Results
- Decisions about product

Targeted Consumer Analysis
- Review of products
- Run through sentiment analysis model
- Results
- +
  - User information
  - Curated recommendation
  - Customer segmentation of product
Innovation in Architecture: Extended the novel GTS architecture to accurately predict all sentiment dimensions in text

BERT's Strength: The larger BERT models yielded superior performance, affirming the strength of this architecture for sentiment tasks.

Performance: Our approaches beat published baseline and remained competitive with other models.

Task Conversion: Converting the structured sentiment task to a text-to-text format enabled utilization of larger GPT models.

GPT Tuning: DaVinci model fine-tuning exceeds baselines, rivals BERT classifiers.

Few-Shot Efficiency: Achieved reasonable performance with fewer training examples.

BERT vs GPT: BERT excels with large datasets; GPT shines when data is limited.

Summary: Exciting advancements achieved in structured sentiment analysis using large language models.
Next Steps and Broad Applicability

- **Advanced Hyperparameter Tuning**: Enhance performance with advanced hyperparameter tuning.
- **Explore Hybrid Models**: Merge GPT and BERT strengths for a performance boost
- **Generalizability**: Our nuanced, multidimensional sentiment analysis approach can extend beyond binary views, capturing emotion intensity, subjectivity, and specific emotional categories. This model's principles are adaptable across various NLP tasks, catering to diverse data types and challenges. The refined sentiment interpretation finds applications in areas like customer service, social media monitoring, and mental health assessment, maximizing our model's real-world impact.
Mission: Enhance structured sentiment analysis with Large Language Models.

Impact: Enhancing business decisions with nuanced sentiment understanding.

Vision: Transform sentiment analysis into a powerful tool for global businesses.
Questions?
Acknowledgements and additional resources

- Dr. Natali Ahn  266 Instructor.
- Wu et al “Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction” 2021
- Hosseini-Asl et al “A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis” 2022
- Wang et al “SUPER-NATURALINSTRUCTIONS: Generalization via Declarative Instructions on 1600+ NLP Tasks 2022
- Scaria et al “InstructABSA: Instruction Learning for Aspect Based Sentiment Analysis” 2023
- Visualization tools at https://voyant-tools.org/
Acknowledgements and additional resources

- Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction.