Unified Approach to Structured Sentiment Analysis

Javier Rondon, Jae Rim, Ratan Singh

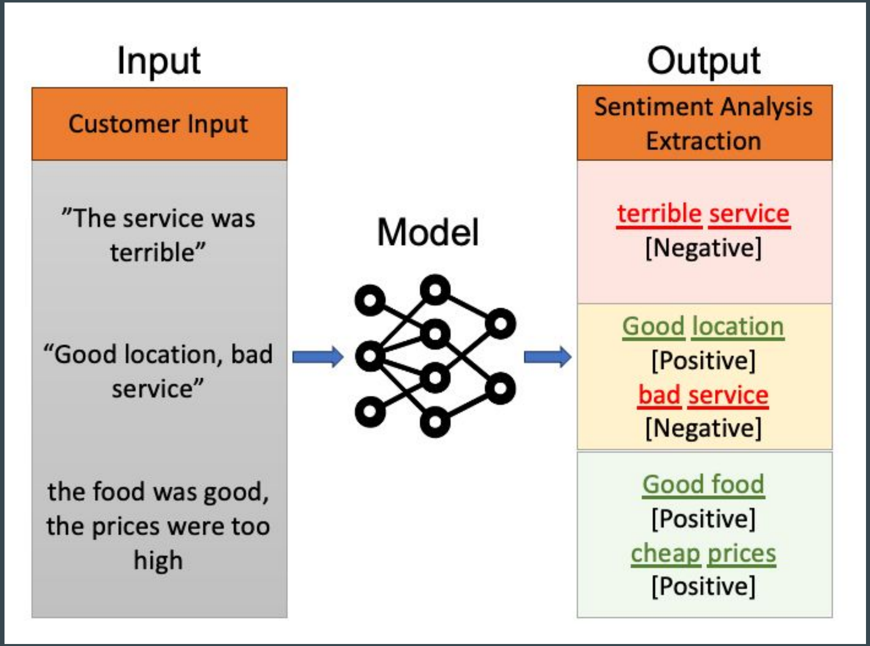
Structured Sentiment Analysis: Understanding Human Emotion and Opinion

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”)

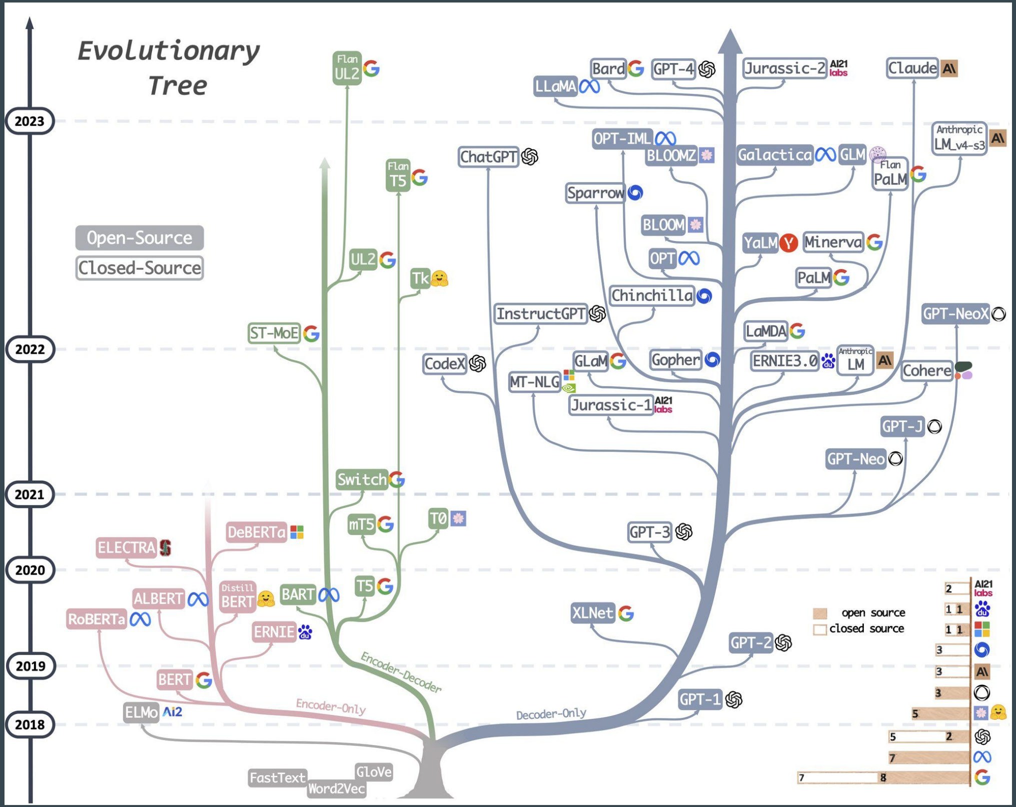
(“I”, “penthouse”, “prefer”,”Positive”)

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

https://twitter.com/ylecun/status/1651762787373428736?s=61&t=vIRCc6kN2k7bJ-gXW7Cw7w

Evolutionary Tree of Large Language Models



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.

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

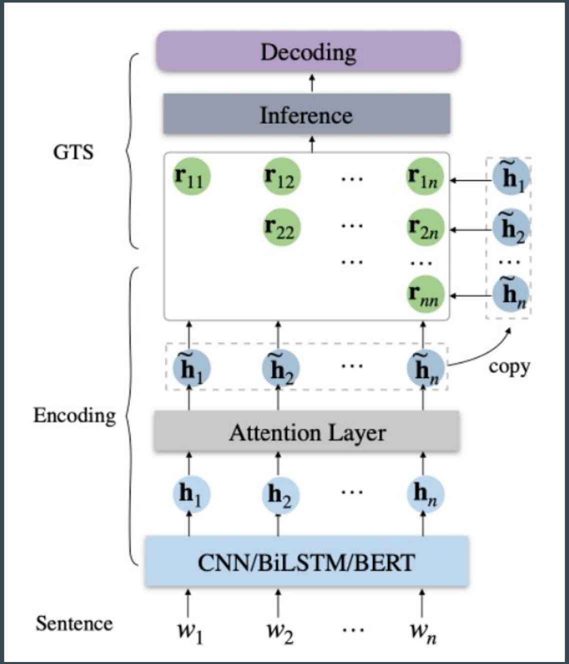
- We classify every pair of tokens with a tag to identify their function and relation in the opinion: holders, targets, expression, polarity

[CLS]	Fantastic	food	and	breathtaking	view	
Implicit Holder	Positive	0	0	Positive	0	[CLS]
	Expression	Positive	0	0	0	Fantastic
		Target	0	0	0	food
			0	0	0	and
				Expression	Positive	breathtaking
					Target	view

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

Pair	Tag
[CLS],[CLS]	Implicit Holder
Fantastic,Fantastic	Opinion
breathtaking,breathtaking	Opinion
[CLS],Fantastic	Positive
[CLS],breathtaking	Positive
Fantastic,food	Positive
breathtaking,view	Positive
food,food	Target
view,view	Target

GTS architecture overview



Given a sentence s ={w1, w2, ..., wn}

Use transformer encoder to generate a representation

Rij of the word-pair (wi,wj)

Inference block

$$
\begin{aligned}
\mathbf{p}_i^{t-1} &= \text{maxpooling}(\mathbf{p}_{i,:}^{t-1}),\\
\mathbf{p}_j^{t-1} &= \text{maxpooling}(\mathbf{p}_{j,:}^{t-1}),\\
\mathbf{q}_{ij}^{t-1} &= [\mathbf{z}_{ij}^{t-1}; \mathbf{p}_i^{t-1}; \mathbf{p}_j^{t-1}; \mathbf{p}_{ij}^{t-1}],\\
\mathbf{z}_{ij}^{t} &= \mathbf{W}_q \mathbf{q}_{ij}^{t-1} + \mathbf{b}_q,\\
\mathbf{p}_{ij}^{t} &= \text{softmax}(\mathbf{W}_s \mathbf{z}_{ij}^{t} + \mathbf{b}_s).
\end{aligned}
$$

Z_{ij}^t Feature representation of the word pair wi,wj)

p_{ij}^t Probability distribution of the word pair wi,wj)

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 F_1 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)

Dataset		Language Model	Sentiment F1 Score	Precision	Recall
OpenER-EN	Dev	BERT_review	0.65	0.69	0.62
	Test		0.63	0.66	0.6
OpenER-ES	Dev	XLM_roberta_large	0.67	0.74	0.62
	Test		0.61	0.71	0.54
MultiBook EU	Dev	XLM_roberta_large	0.69	0.57	0.53
	Test		0.64	0.63	0.53
MultiBook CA	Dev	XLM_roberta_large	0.68	0.7	0.63
	Test		0.67	0.7	0.64
NoReC	Dev	XLM_roberta_large	0.51	0.51	0.48
	Test		0.45	0.47	0.43
Darmstadt Unis	Dev	roberta-large	0.36	0.41	0.33
	Test		0.38	0.44	0.34

	OpeNER		Multibooked		Norec	DS
Dataset	EN	ES	EU	CA	NO	EN
Graph Baseline	0.521	0.495	0.545	0.516	0.272	0.204
Seq Baseline	0.329	0.24	0.365	0.338	0.123	0.06
Our Current Best	0.63	0.61	0.67	0.64	0.45	0.38

Beyond BERT

InstructGPT fine-tuning (Davinci and Courie)

- Dataset processing - Cast the classification task to text-to-text format with training prompts:

{"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

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

Task instruction

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'}]

Input

now complete the following example- "The wellness facilities are perfect , but can be crowded on saturday ."

Output

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

{ "text": "Nice",
"opinions": "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 fined-tune 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.

Model	OpeNER English
Graph Baseline	0.521
Seq Baseline	0.329
GTS model	0.630
fined-tune GPT Davinci	0.598
fined-tune GPT Curie	0.550
Few-Shot ChatGPT 3.5	0.455

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 hotel is nice and clean , but is very far from any nice sorrundings ."

Model	Source	Target	Polar Expression	Polarity
GTS (BERT)		The hotel	very far from any nice sorrundings	Negative
		The hotel	nice	Positive
		The hotel	clean	Positive
fined-tuned GPT Curie		The hotel	very far from any nice sorrundings	Negative
		The hotel	nice	Positive
		The hotel	clean	Positive
Few Shot GPT3.5		surroundings	any nice	Negative
		hotel	nice	Positive
		hotel	clean	Positive
Gold File		The hotel	very far from any nice sorrundings	Negative
		The hotel	nice	Positive
		The hotel	clean	Positive

Visualization of Model Results

- ➢ The 3 components of model output: Holder, Target, Expression are combined to form full sentence for each polarity

Review	Model Output	Sentences Generated
"The room was good, but I prefer the penthouse"	("-", room, "was good", "Positive")	room was good
	("I", "penthouse", "prefer","Positive")	I penthouse prefer

- ➢ Context Windows Generated for "Phrases" and "Collocates" generation from Sentences

I **penthouse** prefer

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)

Term	Count ↓	Length	Trend
hotel very	8	2	
hotel great	5	2	
hotel nice	5	2	
hotel clean	3	2	
hotel rooms	3	2	
hotel fantasic	2	2	
hotel good	2	2	
hotel grounds	2	2	

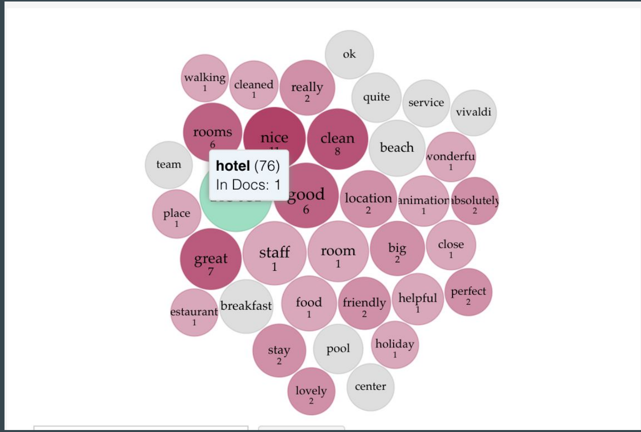
★ Phrases represent continuous set of words in target content window
★ Context window of 2 used above

Collocates for "hotel" (Positive)

Term	Collocate	Count (context)
hotel	hotel	58
hotel	nice	17
hotel	clean	15
hotel	rooms	14
hotel	great	12
hotel	good	12
hotel	staff	9
hotel	room	7
hotel	lovely	7

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

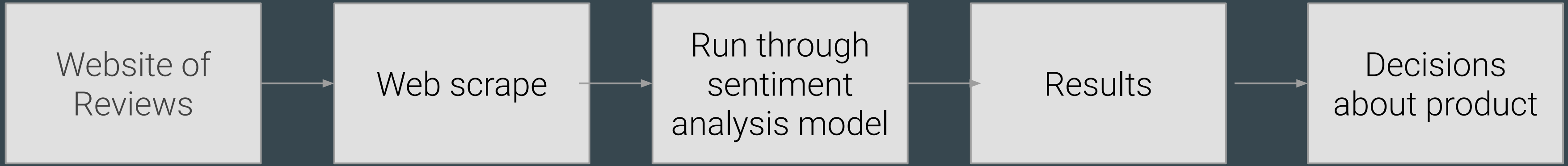
TermsBerry for "hotel" (Positive)



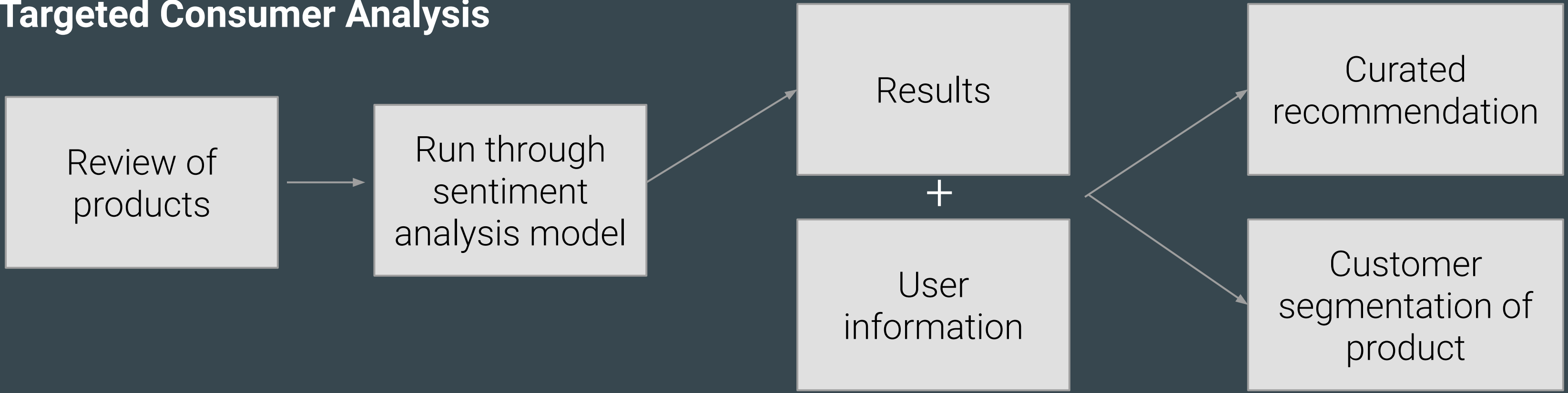
★ Visual representation of collocates across all the sentences

Possible Next Applications

Product Analyzer



Targeted Consumer Analysis



Summary

- **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:
Using LLMs for Enhanced Understanding of human emotions

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

- Generative AI at https://chat.openai.com/ and https://platform.openai.com/playground
- Visualization tools at https://voyant-tools.org/

Acknowledgements and additional resources

- Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing.
- Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis.
- Xinyu Lu, Mengjie Ren, Yaojie Lu, and Hongyu Lin. 2022. ISCAS at SemEval-2022 task 10: An extraction-validation pipeline for structured sentiment analysis.
- Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan,Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction