Pragmatic Perspective on Assessing Implicit Meaning Interpretation in Sentiment Analysis Models

Rashid Mustafin
Norwegian School of Economics
Bergen, Norway
rashid.mustafin@nhh.no

Abstract

Drawing on pragmatic theories of implicature by Grice (1975) and Levinson (1983), according to which speakers often convey more than it is explicitly said, the paper argues that interpreting texts with implicit meaning correctly is essential for precise natural language understanding. To illustrate the challenges in computational interpretation of implicatures, the study introduces a series of illustrative micro-experiments with the use of four transformer models fine-tuned for sentiment analysis. In these micro-experiments, the models classified sentences specifically designed to expose difficulties in handling implicit meaning. The study demonstrates that contrasting qualitative pragmatic analysis with the models' tendency to focus on formal linguistic markers can reveal the limitations of supervised machine learning methods in detecting implicit sentiments.

1 Introduction

Natural language processing models are used widely by businesses and researchers today. With the increasing quality of supervised machine learning, the demand for linguistic expertise in developing these technologies has diminished, especially compared to the earlier time when rule-based approaches were the norm. This tendency has led to a lower level of transparency and explainability. In this paper, the problem is approached through the example of sentiment analysis. It is posited that linguists' attempts to explain the process of intuitive sentiment interpretation qualitatively must persist because the "black box" nature of the state-of-the-art NLP techniques implies unpredictability and risks of affecting decision-making processes negatively. This study presents a pragmatic perspective on implicit meaning in interpreting sentiment and discusses the role of common sense knowledge and contextual understanding that transformer models still seem to lack. A theoretical examination is complemented by a series of illustrative micro-experiments with the use of four transformer sentiment analysis models.

2 Pragmatic Theory of Implicit Meaning

As Levinson (1983, p.97) puts it with a reference to Grice (1975), sometimes people mean more than what is formally stated in the utterance. Levinson (1983) claims that semantic theory is not enough for interpreting such cases because formal semantic analysis does not take into consideration the context and the intentions of the speakers. He uses an example of a dialogue consisting of two utterances (1).

(1) A: Can you tell me the time?
B: Well, the milkman has come.
(Levinson, 1983, p.97)

According to Levinson (1983), should one use the semantic approach for interpreting this interaction, the first utterance can be paraphrased as "Do you have the ability to tell me the time?" (Levinson, 1983, p.98). The second utterance would be decoded as "[...] the milkman came at some time prior to the time of speaking" (Levinson, 1983, p.98). Formally, this interpretation is correct as it reflects the meanings of the lexis and the grammatical structures utilised by the speakers. However, in a real conversation native speakers would extract more information from these phrases than it seems there is semantically. In the first utterance, there is not only a question about the ability to tell the time on the moment of speaking but also a request to do it. The second utterance implies the inability to tell the exact time and instead shares the information that could be relevant for the situation. Levinson (1983, pp.102–103) notes that one utterance can lead to an endless list of inferences, but it does not mean that all of them must be taken into account while interpreting speech. What helps people deduce the relevant implicatures is the assumption

that the participants of communication strive to sustain Gricean cooperative principles (Grice, 1975). Grice's cooperative principles include the maxims of of quality ('be truthful'), quantity ('be informative'), relation ('be relevant'), and manner ('be perspicuous') (Grice, 1975, pp.45–46). As Levinson (1983, pp.102–103) notices, the examples of sentences with implicatures seem to fail in terms of fulfilling the maxims of quantity and relation when interpreted semantically: the reply about the milkman provides information that was not requested instead of what was actually asked, which makes it not informative and not relevant. Assuming that the speaker B is following the cooperative principles, the range of possible implicatures for the utterances to make sense shrinks to only a few, which are then narrowed down to the most likely one in the light of the given context.

The ideas expressed by Grice (1975) and Levinson (1983) are applicable to the problems of natural language processing. Taking into consideration the fundamental role of implicatures in communication, it is impossible to avoid processing texts with implicatures in almost any research or industrial application of NLP models. For example, applying the sentiment analysis perspective, such a review as (2) implies that the tent is sturdy, which is a positive evaluation.

(2) The tent could withstand a hurricane.

There was an attempt to design a rule-based solution for sentiment analysis of implicit judgements (Wiebe and Deng, 2014), but seemingly no published work on fine-tuning the supervised machine learning models specifically to interpreting implicatures for sentiment analysis and no research on the mistakes they make in this regard. Wiebe and Deng (2014) also used Grice's theory of implicatures to suggest a conceptual framework of a system for identifying implied sentiments with the use of a manually annotated lexicon of words. Wiebe and Deng (2014) establish rules for processing certain syntactic patterns, but their system has some significant limitations. The rules and the lexicons are not exhaustive. Judging by the number of citations of this paper, it did not receive much attention by the research community despite the importance of the topic raised, which might have been caused by the decreasing popularity of rule-based language technologies at that time.

Speculating on bridging linguistic insights and computational processing of evaluative language, Benamara et al. (2017, pp.233–236) also briefly touch upon the problem of implicit meaning. They differentiate between three ways of making the sentiment implicit. The first way is describing conventionally favourable or unfavourable circumstances. This type of implicit meaning can be decoded through common sense and general knowledge. One of the examples they give is (3). In this case, it is deforming after a short time that characterises the mattress negatively.

(3) Within a month, a valley formed in the middle of the mattress.
(Benamara et al., 2017, p.235)

The second way of implicit sentiment expression is using objective characteristics that have positive or negative connotations. An example given by Benamara et al. (2017) is (4). This study, however, disagrees on the implicitness of the second type of sentiment expression in Benamara's work. If a word has an established positive or negative connotation, the sentiment is explicit. Benamara et al. (2017) also mention that there are words that can have different connotations depending on the domain: they note that volume is good for hair but bad for things one has to carry in public transport. It is not clear why this kind of examples must be considered separately from the first type of implicit expression of sentiment. After all, it is also a description of a desirable situation in the case of hair, and an undesirable situation in the case of public transport.

(4) Jim is a vagrant.
(Benamara et al., 2017, p.235)

The third way is evaluating an implicit aspect of the opinion target. According to Benamara et al. (2017), (5) exemplifies the third type of implicit expression because it implies a negative evaluation of the aspect of durability. Nevertheless, this type is also questionable in terms of what makes it different from the first one because the example given for the first type, (3), could be also called an evaluation of an aspect.

(5) My new phone lasted three days.
(Benamara et al., 2017, p.236)

Although this study does not agree on the entire categorisation given by Benamara et al. (2017), it accepts the idea of the first type of implicit sentiment expression, i.e. that a reference to a situation that is conventionally regarded negatively is a way to express a sentiment implicitly.

3 Micro-experiments

This section reports on how the four open-source transformers classify sets of sentences that were designed for highlighting potential problematic areas in computational interpretation of implicit meaning. They include the BERT-base model by NLP Town (NLPTown, 2023), the RoBERTa-base model by CardiffNLP (Barbieri et al., 2020), the DistilBERT-base model (HuggingFace, 2022), and another RoBERTa-base model fine-tuned on a wider range of genres and called SiEBERT (Hartmann et al., 2023).

The first micro-experiment poses the question of whether the models are capable of identifying desirable characteristics of two entities and inferring whether a given sentence is indicating a negative or a positive evaluation through comparison. (6) exemplifies a comparison between the volume of the speaker and a phone. There are two possible explanations of how the sentence could be processed: through logic and general knowledge and through formal markers. Operating with general knowledge, a human being would compare how loud an ordinary speaker and an ordinary phone are. Knowing that speakers are usually considerably louder than phones, one would conclude that a speaker that is only insignificantly louder than a phone must be of low quality. Judging by the concrete constructions that could be recurrent in the sentences with a negative sentiment, the pattern that deserves our attention is *barely louder than*. The correct attribution of sentences with the necessity to collate the opinion target properties and the characteristics of other items, like it was shown in (6). Sentences (8-17) replace a phone and a speaker by other entities. The compared entities were altered so that the sentiment orientation varied. Each sentence was also duplicated with the entities from the original sentence swapped.

(6) This speaker is barely louder than my phone.
(7) The phone is barely louder than my speaker.
(8) The stereo system is barely louder than a music box.
(9) The music box is barely louder than a stereo system.
(10) The parrot is barely louder than a fish.
(11) The fish is barely louder than a parrot.
(12) The keyboard is barely louder than the heartbeat.
(13) The heartbeat is barely louder than the keyboard,
(14) The car engine is barely louder than a fridge.
(15) The fridge is barely louder than a car engine.
(16) The neigbours are barely louder than library visitors.
(17) The library visitors are barely louder than the neigbours.

Appendix A includes detailed tables with the expected answers and the labels assigned by the models. In general, (6–17) were attributed to the negative class by all the models. A few exceptions were (9, 16, 17) that were classified as neutral by the CardiffNLP classifier. These exceptions do not seem to have any logical explanation behind, so it can be concluded that the construction *barely louder than* does contribute to the negative sentiment identification. Even when it is more relevant to opt for a positive sentiment, like in (12) or (14), the models still choose negative. Moreover, some non-sensical examples were also classified as negative. The models reacted to a construction that might have appeared in negative contexts and classified all sentences as negative without any apparent consideration for the entities compared.

To investigate the role of the construction *is smaller than* in the same manner as the construction *is barely louder*, the second experiment was designed (18-29). Both bi-class models, DistilBERT and SiEBERT, classified all these sentences except for (19) as negative. RoBERTa attributed all sentences to the neutral class, while BERT classified (19–24) as neutral and (25–28) as negative. In principle, it is possible to assign neutral label to all sentences, although it was intended that (20, 22) were negative, (21, 23–27, 29) were neutral, and (28) was positive. Yet there might be a certain bias to the negative sentiment towards the construction *is smaller than*.

(18) The shower is smaller than a phone booth.
(19) The phone booth is smaller than a shower.
(20) The throne is smaller than a highchair.
(21) The highchair is smaller than a throne.
(22) The pocket is smaller than a matchbox.
(23) The matchbox is smaller than a pocket.
(24) The hummingbird is smaller than a teacup.
(25) The teacup is smaller than a hummingbird.
(26) The portrait is smaller than a coin.
(27) The coin is smaller than a portrait.
(28) The microchip is smaller than a grain of sand.
(29) The grain of sand is smaller than a microchip.

The third experiment included a mandative con-

No.	Target	Construction
1	Attribution of (un)desirable characteristics through comparison	[noun] *is barely louder than* [noun]
2		[noun] *is smaller than* [noun]
3	Negative truth commitment	*I recommend that* [noun] [verb]
4	Adequate quantities	*I sharpened* 100 *colored pencils (multiple different brands, varied shapes) and this sharpener* [*only*] [*took/ate* or *broke*] [numeral] *tip*[*s*] *off* [*a*] *pencil*[*s*].

Table 1: Summary of the micro-experiments.

struction, i.e. a construction that implies a negative truth-commitment of the dependent clause. For example, in (30) the opinion holder expresses a recommendation that the cashier should smile at every customer, which has an implicature that the cashier did not smile at every customer in the moment of their interaction. (30) was classified as negative by all the models. More sentences with this mandative construction (31–35) were tested for a closer analysis.

(30) I recommend that the cashier smile at every customer.

(31) I recommend that the dishes be washed thoroughly.

(32) I recommend that the chef add more salt.

(33) I recommend that the producer use durable materials.

(34) I recommend that the company prioritise quality.

(35) I recommend that the seller communicate politely.

As a result of the micro-experiment, the sentences (31–35) were mostly classified as positive by DistilBERT, BERT, and SiEBERT, and neutral by RoBERTa. (32) was classified as negative by SiEBERT and (34) by DistilBERT, but both look more like anomalies. Again, assigning a neutral label can be also counted as the correct answer if the sentences are analysed more formally. Otherwise, the models seem to fail recognising the implication of a negative truth-commitment, and simply react to such positive markers as *recommend* (30-34), *smile* (30), *thoroughly* (31), *durable* (33), *quality* (34), *politely* (35).

The fourth experiment is about the sense of adequate quantity. Oftentimes, people express implicit evaluation by mentioning the quantities, which correspond to be normal or abnormal in certain situations. In the variations of sentence (36), the number of *tips eaten off* by the sharpener equals to 5, 10, 25, 50, 75, 90, and 100. The original sentence for this experiment was taken from a real product review. All versions were reproduced without the word *only* to discover if this is a formal negative marker of insufficiency. In addition, all these configurations were reproduced with the alternation of the verb: *took/ate* was changed to *broke*. Experiments with number in the versions of sentence (36) demonstrated that the change of the number did not influence the classification process. The models demonstrated a great disagreement again. DistilBERT labelled everything but three seemingly arbitrary sentences as negative. All sentences with the construction *only took/ate* were marked as neutral by RoBERTa, negative by BERT, and positive by SiEBERT. All sentences with the construction *took/ate* without only were labelled as neutral by RoBERTa, positive by BERT, and negative by SiEBERT. The sentences with construction *only broke* were classified as positive and neutral by RoBERTa, as exclusively positive by SiEBERT, and as negative by BERT. The examples with the word *broke* but without *only* were all labelled as negative by all models with a few exceptions in DistilBERT's output. SiEBERT seems to interpret the sentences cases with *only* as positive and those without *only* as negative. Other models appear to be rather erratic in terms of their reactions to changes.

(36) I sharpened 100 colored pencils (multiple different brands, varied shapes) and this sharpener only took/ate 1 tip off a pencil.

Thus, it has been shown how micro-experiments are able to spot the formal markers that transformer models, sometimes erroneously, base their decisions on. For example, the words and constructions *barely louder than*, *smaller than*, *only*, *broke*, *rec-*

ommend, *politely* and others appeared to serve as formal sentiment markers that defined the polarity chosen by the models regardless of the context and the pragmatic common sense interpretation.

4 Conclusion

This paper demonstrates how linguists can contrast qualitative pragmatic analysis with models' orientation to formal markers. Highlighting the discrepancies between these two approaches might be useful in understanding the limitations of the language models based on supervised machine learning.

Limitations

This short paper is not a quantitative empirical study and should not be treated as one. It is not meant to provide any conclusions regarding the quality of concrete models. The micro-experiments presented do not constitute an exhaustive list of possible angles for exploring discrepancies between human perception and the cues that transformer models take into account. Instead, they exemplify a new perspective on the use of pragmatics in model evaluation.

References

Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis Espinosa Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *CoRR*, abs/2010.12421.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Canonical Model Maintainers HuggingFace. 2022. distilbert-base-uncased-finetuned-sst-2-english (revision bfdd146).

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

NLPTown. 2023. bert-base-multilingual-uncased-sentiment (revision edd66ab).

Janyce Wiebe and Lingjia Deng. 2014. A conceptual framework for inferring implicatures. In *WASSA@ACL*.

A Detailed Results of the Micro-Experiments

Table 2: Detailed results of Micro-Experiment 1.

Sentence	**Intended label (if present)**	**DistilBERT**	**roBERTa (CardiffNLP)**	**BERT (NLPTown)**	**SiEBERT**
This speaker is barely louder than my phone.	NEG	NEG	NEU	3 (NEU)	NEG
This phone is barely louder than my speaker.	–	NEG	NEG	2 (NEG)	NEG
This stereo system is barely louder than a music box.	NEG	NEG	NEG	2(NEG)	NEG
This music box is barely louder than a stereo system.	NEG	NEG	NEU	2(NEG)	NEG
The parrot is barely louder than a fish.	POS	NEG	NEG	2 (NEG)	NEG
The fish is barely louder than a parrot.	NEG	NEG	NEG	2 (NEG)	NEG
The keyboard is barely louder than the heartbeat.	POS	NEG	NEG	2 (NEG)	NEG
The heartbeat is barely louder than the keyboard.	–	NEG	NEG	2 (NEG)	NEG
The car engine is barely louder than a fridge.	POS	NEG	NEG	2 (NEG)	NEG
The fridge is barely louder than a car engine.	NEG	NEG	NEG	2 (NEG)	NEG
The neighbours are barely louder than library visitors.	POS	NEG	NEU	2 (NEG)	NEG
The library visitors are barely louder than the neighbours.	–	NEG	NEU	2 (NEG)	NEG

Table 3: Detailed results of Micro-Experiment 2.

Sentence	**Intended label (if present)**	**DistilBERT**	**roBERTa (CardiffNLP)**	**BERT (NLPTown)**	**SiEBERT**
The shower is smaller than a phone booth	NEG	NEG	NEU	3 (NEU)	NEG
The phone booth is smaller than a shower.	–	NEG	NEU	3 (NEU)	NEG
The throne is smaller than a highchair.	NEG	NEG	NEU	3 (NEU)	NEG
The highchair is smaller than a throne.	NEU	NEG	NEU	3 (NEU)	NEG
The pocket is smaller than a matchbox.	NEG	NEG	NEU	3 (NEU)	NEG
The matchbox is smaller than a pocket.	NEU	NEG	NEU	3 (NEU)	NEG
The hummingbird is smaller than a teacup.	NEU	NEG	NEU	3 (NEU)	NEG
The teacup is smaller than a hummingbird.	NEG	NEG	NEU	3 (NEU)	NEG
The portrait is smaller than a coin.	NEU	NEG	NEU	2 (NEG)	NEG
The coin is smaller than a portrait.	NEU	NEG	NEU	2 (NEG)	NEG
The microchip is smaller than a grain of sand.	POS	NEG	NEU	2 (NEG)	NEG
The grain of sand is smaller than a microchip.	NEU	NEG	NEU	2 (NEG)	NEG

Table 4: Detailed results of Micro-Experiment 3.

Sentence	**Intended label (if present)**	**DistilBERT**	**roBERTa (CardiffNLP)**	**BERT (NLPTown)**	**SiEBERT**
I recommend that the cashier smile at every customer	NEG	POS	POS	5 (POS)	POS
I recommend that the dishes be washed thoroughly.	NEG	POS	NEU	4 (POS)	POS
I recommend that the chef add more salt.	NEG	POS	NEU	4 (POS)	NEG
I recommend that the producer use durable materials.	NEG	POS	NEU	4 (POS)	POS
I recommend that the company prioritise quality.	NEG	NEG	POS	4 (POS)	POS
I recommend that the seller communicate politely.	NEG	POS	NEU	4 (POS)	POS

Table 5: Detailed results of Micro-Experiment 4. Part 1.

Sentence	**Intended label (if present)**	**DistilBERT**	**roBERTa (CardiffNLP)**	**BERT (NLPTown)**	**SiEBERT**
I sharpened 100 colored pencils (multiple different brands, varied shapes) and this sharpener only took/ate 1 tip off a pencil.	POS	NEG	NEU	1 (NEG)	POS
[...] this sharpener only took/ate 5 tips off pencils.	POS	NEG	NEU	1 (NEG)	POS
[...] this sharpener only took/ate 10 tips off pencils.	–	NEG	NEU	1 (NEG)	POS
[...] this sharpener only took/ate 25 tips off pencils.	–	NEG	NEU	2 (NEG)	POS
[...] this sharpener only took/ate 50 tips off pencils.	NEG	NEG	NEU	2 (NEG)	POS
[...] this sharpener only took/ate 75 tips off pencils.	NEG	NEG	NEU	2 (NEG)	POS
[...] this sharpener only took/ate 90 tips off pencils.	NEG	NEG	NEU	1 (NEG)	POS
[...] this sharpener took/ate 1 tip off a pencil.	POS	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 5 tips off pencils.	POS	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 10 tips off pencils.	–	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 25 tips off pencils.	–	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 50 tips off pencils.	NEG	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 75 tips off pencils.	NEG	NEG	NEU	5 (POS)	NEG
[...] this sharpener took/ate 90 tips off pencils.	NEG	NEG	NEU	5 (POS)	NEG

Table 6: Detailed results of Micro-Experiment 4. Part 2.

Sentence	**Intended label (if present)**	**DistilBERT**	**roBERTa (CardiffNLP)**	**BERT (NLPTown)**	**SiEBERT**
I sharpened 100 colored pencils (multiple different brands, varied shapes) and this sharpener only broke 1 tip off a pencil.	POS	NEG	POS	1 (NEG)	POS
[...] this sharpener only broke 5 tips off pencils.	POS	NEG	POS	1 (NEG)	POS
[...] this sharpener only broke 10 tips off pencils.	–	NEG	POS	1 (NEG)	POS
[...] this sharpener only broke 25 tips off pencils.	–	NEG	NEU	1 (NEG)	POS
[...] this sharpener only broke 50 tips off pencils.	NEG	NEG	NEU	1 (NEG)	POS
[...] this sharpener only broke 75 tips off pencils.	NEG	NEG	POS	1 (NEG)	POS
[...] this sharpener only broke 90 tips off pencils.	NEG	NEG	NEU	1 (NEG)	POS
[...] this sharpener broke 1 tip off a pencil.	POS	NEG	NEG	1 (NEG)	NEG
[...] this sharpener broke 5 tips off pencils.	POS	POS	NEG	1 (NEG)	NEG
[...] this sharpener broke 10 tips off pencils.	–	NEG	NEG	1 (NEG)	NEG
[...] this sharpener broke 25 tips off pencils.	–	NEG	NEG	1 (NEG)	NEG
[...] this sharpener broke 50 tips off pencils.	NEG	POS	NEG	1 (NEG)	NEG
[...] this sharpener broke 75 tips off pencils.	NEG	POS	NEG	1 (NEG)	NEG
[...] this sharpener broke 90 tips off pencils.	NEG	NEG	NEG	1 (NEG)	NEG