10 Challenges of sentiment analysis and how to overcome them Part 3

AUTHORS Paul Simmering, Thomas Perry

9 February 2023

A guide for evaluation of sentiment analysis solutions for market research: What to watch out for when choosing sentiment analysis software

Analytics 6 min read Share

ARTICLE SERIES

10 Challenges of sentiment analysis

Previously we've looked at what sentiment analysis is and the challenges (both general and specific) those using it face. Now we look at the challenges of grouping and analysing sentiment data.

7. Aggregation

Aggregation means the way sentiment expressions are summarised across larger amounts of text.

Transformer models operate on the sentence or paragraph level. To answer a research question, the results need to be aggregated further to the document level or to a collection of documents. Let's take an example: A product manager asks the question: "What is the overall consumer sentiment towards my product?"

The model starts with an analysis of the first review. The reviewer mentioned 5 distinct positive aspects and 7 distinct negative aspects. A naïve approach is to add the aspects as +1 and –1 terms, but this approach does not consider that some arguments are more important than others. The measurement needs to represent the author's final sentiment or vote. This could be possible if the model can detect a final evaluation by the author ("overall, I think the product is good/bad"). But then it needs to be able to differentiate this final sentiment from the others. Or the model needs to be able to understand the author's priorities – which would be an even harder task than finding sentiment expressions and their targets.

The way to resolve this dilemma depends on the research question at hand. Given enough reviews, counting the number of texts that contain more positive than negative aspects could be appropriate because the product manager inquired about overall sentiment. In other cases, the sum of positive and negative mentions grouped by aspect could be more meaningful.

Whatever aggregation method is used, it must be explained to the audience. This is easier if the model returns fine-grained predictions in the opinion holder à sentiment expression à aspect à target format we discussed previously. A model that only returns a single number in regard to a whole text suffers from a lack of clarity in aggregation.

8. Insufficient intercoder reliability

To learn how to correctly judge sentiment, a model needs a lot of training material to learn from and benchmark against. These examples must be representative of the data the model will analyse in production use.

The training material needs to be created by analysts. For this, the analysts need an annotation scheme that details which types of words or sentences should be labelled. As an example, an annotation scheme could instruct an analyst to label all adjectives with "ADJ". Further, the analysts could be instructed to skip examples where they're unsure about the right label.

The topic of annotation schemes and analyst instructions has not received as much as attention as advances in model architectures. Machine learning research typically runs new models on a benchmark dataset and iterates on the architecture while keeping the dataset fixed. This is perfect for comparing architectures for research purposes. But outside of academia, collecting the right training data specific to the application domain is a critical step that determines the accuracy of the model.

Collecting training data has two main challenges:

1. **Volume**: the sheer amount of material one needs. While dozens of companies recruit online workers for annotation tasks, analyst training and quality control are a concern. Chen (2022) illustrates this with an analysis of Google's "GoEmotions" dataset, which has sentiment labels for 58k Reddit comments. Chen's conclusion: 30% of the dataset is mislabeled, primarily due to the annotator's lack of familiarity with internet slang on Reddit and relevant pop-culture references.
2. **Inter-coder reliability**: Because language and sentiment are subject to connotation, interpretation and understanding, it is difficult to ensure that all annotators interpret things the same way.

Both problems are more difficult to solve the longer the texts are, and the more nuance is needed (e.g., annotating "good" and "very good" instead of just "good"). Complexity negatively influences inter-coder reliability. But also, the content can make a difference, e.g., when complex or conflicting emotions are being described or expressed.

Reviews are generally the easiest media type for sentiment analysis, as their primary intent is to convey sentiment regarding a specific product or service. Social media comments are harder due to unclear author intent, use of slang and the back and forth between participants in a discussion.

Textual complexity can be met with more sophisticated annotation schemes and analyst training, but these measures increase cost and require expert analysts. Training and rule systems have limits, as illustrated by a study by Weber et al. (2018), where these investments showed little benefit in a content analysis task. In the study of van Atteveldt, van der Velden and Boukes (2021), a gold standard training set was created by a group discussion of all items that analysts disagreed on until consensus was reached, a process that we've also found to be effective, though slow.

From our experience, iteration and tight quality control is the key to inter-coder reliability. The first step is to develop an annotation schema that all analysts understand and can apply and which handles almost every conceivable text. It is unlikely that the first draft of the schema works, so it is better to create a schema, annotate some examples and then sharpen the schema. To increase annotation quality, it is key to have multiple annotators check each example and review cases of disagreement. Finally, the data can be checked for errors by using it to train a model and check cases that the model has trouble understanding. In machine learning terms, these are cases where the model predicts a different label than the analyst gave the text. Often, these turn out to be a mistake in the data rather than a fault of the model.

In part 4, we'll help you understand the importance of training material and methods for sentiment analysis.

Analytics Share

Paul Simmering
Data Scientist at Q Agentur für Forschung GmbH

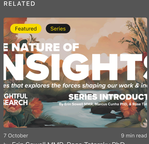
Thomas Perry
Managing Director at Q Agentur für Forschung GmbH

ARTICLE SERIES

10 Challenges of sentiment analysis

- 10 Challenges of sentiment analysis and how to overcome them Part 1
- 10 Challenges of sentiment analysis and how to overcome them Part 2
- 10 Challenges of sentiment analysis and how to overcome them Part 3
- 10 Challenges of sentiment analysis and how to overcome them Part 4

RELATED



7 October 9 min read
by Erin Sowell, MMR, Rose Tatarsky, PhD, Marcus Cunha Jr, PhD

The Nature of Insights: Series Introduction

2 March 2022 6 min read
by Ranbir Ivanic, Adrian Del Bosque, Sarah Parker

The future of AI in market research

8 February 2022 9 min read
by Crispin Beale

The evolution from social listening to digital...

ESOMAR NETWORK

ESOMAR Corporate | ESOMAR Directory | Resource Library by ANA | Trust in Research | Foundation | Job Board