Understanding Customer Sentiment with NLP from Sparse Labelled Data

Daniel Perruchoud and **Joseph Weibel**
daniel.perruchoud@fhnw.ch

Abstract

Although sentiment analysis of texts is generally considered a solved problem, implementing a solution for real-world applications can pose challenges. On one hand, public data typically differ significantly from messy texts of real-world applications, especially if authored by various people. On the other hand, solutions often rely on pre-trained models primarily available in English, and even if models are available in other languages, they tend to have lower quality. Finally, most solutions require labelled data, whose acquisition costs businesses want to minimize. This project systematically investigates methods to address and mitigate these challenges, specifically focusing on how and to what extent the amount of labelled data can be reduced. We develop a sentiment analysis solution for a small Swiss bank using transformer-based models. We use binary classification, since the primary focus is the identification of early signs of negative customer experience, which the bank wants to address. For that, we raise over 10'000 client advisor notes of individual customer contacts, with 300 texts manually annotated by two employees. The notes consist of short German sentences (length mean/std. dev.: 162/105 chars) without fixed structure. For comparison, we also apply our approach to several publicly available datasets. To identify sentiment in these texts, we tried 49 strategies using LLM in-context learning via Mistral 7B, Mixtral 7x8B, Llama 2 and 3, and different prompting strategies including Zero-Shot, Few-Shot, Chain-of-Thought and Reasoning. The experiments show, that given an appropriate prompt, smaller models achieve similar performance as larger models. However, larger models generally encounter fewer issues with different prompt styles. Furthermore, we utilize the k-nearest neighbours (kNN) algorithm and sentence-transformer embeddings for text similarities to get sentiment labels. We vary amounts of labelled data to find optimal parameterization. The most accurate labels are obtained when considering at least 150 samples, at least three neighbours and weighting the neighbours' labels based on similarity (F1: $\approx$0.80). Sentiment identified by LLM is accurate, but also computationally expensive in terms of compute time and infrastructure required for inference. Fine-tuning BERT models with weak labels from LLMs is a method to transfer this knowledge into simpler models. We apply self-supervised learning by fine-tuning various pre-trained BERT models with 10'000 weak labels. Best results are achieved with weak labels based on a Few-Shot prompt with six hard labels/examples and Mistral 7B or Llama 2 7B. The results are slightly worse (F1: 0.76) than using the LLM directly for classification (F1: 0.78), but the number of parameters is reduced drastically from 7B to 0.11B. If instead weak labelling with weighted kNN (k=8) is used for BERT fine-tuning, 40 hard labels are needed to reach the same quality. Also, only one-fifth (32 instead of 160) of the hard labels are necessary to achieve an equally good model than just training on all hard labels (F1: 0.73). By using all 160 hard labels for finding similar texts, the quality of the fine-tuned BERT model can even be further improved (F1: 0.82). Applying our approach to several publicly available English datasets including financial news headlines, product and movie reviews, we consistently found results superior to the ones reported above for the bank. This indicates that published results cannot be directly transferred to real-world scenarios. One reason is that publicly available pre-trained models for German show lower performance. Additionally, quantitative analyses show that the bank's client advisor notes employ a broader vocabulary than e.g. financial news headlines and therefore require more data to generalize. Moreover, labelling is not straightforward for humans, as a large portion of the texts does not contain sentiment and abundant neutral texts complicate the identification of negative texts. Through our systematic evaluation, we were able to examine the factors contributing to the quality differences between the datasets and determine which weak-labelling strategy yields the best results in each case. We plan to publish our code so that tests can be conducted for other real-world scenarios.