ABSTRACT

Online product reviews have a substantial impact on purchasing decisions, which makes manipulative or artificially generated reviews an increasingly relevant problem. The rising availability of powerful LLMs enables the generation of synthetic reviews in high quality and large scale, further complicating the detection of opinion spam. The aim of this thesis is to create a high-quality, labeled dataset of German product reviews consisting of both real and LLM-generated texts, which can serve as a research basis for the development of detection methods.

For this purpose, genuine Amazon product reviews as well as synthetic reviews generated by different LLMs via few-shot prompting and fine-tuning were used. Stratified sampling was applied to ensure that the statistical distributions of both subsets are comparable with respect to categories, rating scales and text length. The quality of the final dataset was assessed both statistically and linguistically. To evaluate the realism of the generated reviews, two classification models (logistic regression and RoBERTa) were trained and compared.

The results show that RoBERTa achieves an accuracy of 90 %, outperforming the linear model in distinguishing between real and generated reviews. The hardest to detect generated reviews originate from the fine-tuned *GPT-4.1 Nano* model. Overall, this thesis demonstrates that the constructed dataset exhibits realistic distributions and is suitable for the evaluation of modern classification models. The resulting dataset closes an existing research gap in the German language domain and provides a basis for further work in automated opinion-spam detection.

Keywords: Opinion Spam, Large Language Models, Few-Shot Prompting, Fine-Tuning, RoBERTa, Logistic Regression, German-Language Dataset

ZUSAMMENFASSUNG

Online-Produktrezensionen haben einen erheblichen Einfluss auf Kaufentscheidungen, wodurch manipulierte oder künstlich erzeugte Bewertungen zu einem wachsenden Problem werden. Die zunehmende Verfügbarkeit leistungsfähiger LLMs ermöglicht es, synthetische Rezensionen in hoher Qualität und großer Menge zu generieren, was die Erkennung von Opinion Spam zusätzlich erschwert. Ziel dieser Arbeit ist die Erstellung eines qualitativ hochwertigen, gelabelten Datensatzes aus deutschsprachigen Produktrezensionen, der sowohl reale als auch durch LLMs generierte Texte umfasst und als Forschungsgrundlage für die Entwicklung von Detektionsverfahren im Bereich Opinion Spam dienen kann.

Hierzu wurden echte Amazon-Produktrezensionen sowie synthetische Rezensionen, die durch verschiedene LLMs mittels Few-Shot-Prompting und Fine-Tuning generiert wurden, verwendet. Mittels stratifizierten Samplings wurde sichergestellt, dass die statistischen Verteilungen beider Teilmengen hinsichtlich Kategorien, Bewertungsskalen und Textlängen vergleichbar sind. Die Qualität des finalen Datensatzes wurde sowohl statistisch als auch im Bezug auf sprachliche Diversität bewertet. Zur Prüfung der Realitätsnähe der generierten Rezensionen wurden zwei Klassifikationsmodelle (logistische Regression und RoBERTa) trainiert und miteinander verglichen.

Die Ergebnisse zeigen, dass RoBERTa mit einer Genauigkeit von 90 % besser zwischen realen und generierten Rezensionen unterscheiden kann als das lineare Modell. Die am schwersten erkennbaren generierten Rezensionen stammen dabei vom feinabgestimmten *GPT-4.1 Nano*-Modell. Insgesamt wurde deutlich, dass der erstellte Datensatz realistische Verteilungen aufweist und sich zur Evaluierung moderner Klassifikationsmodelle eignet. Die bereitgestellte Datengrundlage schließt damit eine Forschungslücke im deutschsprachigen Raum und bildet die Basis für weiterführende Arbeiten im Bereich der automatisierten Opinion-Spam-Detektion.

Schlüsselwörter: Opinion Spam, Large Language Models, Few-Shot-Prompting, Fine-Tuning, RoBERTa, logistische Regression, deutschsprachiger Datensatz