

Sentiment Analysis of Self-Published vs. Traditionally Published Books using Machine Learning

Jayasundara J.M.G.N.^{1*}, Adeeba S.¹

¹Department of Computing and Information Systems,
Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka
**jmgjayasundara@std.appsc.sab.ac.lk

The rapid growth of the self-publishing channels, such as Amazon Kindle Direct Publishing (KDP), has greatly changed the model of distribution of books across the world today because authors are able to bypass the traditional publishing framework. The conventional publishers have well-established editorial and marketing procedures, but the self-published authors have full creative freedom with more or less quality control. However, in spite of this change, there is a shortage of academic studies that use computational sentiment analysis to compare the perception of books by the readers under these two models in a systematic way. In this work, this gap is filled by comparing the attitudes of readers to self-published and traditionally published books based on the large dataset of Goodreads reviews. The study aims at (1) determining the patterns of sentiment between the two publishing models, (2) identifying the critical themes that determine the perceptions of the reader, and (3) assessing the contribution of platform visibility and metadata in modulating the trend of sentiment. Text normalization, tokenization, and publisher classification by metadata preprocessed the reviews. In identifying the reviews in self-published and traditionally published classes, a TF-IDF vectorizer and a Logistic Regression classifier were used. This model was able to accomplish an accuracy of 0.80 with a test sample of 125,757. The performance measures showed a precision of 0.82 and a recall of 0.78 for self-published books and a precision of 0.79 and a recall of 0.82 for traditionally published books. Furthermore, a DistilBERT model was used as an additional robustness test. The findings indicate that the sentiment of readers is fairly equal on both publishing models; however, selfpublished books have a greater diversification of sentiment distribution. The consistency of traditional books is probably higher because of the professional editing and the publication organization. The research has implications for those publishing and those being published in terms of marketing approaches, content suggestions and implications to authors in their choice of publication pathway.

Keywords: *Goodreads; Machine Learning; Sentiment Analysis; SelfPublishing; Traditional Publishing;*