Topic: Amazon product review spam detection
Project Abstract

Course: Statistical Natural Language Processing

Student name:

- Khoa Nguyen (khoa.nguyen@aalto.fi)
- Huyen Pham (huyen.pham@aalto.fi)
- Quan Nguyen (quan.nguyen@aalto.fi)

1. Problem & motivation:

E-commerce platforms like Amazon have suffered from fake review problems as some shoppers accept gifts and other incentives in exchange for positive reviews, despite a ban on these activities. Hence, having a model that assists companies in detecting such issues would help customers shop with confidence knowing the reviews they read are authentic and relevant. Engaging in spam detection might also create a competitive advantage for firms regarding customer experience.

2. Dataset:

Link: https://www.kaggle.com/naveedhn/amazon-product-review-spam-and-non-spam
The dataset contains Amazon product reviews with spam and not spam labeling. This is a large corpus that contains 26.7 million reviews and 15.4 million reviewers. The class label is spam and not spam, where "0" indicates not spam and "1" indicates spam reviews.

3. Aim of study:

The aim of the study is to predict whether a review is spam or not by applying different models and evaluating the results from the best model.

4. Methods:

Since this is a classification problem separating spam and non-spam comments and also potentially classification based on which products the comments are for, we think that there are some available models such as BERT or LSTM that are suitable for this task.