Topic modelling online customer reviews

Radoslaw Kowalski[1], Slava Mikhaylov[1], Tom Cusak[2]
[1]University College London, [2]Argos

Project Background

Many published studies inform how to use topic modelling to understand customer review data, but no prior studies attempted to implement a single topic model across a wide range of product categories. Meanwhile, retailers need a solution that would allow them to analyse customer review data on all of the products they sell simultaneously. Retailers' customer reviews may span hundreds of product categories and tens of thousands of products, and topic modelling for individual product categories or products would be very costly. Moreover, it would be hard to compare topic model results between product categories as each model produces a different set of topics. Therefore, a single topic model for all product categories can give retailers more insight than many topic models for individual product categories or products.

Data and Methods

Latent Dirichlet Allocation (LDA) model with Gibbs sampling was used in this study to see whether topic model results can be a reliable tool for summarising customer reviews across product categories. The study used 650,000 customer reviews and user-generated sentiment scores obtained from Argos plc. The analysis answered the research question: How to implement topic modelling on reviews of a wide range of products or services simultaneously? The tested hypotheses were:

1) A combination of select parts of speech is best for identifying distinct product features and benefits
2) There is an optimum number of topics to be found in customer reviews

Key findings

The first two hypotheses were evaluated through inspection of the top 5 words to see if the words signified identifiable product benefits and drawbacks. Both hypotheses were confirmed. Optimum variants of LDA have been identified for the choice of parts of speech, n-grams and target number of topics. The third hypothesis was not confirmed. Weighted average of sentiment scores for topics cannot effectively measure sentiments associated with each topic. Nonetheless, LDA topics representing product benefits have the same meaning across product categories because LDA takes words' context into account.
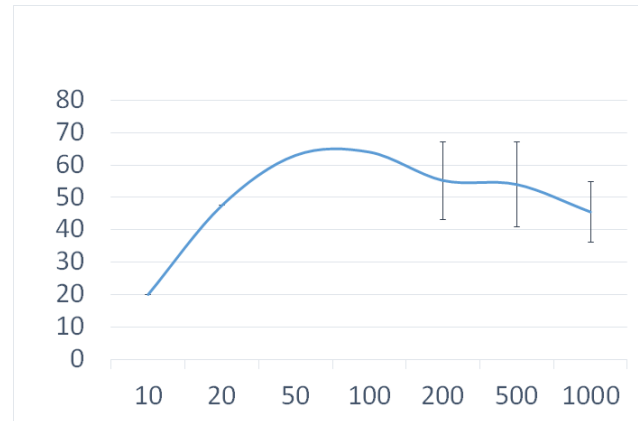
Accuracy of topics (%) depending on the number of topics generated by an LDA topic model

Value of the Research

Future topic modelling analyses need not be constrained by arbitrarily defined product categories. Topic modelling can be carried out across product categories without sacrificing the quality of modelling results. Research findings suggest that retailers can use topic modelling to obtain new insights. For example, topic model composition can be compared between products and product groups to discover possible correlations with sales as well as with customers' demographics and online behaviour patterns. Moreover, analysts can freely investigate the topic structure from customer reviews according to their own preferences (e.g. organizing reviews according to posting date as opposed to membership in a product category) without need to recalculate the topic model again.