Unlocking the power of web text data

December 8 2017

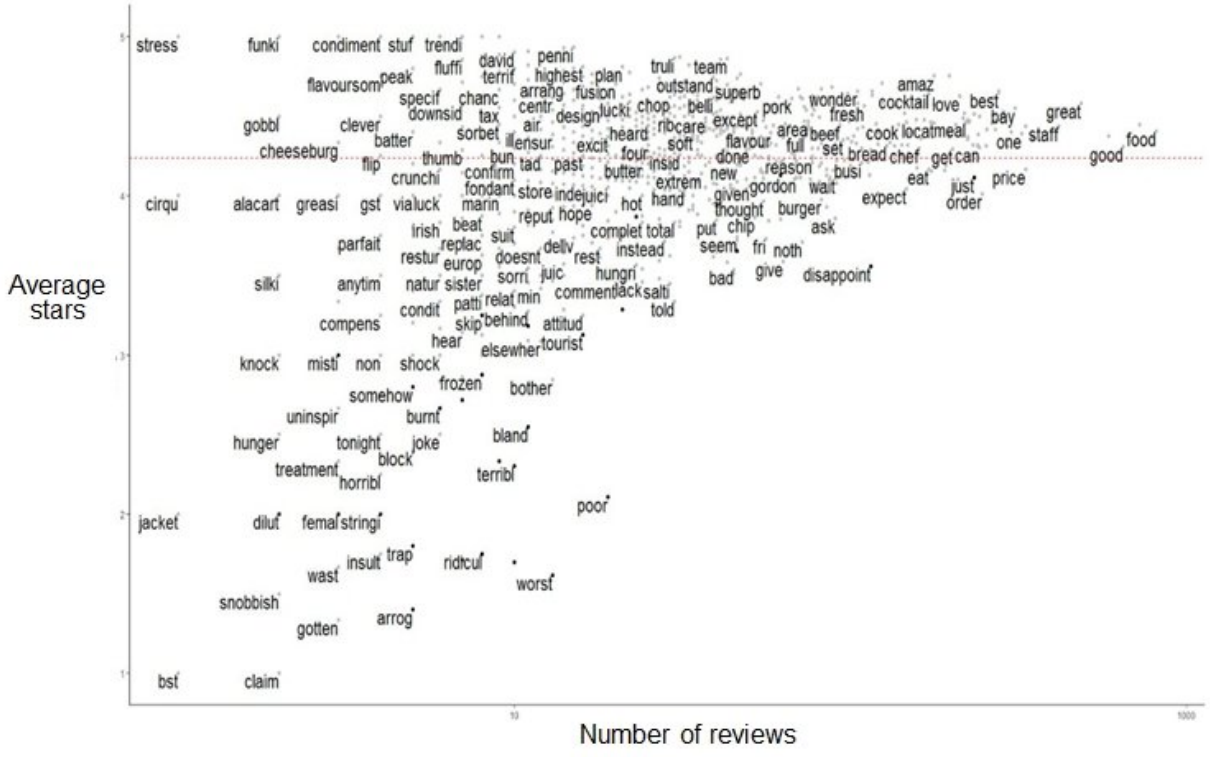

Figure shows the word features which were found in online customer ratings and reviews for nine major restaurants managed by a hotel chain in Singapore. Each word feature in the corpus is displayed as a grey dot. The features selected by the RTL classifier are highlighted as black dots. Credit: National University of Singapore

NUS statisticians have developed the Regularised Text Logistic (RTL) regression model to extract informative word features from digital text

for decision-making.

The world is increasingly becoming connected through the internet and social media applications, creating vast amounts of data. With the massive increase in web posts, user reviews and feedback around the world via electronic word-of-mouth, web text data has been shown to provide important information for content analysis, as well as create an impact on decision-making processes. Businesses and organisations need to be able to analyse and make sense of data to remain competitive and relevant.

Prof CHEN Ying from the Department of Statistics and Applied Probability, NUS and her research team have developed a text mining and analysis model which can identify and extract informative textual data of interest automatically from public postings on the internet (e.g. social media comments etc). This is known as the Regularised Text Logistic (RTL) regression model.

Online web textual data comes from many distributed sources and is often unstructured. This makes it difficult to analyse using conventional approaches. The RTL regression is a machine learning classifier that helps to accurately classify customers' review polarity (positive or negative) based on the textual content. It is also capable of automatically detecting a small set of informative word features that help business decision-makers pinpoint the key aspects of customer reviews easily.

Prof Chen said, "This automated feature saves time which would otherwise be spent reading the review information online. With this feature, business decision-makers can obtain immediate feedback on customer sentiments towards their products or services, so that they can tailor their offerings to improve the customer experience."

"From our knowledge, the RTL model is the first supervised sentiment

classifier for large amount of web-based text using the logistic regression framework with theoretical derivation," added Prof Chen.

More information: P Liu; Y Chen; CP Teo, "Sentiment Analysis for Online Reviews with Regularized Text Logistic Regression" working paper (2017).

Provided by National University of Singapore