Data-driven ranking and visualization of products by competitiveness

Sheema Usmani, Mariana Bernagozzi, Yufeng Huang,
Michelle Morales, Amir Sabet Sarvestani, Biplav Srivastava
IBM Chief Analytics Office,
Armonk, NY, USA 10504

Abstract

Competitive analysis is a critical part of any business. Product managers, sellers, and marketers spend time and resources scouring through a huge volume of online and offline content, aiming to discover what their competitors are doing in the marketplace and to understand what type of threat they pose to their business' financial well-being. Currently, this process is slow, costly and labor-intensive. We demonstrate *Clarity*, a data-driven unsupervised system for assessment of products, which is currently in deployment at IBM. *Clarity* has been running for more than a year and is used by over 1,500 people to perform over 160 competitive analyses involving over 800 products. The system considers multiple factors from a collection of online content: numeric ratings by users, sentiment towards key product drivers, content volume, and recency of content. The results and explanations of factors leading to the results are visualized in an interactive dashboard that allows users to track their products performance as well as understand the main contributing factors.

Introduction

Every business wants to know how their products are doing in comparison to its competition. In the field of market research, this has mostly been a manual process with researchers scanning through a large volume of online and offline content, keeping track of key drivers of interest for each product, deciding whether a mention represented positive or negative feedback by manually annotating each mention etc. Researchers would repeat this process for every additional public domain forum, for every new driver. It is time and labor-intensive, error-prone, slow and costly process. Furthermore, as competition and feedback from users continue to evolve, any analysis done previously needs to be frequently updated to ensure accuracy.

Recently, there have been concerted efforts towards automating parts of the above process by leveraging natural language processing and machine learning. In Joung et al. (2018), the authors use text mining methods to analyze customer complaints and find gaps in the company's products. In Afful-Dadzie et al. (2014), the authors perform text analysis on user comments posted on social media to compare telecommunication providers in Ghana. In (Bhatt, Mcneil, and Patel 2014), the authors track general sentiment overtime for products by calculating a sentiment score based on user-generated content such as reviews and comments.

Our system builds upon previous work by introducing a novel competitive metric (shown in Figure 1) that encompasses sentiment towards key drivers as one of its contributing factors. The competitive scores and contributing factors are visualized in an interactive dashboard (described through a running example in the video) that allows users to track their products performance. Due to business reasons, product names are anonymized in the demonstration.

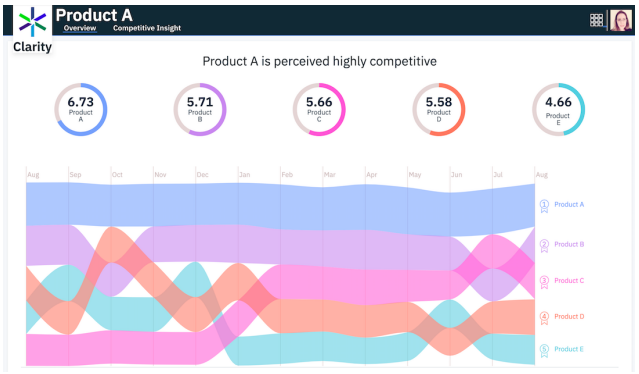

Figure 1: Sorted stream graph to visualize products competitiveness. X-axis: time, Y-axis: product rank and thickness of line: absolute competitiveness score.

System Overview

The main steps of Clarity are:

1. Prepare review data of products p_1 to p_N from sources d_1 to d_M (offline)
2. Process request for analysis for product p_i (online)
3. Visualize analysis results (online, optional)

The steps involved in the computation and visualization of the competitive score for a product are described below.

Data preparation As input, the system aggregates unstructured text from an undisclosed set of public forums and review sites, where comments are widely shared by users

and experts. The system filters the aggregated data by product names provided by the user, to identify the relevant content for each product in the inputted list. The first component of our system prepares and processes the textual data by tokenizing the document into words, applying Word2Vec model to find the numerical representation of each word and saving the corresponding word vector. We also record the number of reviews/posts and star ratings.

Analysis As part of the second step, the system generates the *Clarity Score*, which denotes the perceived competitiveness of the product. The score is based on the number of online reviews, numeric rating, sentiment towards key drivers of the product and recency of content. The first two factors are calculated as part of data preparation.

The calculation of the average sentiment towards each driver includes an NLP engine which processes the text in order to understand how the product is performing across any number of drivers/topics. Using the text as input the engine will extract keywords as well as calculate the targeted sentiment towards those particular keywords, using the Watson Natural Language Understanding API. The pre-trained word vector model is then used to convert the keywords as well as the dimensions to word vectors. After the words are converted to word vectors, distance measures (e.g. cosine distance) are used to determine which keywords are most related to which driver. Distance metrics, as well as tuned thresholds, are used to assign keywords to particular dimensions. After assignment, the average sentiment is calculated for each dimension.

The system scores each factor using its percentile score, computed using the average value of the factor for a particular product x compared against the entire distribution of values of that factor for all products. This percentile score will serve as a score for how that particular factor is performing as compared to the competition.

Percentile scores for each factor and data source are then combined via a weighted sum, where the weights represent the count or volume proportion for that particular data source, as compared to the other data sources. Use of percentiles scores account for the difference in the distributions of the key factors among different data sources. This process will be performed various times using different time frame windows. To reflect the changing product life cycle phases, scores can be computed on 3-month frames, dating back as far back as 18 months. Then to combine across time linear weights are used, to weight more recent time frames higher than past frames. After aggregating across time frames, the system will output one score per product, which holistically represents its competitiveness across the key factors considered. The system can also show the score over time.

User Interface To better visualize the changes of the scores over time, we used a variant of the stream graphs called *sorted stream graphs*, (see Figure 1), which more appropriately convey the intended insights.

The x-axis, naturally, represents time; each stream represents a product; and the height of each stream represents the score of the product at that particular point in time. The visualization is highly interactive, allowing the user to highlight a stream to better understand the historical changes of its ranking. Also, hovering over the streams brings up a *tool tip* displaying the exact score at that point in time as well as the deltas for the percentage change of the score and change in ranking with regards to previous month.

This visualization is presented in the context of a web dashboard. The dashboard also showcases other visuals that convey detailed information about the contributing factors. All the dashboards visualizations, including the one described in this article, were created using Data-driven documents (D3) (Bostock, Ogievetsky, and Heer 2011)

Along with the visualization, *Clarity* supports invocation via Application Programming Interfaces (APIs). This enables the core capability of data-driven comparison of products to be integrated and reused in various applications.

Conclusion

In this demonstration, we considered the problem of comparing products in a marketplace automatically from online content. This is an important business activity that marketers, sellers and product managers conduct regularly. Unfortunately, currently it is mostly a manual, time consuming and costly process which can be particularly challenging for businesses with large product portfolios and fast-changing customer environment. In response, we presented *Clarity*, an unsupervised data-driven system for assessment of products. Our contributions are the following:

- A novel unsupervised approach to assess the competitiveness of products in a marketplace along factors learned from data.
- A novel approach to explain factors affecting competitiveness score of products and visualization of the results.
- Evaluation of the deployed system, *Clarity*, showed that our system is aligned with the competitive analysis by the market experts (such as Gartner Magic Quadrants and Net Promoter Score).

Clarity has thus proven to be an excellent example of an AI-based system that has been integrated and reused in various applications such as product pricing recommendation, and talent management and has performed extremely well in critical business activities.

References

Afful-Dadzie, E.; Nabareseh, S.; Oplatková, Z. K.; and Klimek, P. 2014. Enterprise competitive analysis and consumer sentiments on social media - insights from telecommunication companies. In *DATA*.

Bhatt, D. A.; Mcneil, K. E.; and Patel, N. A. 2014. Time-based sentiment analysis for product and service features. In *Journal Software, VOL. 9, NO. 2,*.

Bostock, M.; Ogievetsky, V.; and Heer, J. 2011. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Joung, J.; Jung, K.; Ko, S.; and Kim, K. 2018. Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development. *Sustainability* 11(1):1–14.