Richard Muñoz
EECS 6898
Week 4: Oded Netzer

Mine Your Own Business: Using Text Mining in Business Applications

Dr. Netzer focused his talk on two projects. The first project concerned the application of text mining to analyzing social media in order to infer market structure. The second project concerned the prediction of loan default from the text of loan applications. Dr. Netzer also briefly mentioned two other topics that he did not have time to discuss: automatic idea analysis and predicting side effects of drugs based on adverse event data.[1] Before diving into each project's details, Dr. Netzer briefly motivated the application of text mining to solving business problems.

Dr. Netzer described his research focus as using data to make business decisions using consumer data. Thus, his focus is less on the details of text mining methods and more on the application to the business world. He argued that managers are now enabled with views of lots of data, but opportunities remain to find in the data what competitors have not found. He also argued that more than simple tracking of social media "hits" (e.g., likes and views) is needed in order to have an impact. In particular, the most important statistics need to be identified (through both business judgment and statistics) and focused on. Dr. Netzer shared some success stories of companies using social media to have a direct impact on consumers from companies such as Southwest and Zappos. These companies in particular were able to use their text mining systems to proactively identify consumer issues and respond to them.

Following this introduction, Dr. Netzer described the topic of analyzing social media to infer the market structure for the automobile industry. Specifically, Dr. Netzer applied text mining to approximately 900,000 consumer postings to an online automobile forum (Edmunds.com). He then used the results of the text analysis to analyze and visualize the co-occurrence of brands and models. He also applied similar analysis and visualization to the co-occurrence of brands or models with specific words they are mentioned with. Dr. Netzer enumerated the major issues encountered during the text mining process: handling negation, understanding sentiment, anaphora resolution, and catching meaningful phrases. In addition, he mentioned some marketing-specific difficulties encountered the process: identifying brand and model names. Using handwritten rules largely resolved the identification of brand names. He also stated that while his group has had some success replacing the handwritten rules with learning algorithms in more recent work, there remains an opportunity to continue to reduce the number of rules that are made by humans. In contrast to brand names, Dr. Netzer described the latter as particularly difficult, due primarily to the large variance in names and abbreviations. In either case, I think that we might be able to apply larger corpuses of data to train learning models on. In particular, I think that data sources such as Twitter may be particularly well suited for such model training. I bring up Twitter in particular since each tweet's word limit may encourage abbreviations and variations on the naming of brands and models. An additional challenge for using such general data sources is curating the datasets in order to identify relevant posts and identify the correct brand and model name labels.

[1] As a brief aside, researchers from Columbia University have been in the news for a recently published article on this topic with a focus on identifying drug-drug interactions.

As a result of the text mining, Dr. Netzer created a market structure network diagram with approximately 170 models. A clustering algorithm uncovered three main clusters: compact, family, and luxury cars. A similar map and clustering for approximately 30 brands uncovered three groups: luxury, American, and Japanese/Korean vehicles. A particularly striking result from this analysis was that Cadillac and Lincoln were pushed toward the luxury group and away from the American group, which Dr. Netzer hypothesized correlated with their marketing pivots as an attempt to established themselves along with the luxury cars. Dr. Netzer also described a time-series analysis of the consumer forum and consumer brand switching data that supports this hypothesis. This was an interesting combination of analyzing both data sources, and I think this begs additional investigation to see if additional connections between the text mining analysis and sales data can be uncovered. For example, Dr. Netzer brought up that his group was able to uncover that Honda cars had been experiencing oil problems. It would be interesting to see if the timing and magnitude of the discussion surrounding such problems correlate with sales data. It seems that Dr. Netzer has much of the data that would be necessary to begin to look into this analysis since he has both the consumer forums text and the sales data. There may be additional challenges to identify major car issues in the text and to determine the correct alignment of the sales data with the text information (i.e., depending on the frequency of the data points for the sales data – monthly, semi-yearly, etc.).

The second topic that Dr. Netzer discussed was predicting loan default from the text of loan applications made through the website Prosper.com. These online loan applications contain both traditional loan application measures such credit score, amount requested, demographic information, and a text description from the applicant as to why they are requesting the loan. Dr. Netzer described the benchmark for this task as particularly high, since the credit score is already highly predictive of default. The researchers investigated the performance of logit and decision tree models and they settled on a mixture of two logit models for predicting default. This ensemble achieved an improved predictive ability of 4-5.7% (10-fold CV) over not using the text. Dr. Netzer also investigated the words and writing styles that predict default. He also responded to a student during the presentation that in a separate analysis they controlled for the purposes of the loans. This was my question as well, since I think that the purpose of the loans may have a big effect on the default outcome.

I think that an extension of Dr. Netzer's loan analysis could focus on using information from traditional loan applications (such as notes from a loan officer) to inform the default prediction of the online loan applications. It's likely that companies keep such documents from their loan officers that interview applicants during a traditional loan application process. It might be possible to extract the loan officers' analyses and apply those to the text of the online applications in order to create a "loan officer score." I think that this framework could improve the models' performance, since it theoretically would also be informed by the experience and training of loan officers.