Patrick Rogan
UNI: psr2125

EECS 6898 Homework 3

Dr. Netzer presented on two areas of research: using web-based text mining as a means of generating business insights and using text data as a predictor of whether an individual will default on a loan. He also mentioned two other studies that use text based approaches to identify prescription drug side effects and to assist individuals in their creation of novel, high quality ideas.

Dr. Netzer first discussed his paper Mine Your Own Business: Market-Structure Surveillance Through Text Mining. Being from the business community (as opposed to the computer science/data science field), Dr. Netzer described how business leaders need to understand the process of deriving insights from data and using these insights to make data driven decisions. This is especially true now that managers are exposed to large quantities of data (eg. data gathered from social media), but may not be able to effectively use it to arrive at meaningful insights. He stressed the need for executives to view data as more than an IT issue.

He expanded on this discussion by describing several case studies of businesses using and responding to social media data. For example, Southwest Airlines was able to actively respond to a customer who had Tweeted about the uncomfortably hot environment he was stuck in as his aircraft was waiting for takeoff. Dr. Netzer highlighted how social media influence has shifted power away from companies and into the hands of customers. An example of this is companies using metrics such as Klout Score (a measure of social media influence) to give customers better service in order to prevent far reaching social media outbursts. In many cases businesses are now using data to make better decisions and improve customer satisfaction.

Dr. Netzer then described the specifics of his investigation which sought to answer two main questions: (1) can the web be used as an accurate source of market research data? (2) Can web data be used to infer market structure? These questions were answered by performing text mining on web data and performing network analysis.

There are many opportunities that come from using text mining as a means of gathering market data. For example, there is a large quantity of data that can be obtained passively (no direct contact with customers) which has been timestamped allowing for the study of trends. Unfortunately, given the organic nature of the forum-based data that was mined, the data tends to be messy and it was not initially obvious if the data was representative of the true population of customers or if the topics discussed were relevant.

He then discussed how the text mining process is made possible by vast amounts of automation. Dr. Netzer followed this with an overview of the text mining process, noting that the activity has two main components: the gathering of data and performing basic cleaning and segmentation and a much more difficult component where relevant information is extracted and the questions that originally inspired the research are investigated.

Dr. Netzer described some of the issues in the final two stages of the process. He gave examples of simple and complex sentences that involve multiple products and assessments of quality. An issue with text mining comes from the inability to accurately group nouns and adjectives, understand negation and detect tone, such as sarcastic. While the technology is still improving he noted a young child is still likely to be able to better understand and contextualize sentences. He also mentioned that while identifying brand names is simple (Audi, BMW, Ford, etc.) working with models is much more difficult. Nonetheless, the process his research group used was still able to achieve high levels of accuracy ($\sim$70-98%).

He then described how his team was able to build and associative network of car brands, models and associated attributes. When compared to truth data (obtained from individuals changing the make and model of car they own), the networks built from text mined data and truth were highly correlated.

This lead to him describing a inconsistent point in network-clustering analysis: the Cadillac brand. The investigation into this gave insight showing that the online impression of the automobile had shifted to that of a luxury vehicle at a faster rate than the market as determined by customer car model switching.

Dr. Netzer's work on this study was done by downloading the entire forum from the website edmonds.com. I asked Dr. Netzer what his team does when product and brand data is not so nicely concentrated in one location. He indicated that there is a large, sophisticated and constantly improving set of web scraping tools available. He indicated that while non-centralized data does require these tools and additional processing (the front end of the text mining process) the subsequent work on the cleaned data does not change substantially.

Dr. Netzer then discussed unpublished research that seeks to determine if the text provided by an individual in a loan application can be used to determine if they will default. To accomplish this, data was gathered from prosper.com, an online peer-to-peer lending platform. He briefly described Prosper's loan application process and how the default rate tends to be high ($\sim$33%) when compared to traditional banks (7%).

This leads to the question of whether one can accurately determine if an individual seeking a loan is likely to default. To answer this question, ensemble learning techniques were applied to basic characteristics of the loan applicant (credit score, personal financial information, etc.) along with features extracted from text

provided by the applicant describing their situation and intended use of the loan.

The analysis showed that this approach was able to better predict whether or not an individual would default on a loan than analysis that did not incorporate this text. Dr. Netzer highlighted some of the types of words associated with those who went on to default (words describing hardship, desperation, external influences, and indicators of lying) and words used by those who did not default (these tended to indicate financial literacy and describe a brighter future).

A second approach was used to mine the text provided. Instead of the text as a whole, LIWC (Linguistic Inquiry and Word Count) was used to derive features of the text that were included in a machine learning algorithm. Use of LIWC features also improved the accuracy of the model allowing for better predictions of individual default.

Dr. Netzer concluded by noting that this research seems to be unique in that most text mining tends to be backward looking and focuses around larger concepts (eg. using aggregated sentiment to predict the stock market) while this study can be used to be used to predict future behavior of individuals for years to come.

The research presented focused heavily on using customer data to make informed business decisions. There are, however, other aspects of a business that could potentially gain from the use of some of the text mining techniques mentioned.

While employed in industry in a knowledge worker role, I observed many sources of easily accessible text data. These include emails routed through company servers, instant messages (both individual and group conversations), and internal forms of social media (something like Facebook, but only accessible to employees). It was no secret that this information was captured and stored for various purposes.

Combining such a large store of text data with sentiment analysis from social media presents a business with an opportunity for additional insight. Not only would they be able to examine the health of their organization they could gain a deeper understanding of how top level decisions impact employees and subsequently customers. A few possible applications are:

- **Employee social network graphs:** This would be likely be the simplest tool to construct. Employees who communicated on corporate assets would have connections in a graph. The edge weights would then be determined by metrics such as frequency of communication and length of message. Such a graph could be used to identify employees who do not communicate and provide a means of introduction or other purposes.
- **Evaluating policy impacts on employees:** Sentiment of the company as a whole could be tracked over time. This could be used to form a metric for a morale baseline. Then, the sentiment of employees could be monitored following policy changes (eg. the company increases share repurchasing and dividend payouts while freezing employee raises). After performing this analysis following several policy changes, the company could gain an understanding of how certain actions impact the morale of the company and use this information to gain a more complete understanding of the cost and benefits of implementing a new policy.
- **Evaluating policy impacts on customers:** As mentioned by Dr. Netzer, companies can mine text data from the web to track the market's opinion of a brand or model over time. This could be combined with the company's internal evaluation of employee sentiment to determine if a policy that had a substantial impact (positive or negative) on employees translated to a better or worse product or service being provided to the marketplace.
- **Identification and prevention of toxic environments:** Given that companies have information about what projects and programs their employees are working on, multiple sources of text could be aggregated from everyone belonging certain team. The sentiment of those on the team could be evaluated and tracked over time. Management could then identify teams with decreasing sentiment and take measures (eg. provide incentives, shake up the personnel, etc.) to improve group performance.

One of the drawbacks of these solutions is the attempt to answer fairly complicated questions that have many predictors by evaluating a single sentiment variable. This could result in correlations which may not be indicative of causation, leading managers to draw conclusions and implement changes which may not be in the company's best interest. Nonetheless, by using and refining text mining techniques to assess sentiment both within a company and online, a company may be able to make better informed, data driven decisions.