Jim Sears
ADS652Z2 – Text Mining
Final Report

The insurance industry is one that is increasingly competitive, evolving and adapting. My wife works for an independent, non-profit research firm that gathers data from the entire industry and elsewhere and creates products for their member insurance companies. Their applications of text mining techniques are on the rise and include two interesting products that are currently in development. As they are unreleased and proprietary, the readers' discretion is requested. Due to their sensitive nature, the only request of their project managers was to know the likely inputs and the desired outcomes. These and the plausible techniques incorporated are discussed below.

Candidate Sentiment Analysis

Hiring financial advisors (insurance salespeople) is a laborious and often expensive process that can involve multiple personnel and departments, and sometimes outside consultants. Ensuring that the recruitment process is effective, but also efficient. There are several reasons why a candidate may accept or pass on a position, and most are not actually related to the financials of the offer. Often it is the timing, the expected work-life balance, fit of the company culture, or even the fit of the position for those new to the industry. Then there are the many cases where it was the actual recruitment process.

The way a company handles this process is very insightful for a candidate imagining what their future there may be like. Unresponsiveness, forgetfulness, tardiness, and other negative actions can make the decision easier for the candidate to move on, whereas opposite actions may inspire them to continue. For a national-level insurance company, thousands of candidates may go through the recruitment process on a yearly basis. With the resources allocated to the process, and the competition for quality candidates, it is unacceptable to lose one because of a mismanaged process. It is also difficult to confirm that was the reason.

The product in development centers around determining the basic nature of what candidates have to say about their recruitment experience. Simply put, was it positive or negative? Candidates' posts on social media and job search and recruitment sites such as Indeed and Glassdoor are subjected to a text mining methodology called sentiment analysis which applies a general, yet insightful, label of positive or negative to each post. The approach of doing so is a typical classification model that likely also incorporates word scoring.

A classifier uses supervised machine learning to associate each of the words in a document with its previously and manually labelled class; here the classes are the sentiments: positive or negative. This enables predictions of the sentiment of new and unlabeled posts based on the words they contain. However, a large collection of posts will contain a lot of different words that will result in a multidimensional vectorization that may contain thousands or tens of thousands of features. With this many features, the probability of any word being associated with one class (sentiment) or another is too low to be significant or reliable. The potential solutions are to implement a different and/or additional vectorizer, have enormous amounts of data to outweigh the vectorization's features, or utilize a model that has high bias and low variance. The latter describes a Bayesian classifier, a model that will find the parameters (words) of the collection of posts that maximize the probability of a post being positive or negative. The more often a word is found in known positive posts, the more likely a new post that contains that word will be positive too. And the more often a word is found amongst all posts, the lower the probability of it being an indication of either sentiment; it has diluted its own importance.

This classifier, like any supervised machine learning model, requires a lot of labeled data to be reliable. In this case, that means thousands of posts need to be collected, reviewed, and labeled as positive or negative. As there may not be enough data available, or the personnel and time available to manually process it, a word scoring technique may be implemented to bolster the Bayesian classifier.

The technique of word scoring utilizes an existing list of words that are appropriately scored based on their inherent sentiment. Positive words like “good”, “better” and “best” have positive scores (1 to 5) that increase with their relative positivity. Conversely, negative words have negative scores that decrease with their relative negativity (-1 to -5) . As most words are relatively neutral, they would not be found on these lists and therefore have a score of zero. If this technique were used alone, each word in a post would be assigned a score and the net score of the post would indicate its sentiment. A known drawback of this technique is its reliance on the comprehensiveness, robustness, and reliability of the list of scored words and their scores (as well as how current the list is). However, this may provide an opportunity for a domain expert to improve the list. They may want to ensure that certain words are included or given higher or lower scores to increase their influence on the net score. In the context of a candidate’s comments, this may include words that are more germane to the recruiting process. The stemmed or lemmatized forms of words “unresponsive”, “forgot”, “uncertain”, “delay”, “cancel”, “disrespectful”, etc. could be added or have their scores reduced to negative four or five.

A hybrid model that incorporates a Bayesian classifier that either utilizes customized word scoring to weight certain words’ probabilities and/or corroborate its findings is a plausible technique for the sentiment analysis in the context of a job candidate’s recruitment experience. Their posts on job search and recruiting websites are like movie, restaurant, product, etc. reviews, but should also include vernacular that is highly relevant to the domain. This could be a useful product for a member insurance company as the percentage of positive and negative sentiments could be factored into a benchmark score so that they could measure themselves relative to their peers and themselves should they determine that improvements in their recruitment process are necessary.

Emerging Topic Modeling

The products that my wife’s team focuses on are those related to talent. The previous section is an example of one related to recruitment where they can help their member companies understand the perception of themselves from the candidates’ view. They also develop tools to assist in the selection of candidates to recruit, as well as the retention of those employees. There is a lot of turnover for financial advisers and the consistent theme is that the candidate is trying a different or new career path and they are underprepared to be successful. This causes them to lose faith in themselves and/or their employer, or vice versa, and the relationship dissolves.

The key to fixing this is providing the financial advisers with the proper education and training. There are traditional and generic products available to train them to be more effective with certain products, demographics, or industries, or to educate them on ethics, compliance, and regulation. But within those realms, there are topics that are evolving or emerging. This could include a new technology that may have potential for the insurance industry, or a new regulation that is likely to be enacted. There are many topics from within the industry and many from the outside. The research team is tasked with determining all these potential and emerging topics. The product managers then must determine which topics to develop tools for as there are not enough resources to create them all. There is also the matter of timing; a new product takes from months to a year to develop and that requires the foresight to have it ready in time for the member companies.

The recurring pitfalls include not selecting the topics that the members come to need, developing products that are not essential (not relatively profound or impactful enough, but nice to have) or are too far ahead of their time. An example of the latter is the concept of team selling came to popularity in the insurance industry about six years after a product had been developed. Due to a lack of interest at the time, it was subsequently discontinued, only to be revived years later when it required an overhaul for image, content, etc. In retrospect, the product should not have been developed because the members did not have a need at the time.

To become better topic selectors, the product developers have traditionally relied on word counts from various sources that are essentially keyword searches. These include searches for products, tools, research and information on the company website and resource library. They also post articles or host webinars on emerging topics and count clicks and registrations. The results have been mixed and the product developers are seeking a tool to mine the text generated by their own sales team as well as the member companies they service. The sales team members and researchers (approximately two hundred people) generate copious amounts of activity notes on Salesforce, each documenting their conversations and interactions with their own contacts in the various regions where their member companies (hundreds) operate. Performing topic modeling on these notes could provide the required insight as new or emerging topics would stand out against the traditional, recurring and/or generic topics such as recruiting.

Topic modeling is an unsupervised machine learning process that finds a desired number of latent themes within a collection of documents, or activity notes in this case. It is a similar process to clustering in that it seeks to provide a structure (grouping) to data that is not readily obvious due to the volume and/or variety in which it is available and/or the velocity at which it arrives. While the latter assigns each data point to a single cluster, topic modeling allows not only each word or token to be associated with multiple topics, but each document may be associated with multiple topics as well.

These distinctions of “sharing” are critical as a single word or token can have multiple meanings or be associated with different areas of interest, just like many of these activity notes are likely to cover the many areas of interest discussed at a meeting. The sharing is an exercise in probability, specifically a Bayesian model that differs slightly from that of the classifier discussed in the previous section. While the classifier also uses probability to group the words or tokens, albeit just two, it has the benefit of previously labeled documents to make those predictions more confidently. Without labelled data, the Bayesian model used would likely be a generative one that involves Latent Dirichlet Allocation (LDA).

LDA tries to model the process that generates the data and pick the parameters of the model that maximize the probability of generating that data. Each topic has a different probability of generating different words, and each document has a different probability of generating different topics. So, the model iterates through each document and groups words that tend to be found together in district documents, or other documents that contain some of those same words. As these words are more likely to be found together and are found together in different documents also, they are more likely to be associated with the same topic. The model repeats the process of going through each document over and over to assess the strength of the topics (or word groups) it has previously created and adjust them based on what has been learned. Then each document can be assigned one or more probable topics based on the words it contains.

In the case of the Salesforce activity notes, a topic containing the following tokens would likely have been created: regulation, NAIC (National Association of Insurance Commissioners), DOL (Department of Labor), best interest, full disclosure, compensation, commission, incentive, junket, kick-back. Some notes may have just contained “NAIC”, “full disclosure” and “commission”, and others just “NAIC” and “junket”. Both would be referring to an upcoming regulation that requires financial advisors to disclose their commission on policy sales, and forbids the practices of gift-giving in the form of kick-backs, trips/vacations, etc. The topic modeler elevates each of these word’s importance regardless of their frequency relative to the collection of documents because they are related to a collection of associated words (topic) that is more pervasive than others.

The model requires the user to select the number of topics and therefore there is some trial and error in optimizing that number. That involves the manual process of reviewing the topics to ensure their associated words are pertinent and plausible. As this entire process would be run periodically, it also affords the user an opportunity to find topics not seen previously, or those that were only of interest for moment in time. It is likely that a proportion of the topics found will be the same each time the topic modeling project is done. These include the more prevalent areas discussed, such as the sales process, recruitment practices, etc. That said, any newly discovered topic is inherently important and gives the product managers an idea of a topic they may not have known about previously or they had previously believed to be insignificant.

This technique could also be expanded to include emails from higher level employees from the home offices of the member companies. This large volume of correspondence is another source for potentially emerging topics as it contains requests from CEOs, CMOs and other positions of leadership and management that are looking for insight and research on what they deem as important. Gaining insight from both the management level and the salespeople in the field provides two distinct perspectives of the industry that should result in identifying emerging areas of interest. Exploring the results should lead to the efficient and timely development of products and tools that can help the member companies prepare their employees to be more successful.