CS410 Technology Review Submission – Fall 2021

Survey of text mining and analysis toolkits

Review of available tools and how they can be used in real world large enterprise setup to improve customer conversions & satisfaction

Author : Sathish Rama (NetID : sbrama2)

Introduction

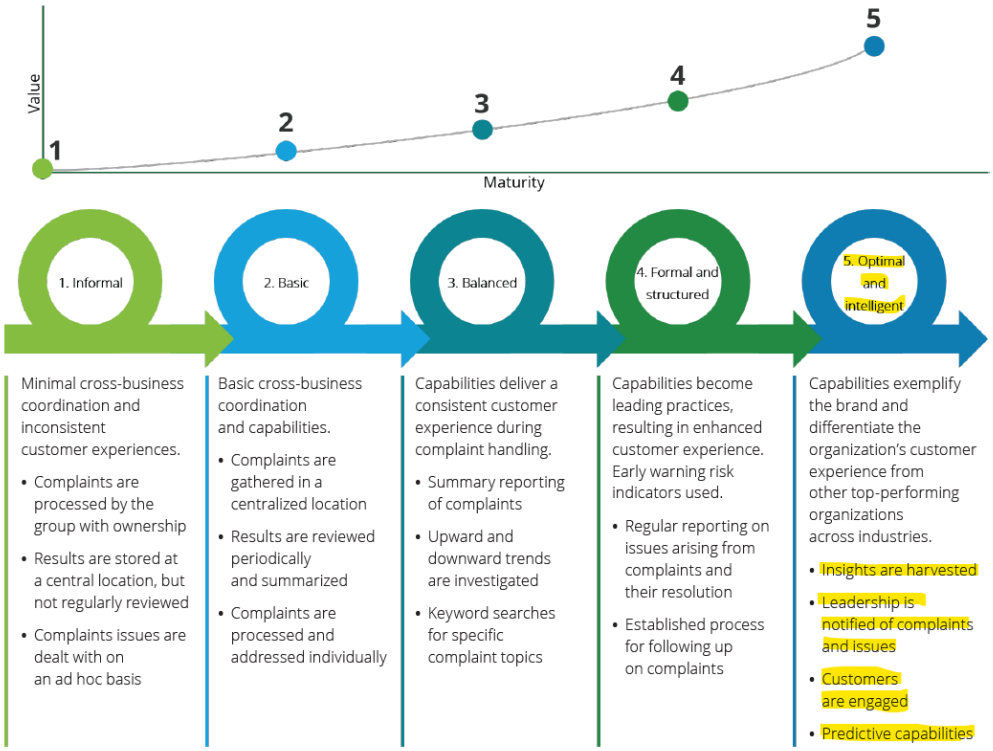
Text Mining and Text Analysis field has gained a great deal of attention in recent years in large and small enterprises due the tremendous amount of text data, which are created in a variety of forms such as social networks, customer conversations, communications & interactions, patient records, health care insurance data, news outlets, etc. Most of this text data is unstructured and posses significant challenges to process & interpret and make practical use of the analysis to apply to business context and automate business processes. Hence the technology toolkits have evolved tremendously and there are plenty of libraries, tools , platforms, software as service & application programing interface (API) offerings in the industry for variety of situations and use cases. This Technology review focuses on tools available & current challenges in real world large enterprise setup. Given the large variety of tools & their applicability, it is not possible to cover all tools and their detailed review so I will be focusing on well-known text analytics/NLP tools in customer conversions business process area.

Text Mining and Analysis is task of extracting meaningful information from unstructured text from various sources. There are several techniques and steps used in this process such as text retrieval, text representation, text encoding and text classification & text clustering to eventually derive meaning insights. This process usually starts with Text Retrieval, extracting structured text from unstructured text for natural language understanding using Named Entity Recognition (NER) & using Markov/probabilistic models for predicting neighboring text words and extracting relations for seeking semantic relations between entities(places, persons, products, companies etc) in the text. Then the text is represented & encoded which includes tasks such as Tokenization, Filtering, Lemmatization, Stemming, Part of Speech Tagging and eventually text is classified & clustered using Naïve Bayes classifier, Nearest Neighbor Classifier, Decision Tree Classifiers, Support Vector Machines, Clustering methods to help with sentiment analysis, topic selection or intent detection solutions that are used to predicting customer churns, classifying high priority customer complaints, routing prioritized complaints to call center associates, developing chatbots to interact with customers or help find or answer questions related to products or services etc.

Customer Conversions is one of the core business process for many organizations that focuses on improving the rate of converting business prospects/leads into customers thus generating sales/revenue for an organization. There are several strategies and approaches employed in this process with primary focus on providing smooth & great experience to prospects/leads by proactively identifying and clearing blockers, reducing friction in sales process & improving overall time to complete the purchase of a product or service. Text Mining & Analytics plays a significant role in this process and several organizations are leveraging Text Mining & Analytics to detect customer sentiment and reduce churn, proactively identify customer concerns, improve sales associates' productivity by appropriately routing customers to right sales agent

quickly, predict customer friction and attempt to remediate friction upfront, leveraging chatbots to provide customers information quickly, automate sales process through chat & social media channels etc.

Figure: Customer Complaints Maturity Model showing transition to Maturity Level 5 using Text Mining & Analytics



Tools Review

There are several tools or software's available for variety of text analysis tasks. These tools can be grouped in general into libraries, platforms (software as service offerings) & API's (Application programming Interfaces).

Libraries: Libraries are software packages that can be used to develop applications, models or platforms. The popular libraries are Apache OpenNLP, Natural Language toolkit (NLTK), MITIE(MIT Information Extraction), text2vec, Gensim, CoreNLP, Spacy, PyTorchTextBlob, Google Cloud NLP, Mallet, LingPipe, NLP4J etc. Based on the programming language such as python, java or C++ one can choose the appropriate toolkit to be used for text analysis. Libraries provides the flexibility on how to develop the text analysis solution but involves significant amount of effort to develop & integrate with existing enterprise systems. OpenNLP, NLTK & MITIE are popular for Java, Python & C++ respectively. They all offer good features such as Tokenization, parsing, Classification, Stemming, Part of Speech Tagging, Semantic reasoning etc.

Platforms: Due to large need across several organizations, several software vendors have developed offerings in this space to help organizations quickly develop & adopt text analytics

solutions into their business processes. Platforms provide advantage in reducing the amount of code to write using libraries and helps in focusing on solution and time to develop & deploy the solution. Some popular platforms are Amazon Comprehend, Google NLP, Microsoft Text Analytics, Nice Call Centers Solutions, IBM Watson, Monkey Learn, Aylien, Thematic, Lexalytics etc.

- IBM Watson popular as AI platform
- Google Cloud NLP is popular for Custom machine learning models
- Amazon Comprehend is popular for pre-trained text mining models
- Nice has several offerings centered around Call Center Management
- MonkeyLearn popular for User-friendly text mining
- Aylien popular for Simple API for text mining
- Thematic used for Text mining for customer feedback

Since these are platforms, we still need to invest effort in selecting the platform that suits well for the enterprise considering factors such as sourcing text into these platforms, type of problem to solve, the text retrieval/classification techniques/approaches needed and whether the pre-built models/features meet the requirements or custom models/features needs to be developed.

Text API's: Due to rapid change in social interactions, communication channels and common problems faced by most organizations - there is a rise of industry/solution specific offerings hosted as API's which enables organizations to rapidly integrate a text analytics solution without spending large effort in development of such text analytics solution. There are several API's available in this context. RapidAPI, a API platform, has several API's of which ADA NLP, Text Analysis API, Twinword Sentiment Analysis API appears to be popular for sentiment analysis, tokenization, text extraction, Named Entity Recognitions etc.

Conclusion

To summarize there are plenty of opportunities to leverage Text Mining & Analytics to improve an organization's customer conversion process & overall customer management. There are several tools, software's & API's to achieve this and depending on the enterprise technology landscape, budget & time constraints, one needs to choose the right tool & approach

References

1. Best Text Mining Tools of 2021, Rachel Wolff, March 2020, https://monkeylearn.com/blog/text-mining-tools/
2. "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut , July 2017, https://arxiv.org/abs/1707.02919
3. Top Free software for text analysis, text mining & text analytics , Web article , https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/
4. Enterprise NLP Challenges , by IBM Research , June 2020, https://www.ibm.com/blogs/research/2020/06/advancingnlp2020/