Natural Language Processing/ML techniques for Life Sciences companies

Ch. Yugandhar
Department of Business Analytics
Vignana Jyothi Institute of Management
Hyderabad, TS, India
yugandhar.c@vjim.edu.in

***Abstract*—This paper presents the success story, findings and learnings related to the implementation of a system that classifies conversations among medical professionals into various categories. A thorough scrutiny of source data, processes, and technological methods was conducted before the system was developed with the intent to implement in production with a well-established maintenance and support infrastructure.**

Keywords— Natural Language Processing, Logistic Regression, Random Forest, Production implementation, Machine Learning

I. INTRODUCTION

Off-label promotions are a big risk and major internal process nightmare for any Pharma/Life Sciences company. The term "off-label use" means the use of a drug for indications beyond those formerly evaluated by the manufacturer and approved by the FDA (1). Even with all the existing processes with lot of checks and balances in the systems, companies end up in situations that force them to face ill consequences caused by conversations or statements that slipped through the cracks. Out of total prescription-drug sales that account for about 216 Billion USD, about 40-50% of the share is through off-label promotions by various stakeholders (2). To mitigate these unforeseen issues US Department of Justice (DoJ) mandated the companies to comply with various procedures that promote "prevention and detection of criminal conduct" within the companies. One of the top five Life Sciences companies in the world implemented a quick solution that involves several analysts performing keyword searches through all the available content and weed out the claims that were not approved by FDA. This activity was successful to an extent for a time, but when the content grew over several years and regulatory bodies demanded more evidence to demonstrate the improvements in compliance process, an innovative and robust solution had to be thought through. This system implementation was a successful attempt to meet the above requirements.

II. FUNCTIONAL REQUIREMENTS

A. *Challenges*

Main challenges faced by the company can be broken down into the following few statements.

- Need assistance to enhance the review process of unstructured data in documents.
- Keyword based search by humans is becoming too tedious, time consuming, and inaccurate.
- Since the corpus has increased uncontrollably, the level of confidence in using the existing methods is fading.
- Incremental addition of workforce is leading to higher expenses as well as the computer systems are becoming chaotic which cannot be well managed.
- Difficult to prove the government auditors the effectiveness of the existing system in use.

B. *High level requirements*

The following are the list of high level requirements identified by the leadership in order to have long sustainable system that meets the current needs and can also scale to the future requirements.

- System should identify the claim-like sentences with a high level of confidence.
- Analysts should be able to tune the system to improve the accuracy.
- Continuous data input should be taken from new ongoing conversations and learn from them.
- Should be flexible to add or remove modules based on the need in demand.
- Should archive its tagging reasoning of data up to last six months for future review.
- Analysts should be able to demonstrate the effectiveness of the system to the auditors.
- The architecture of the system should support more than twenty medical drugs seamlessly.

III. SOLUTION ROADMAP

The roadmap that was suggested considered the goals of the management, customer base of the company, and the culture of the company. The following was a high-level roadmap and it was carried out successfully.

- Identify a solution which is innovative and that would meet the current needs.
- Create a proof of concept to demonstrate the idea in the short term.
- Once it is tested and proven expand the solution to lay a foundation for the future needs.
- Plan and perform phased deployments in production environment.

Taking all the requirements and constraints into account a proof of concept was carried out with a sample dataset. Since the task at hand is to analyse the text documents that were used for communication among various stakeholders a solution that deals with Natural Language Processing using Machine Learning was considered. Data was acquired from disparate source systems and fed into an engine whose high-level work flow is depicted as follows.

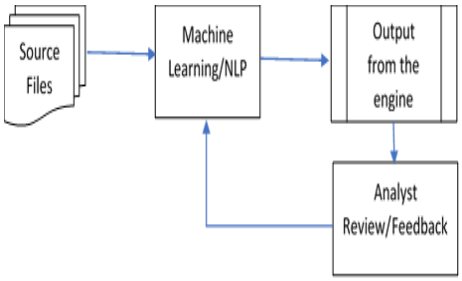

Figure:1

IV. Solution Implementation

A. *Proof of concept learnings*

For the sample set of data that was acquired, various technologies and methods were tried and tested. For text analysis data cleansing and data preparation steps were conducted before the data could be examined and fed into the system. The file types that contained the data included Microsoft outlook emails, MS Excel, MS Word, MS PowerPoint, Notepad files etc. High level tasks included the following.

- Data preparation
- Feature Engineering
- Model choice and training
- Model accuracy and performance tuning

a. Data Preparation
While preparing the data for pre-processing various data issues related to character encoding were encountered. To resolve these issues, the programming language libraries and software versions suitable for the accepted data encoding were used. In addition to this the source data curators were informed to provide all the data in the desired format only.

b. Feature Engineering
Understanding the data was the big part of the implementation as the language used by the Physicians and medical representatives and other stakeholders was related to the pharma field and there was certain lingo that was apt for a specific product and its competitors in the market place. Topic modelling, Phrase modelling techniques were used to understand the usage of data and their importance in the system. Count Vectorization, Parts of Speech tagging, Named Entity Recognition, Lemmatization among various other techniques were used to understand the data precisely, so it can be prepared and fed into the algorithm.

c. Model choice and training
As depicted in the Figure 1 the system receives feedback from the analysts based their knowledge from various meetings, conferences and other modes of communication with industry experts and their peer group. Such activity can create hidden feedback loops (3) if it is not monitored well. Much care was taken to avoid making the data too ML-friendly which otherwise creates “Pipeline Jungle” (4). This initiative proves to be wise in the long run for a production system and further various new medical products can be added to the same pipeline. After great deliberation and various model performance comparisons Logistic Regression and Random Forest were finalized as algorithms of choice for this implementation.

d. Model accuracy and performance tuning
Logistic Regression gave an 80% accuracy in predicting a sentence to be a Claim or No-Claim and the system took couple of hours to run the entire pipeline to process one medical product that has maximum amount of data which is about 50 M records.

B. *Method of automation and the selection of Technology stack*

After examining the requirements and constraints at hand, there has been a thorough investigation of the currently available methods and technology stack. The factors included in finalizing the processes and the technology stack are listed below.

- Culture of the company: There are mix of professional stakeholders that communicate with each other in various settings. So, influencing factors that lead to a specific style of communication are considered. The primary stakeholders are physicians, field representatives, and analysts from various departments like Compliance Monitoring Bureau.

- Type of business: Since the Life Sciences industry is enormous in size there is huge competition in the market place. There is a mix of Science, Technology, and R&D efforts along with marketing efforts for the discovery, development, distribution, and sale of products. So, data from all the above-mentioned stakeholders is collected for all the relevant products to make the discovery of the claim-like sentences highly effective.

- Method of automation: As per the current technology innovations and standards the company chose Machine Learning techniques for Natural Language Processing to carry out the development of the system. Various existing proven algorithms like Logistic Regression, Random Forest were considered for implementation against the newly available Neural Networks or Deep Learning for the two main reasons 1) There is still lot of research underway for the latest methods and too risky to consider at this time 2) The process that flags a document or sentence as a claim or no-claim has to be explainable to the auditors. Hence black box like approach cannot be taken or demonstrated.

- Technology selection: Proven, easily available, well maintainable software has been chosen to implement the system. This included the following factors.
    - Platform that can handle high volumes of data and that can support future data loads
    - Environment that has good governance strategies along with proven efficient architectures like expandability of the system as the data grows and pay per use like model for the compute power which requires payment only when used.
    - The interface that analysts use must be user friendly and expandable to add more features as well as products in the short term and long term.
    - Data storage, backup, and archival methods and data gathering systems in production should be integrated seamlessly.

Based on the above requirements the following suite of products was chosen.

- MAPR Spark ML platform for data processing
- AWS S3 buckets for file storage from various third-party vendors
- PySpark, Python for business logic
- Python, PHP, HTML, AWS RDS(with MySQL backend) to build the user-interface for analysts and APIs for cross-communication among entities
- Most of the libraries were from Scikit-SKLearn and NLTK
- SpaCy was chosen for syntactic and semantic processing
- Genism was used phrase and topic modelling

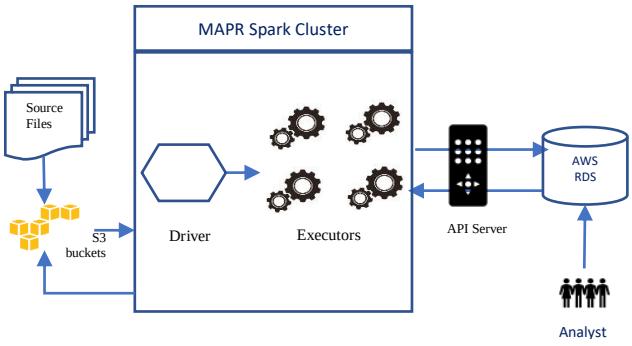

Figure:2

After applying the learnings from the proof of concept phase, standardization in the following areas took place

- A robust but flexible pipeline was created for taking the data through as shown in figure 2
- Interpretation of the results

C. Flexible and robust pipeline

The MAPR Spark cluster was divided into two parts 1) That stays on permanently in the host environment and 2) the other is stood up only when the training of the model is being done. This saves lot of cost while doing the development as well as while the system runs in production. Documents are fed into the system for a given product and are taken through the data preparation process of Count Vectorization, PoS tagging, NER, Lemmatization etc. After which the resulting corpus is fed into two algorithms i.e. Logistic Regression and Random Forest. Each sentence in a document is given score (0.01 to .99) based on the training data that is created by the subject matter expert. Each product is given a certain threshold for the scores based on the stakeholder's expertise. The sentences that are above or below this threshold will be considered for review by the compliance group based on their demand. Through the given user interface, the analysts from the compliance department review the score attributed to each sentence and either approve or disapprove the classification on a periodic basis. At the time of review the analyst can choose to add a given sentence to the training data and/or flag the same sentence to send it to the author to refrain from using certain verbiage.

The above process went through few cycles before the system was made ready for production use. The same process designed above was tried and tested for multiple products with few changes related to the specifics of the product and then the system was implemented.

D. Changes to model pipeline based on product and cyclical events

There were lot of individual observations for various products that are baked into their respective models.

- Phrase modelling: From a stream of sentences automatically detect those phrases that frequently co-occur. The text for a given medical product contains certain phrases that frequently co-occur in various contexts. To name a few, phrases like "upset stomach", "dry mouth", "severe diarrhoea", "blood clots", "internal bleeding", "blood pressure", "high blood pressure", "prescription drugs", "nose bleeds", "bleeding gums", "severe headache", "bloody stools", "pulmonary embolism" or PE, "back pain", "joint pain", "chest pain" are together generally. Also, there are many other bigrams, and trigrams, or n-grams used by physicians related to a specific medical product and side effects when the drug is administered to a patient.

$$\frac{count(XY) - Count(min)}{count(X) * count(Y)} * N < threshold$$

Count(X) => Number of times token X appears in the corpus

Count(Y) => Number of times token Y appears in the corpus

Count(XY) => Number of times tokens X and Y appear together in the corpus

$\boldsymbol{Count_{min}}$ => Accepted phrase occurs a minimum number of times(user-defined)

N => total size of corpus vocabulary

threshold => The strength of relationship between tokens to be accepted as a phrase together

- Topic modelling: A corpus can be divided into a group of topics to gain understanding of the corpus. Topic modelling is used for the purpose of dividing the corpus into a given number of topics and examine the data for further exploration. It basically reduces the dimensionality through unsupervised modelling like techniques by clustering the tokens into various groups with a certain weight attached to each token for a given topic. Document classification assigns a single category to a token where as topic modelling assigns multiple categories to a token where applicable. These techniques are used for Text classification, Recommender systems(for recommending various articles to users based on their current reading history), and uncover theme of a given blob of text or a given document. The most popular algorithms that are available are Latent Dirichlet Allocation(LDA), Latent Semantic Analysis or Latent Semantic Indexing.

  LDA model addresses the shortcomings of tf-idf scheme (5) and is represented in the figure 4.2 below.

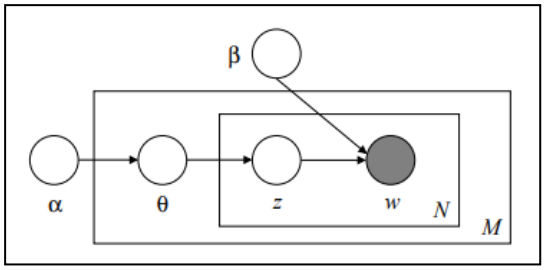

Figure 4.2

α and β are corpus level parameters; θ is sampled once per document an z and w are sampled once for each word in each document. LDA algorithm is proven to be highly successful in creating topics for any given corpus. For the current situation, using LDA, topics were observed to know if there are any cyclical events year over year such as medical conferences, press conferences etc. Analysts used their discretionary judgement to add any sentences to the training corpus if a new topic is observed due to the above. Also, if new product gets announced or released into the market either by the company or a competitor, the topics of discussions would change very much, and it is again at the discretion of the analysts to choose what should be added to the training corpus.

E. Results and the conclusion

The results were measured through confusion matrix and ROC AUC techniques. Accuracy levels varied from 65% to 80% for various medical drugs yielding an RoI of over two million per year. The models could be further tuned to improve the performance but the data that flows in from the real-world discussions is very dynamic in nature, so the company was comfortable to keep at this level of accuracy and have some human oversight to comply with the regulations of the Federal government.

F. Scope for future work

- *Current research vs. current implementation:* Natural Language Processing technologies have grown exponentially in the recent past starting from punch cards technology to batch processing and now can process multi-million web pages in less than one second(6). These tasks include, parsing, PoS tagging, machine transalation etc. to name a few. The system that was built made a conscious decision to use models that are trained on very high dimesional sparse features since the problem at hand is simple classification problem and also the model that is built to solve the problem should be transparent in nature and explainable to the auditors.
- *Extendibility of the system:* Tasks in Syntactic layer, as well as semantic layer are well handled by many tools available in the market and they are very matured. Now that the data of the company is well understood, and process has been streamlined along with the technicalities like character encoding, data pipeline, and feedback mechanisms, the processes that were followed could be replaced by various Neural Networks algorithms based on dense vector representations. A Word embeddings model like word2vec for Life Sciences industry would definitely come to good use to experiment with deep learning models. Mikolov and his colleagues proposed the idea of continuous bag- of-words(CBOW) and Skip-gram models to construct high-quality distributed vector representations(7). What can also be added to the current model is of two aspects that are very much applicable to the dataset namely, 1) Sarcasm detection 2) Metaphor understanding. Since the goal of the project is to find those sentences that might cause trouble to the company, inform all the stakeholders to avoid statements that are uncalled for, and at the same time avoid verbiage that is sarcastic in nature. Furthermore, in a given dialogue about 5% to 20% of the words are used metaphorically (8) demands that the Metaphor detection is highly necessary.

REFERENCES

[1] N.Repucci, "An Examination of Off-Label Marketing and Promotion: Settlements, Issues, and Trends", Journal of Health Care Compliance – September - October 2011

[2] David Armstrong and Anna Wilde Mathews,"Pfizer Case Signals Tougher Action on Off-Label Drug Use", Wall Street Journal , Eastern edition; New York, N.Y. [New York, N.Y] 14 May 2004: B.1.

[3] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, "Machine Learning: The High-Interest Credit Card of Technical Debt" *(page3)*.

[4] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, "Machine Learning: The High-Interest Credit Card of Technical Debt" *(page6)*.

[5] David M. Blei,Andrew Y. Ng, Michael I. Jordan,"Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) 993-1022.

[6] Tom Young, Devamanyu Hazarikaz, Soujanya Poria, Erik Cambria, "Recent Trends in Deep Learning Based Natural Language Processing", arXiv:1708.02709v5 [cs.CL],2018

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119.

[8] Ishaan Kumar, Rajat Kumar, "Natural Language Processing Metaphor detection", IIT Kanpur, October 4, 2015