Comparing SAS® Text Miner, Python, R

Analysis on Random Forest and SVM Models for Text Mining

Arun Jalanila, Nirmal Subramanian
Comprehensive Health Insights
Humana Inc.
Louisville, KY, USA
{ajalanila2, nsubramanian1}@humana.com

I. INTRODUCTION

One of the common discussions around an organization is the preference of one tool over another and the various factors such as current skill sets available in the organization, users learning ability and capacity of the tool to handle visual capabilities that leads to the selection and utilization of these tools.

In order to answer some of the questions around performance and ease of tool usage and visualization, a comparison between SAS® Text Miner, Python and R Programming tools was conducted. We incorporated the data that were provided as part of ICHI's data analytic challenge, which addresses categorizing user questions in a healthcare forum into predefined categories. The purpose of this study was to evaluate some of the tools available to perform these tasks based on our user experience.

SAS® Text Miner is a data mining tool used for finding patterns across text data through predictive modelling. Python and R programming tools (both open source tools) are used for statistical analysis and data interpretation.

II. METHODS

To compare these three tools, we used two models that are available across all three tools; the Random Forest (RF) Model and the Support Vector Machines (SVM) model.

As a preprocessing step, the data were cleaned using Java regular expressions and processed using an industry standard medical taxonomy system called Unified Medical Language System® (UMLS) MetaMap, which identifies biomedical concepts. These concepts are the keywords which are useful in identifying the predefined categories.

Keywords obtained from MetaMap were tokenized and function words (e.g., 'a', 'the') were removed using the stop word removal process. The remaining keywords were sent into a heuristic process that stems the word to eliminate different forms of recurrence (e.g., pluralization).

The term frequency-inverse document frequency (TF-IDF) algorithm was used to convert qualitative text content into quantitative representations based on weightage of each term relative to its frequency in a set of documents. This converts the text into vectors which are easily identifiable and used by SVM and RF models in classification. Figure 1 shows the process flow for implementation of the SVM and RF models in SAS® Text Miner., This process flow addresses data mining concepts and has inbuilt functions for tokenization, stemming and TF-IDF. Text topic node was implemented for topic recognition and discovery. The process flow in R and Python follow the same path as shown below.

Figure1: Process Flow in SAS® Text Miner

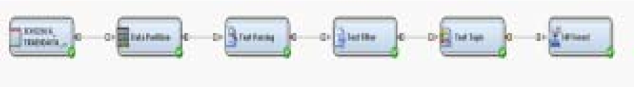
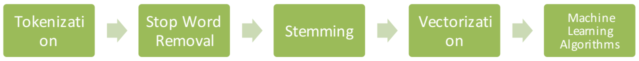


Figure 2: Process flow in Python and R



III. RESULTS

Table 1 illustrates the comparison of key factors; For the RF model, all three tools achieved the same level of performance. For the SVM model, SAS® Text Miner was less accurate than the Python or R implementation.

Table 1. Comparison of Key Factors

	Python	R	SAS® Text Miner
SVM	0.62	0.633	0.53
RF	0.6	0.619	0.64
User expertise	Experienced	Intermediate	Beginner
Ease of use	Fair	Fair	Good
Performance	Good	Fair	Good
Visualizations	Fair	Fair	Good

Our analysts possess the skillset of Experienced, Intermediate and Beginner levels in Python, R and SAS® Text Miner respectively. Table 1 indicates that while SAS® Text Miner has better ease of use with point and click functionality, Python was able to process the algorithm as fast as SAS® Text Miner. SAS® Text Miner has better visualizations which help the users to better interpret the results where as in Python and R, the user has to code manually.

IV. CONCLUSION

We believe that by combining R and SAS® Text Miner, it would be possible to achieve better results for the volume of data used in the analyses, especially using R to perform preprocessing and modelling and SAS® Text Miner to perform better visualization. However, based on the validation results, we believe that more work is needed increase the accuracy of the categorization results.

978-1-5090-6117-4/16 $31.00 © 2016 IEEE
DOI 10.1109/ICHI.2016.58