UNRAVELING BI-LINGUAL MULTI-FEATURE BASED TEXT CLASSIFICATION: A CASE STUDY

Aziz Yarahmadi, Mathijs Creemers, Hamzah Qabbaah, Koen Vanhoof

Abstract: *Extracting knowledge out of unstructured text has attracted many experts in both academia and business sectors like media, logistics, telecommunication and production. In this context, classification techniques are increasing the potential of Natural Language Processing in order to produce an efficient application of text classification in business context. This method could extract patterns from desirable text. The main objective of this paper is implementing a classification system which can be widely applied in commercial product classification problem solving. We have employed various applications of Natural Language Processing and Data Mining in order to solve parcel classification problem. Furthermore, we have investigated a popular case study which is associated with parcel shipping companies all around the world. The proposed methodology in this paper is part of a supervised machine learning project undertaken in order to gain domain specific knowledge from text.*

Keywords: *supervised text mining; commodity description classification; shipment classification system; natural language processing*

1. Introduction

Supervised machine learning frameworks have been introduced to business decision making in several domains in order to simplify tremendous complex datasets which represent few or no guidelines in data interpretation. On the other hand, more and more organizations are interested to apply such frameworks to documents in which lies patterns that can be modified towards growth and clarification. Text mining, which intensifies the hidden structure in documents, is recently well integrated in classification functions. Feldman and Sanger [Feldman, Sanger, 2006] express the similarities between Text Mining and Data mining due to the fact that text mining is originated from data mining and inherits much of data mining techniques in the text processing context.

Recent research has focused on applications related to text pattern recognition in business planning and operations. Current examples are deepened in sentiment analysis applications in marketing, social media mining for product and service penetration, fraud detection in financial statements, customer service and many more. Unique potentials of pattern discovery in documents, from structured to totally unstructured, empowers business administrators to extract rules even from informal and manipulated sentences. However, applications in this context are sensitive to the text structure and vary on the use and the expected results. An attractive case study for companies producing and transferring commercial goods is to employ a classification system for the items they produce, send or receive. Such a system could be highly valuable to inbound and outbound logistic services as well [Shankar, Lin, 2016]. The combination of text mining and supervised machine learning is known as corpus-based document classification. Furthermore, in this context, documents are being annotated manually in order to create corpora, that acts as a relational database of documents, categorized and supervised to be used later on, to assess and analyze consecutive document datasets.

Commercial commodity categorization using text mining implements a classification framework using pre-labeled items to train a classifier [Kotsiantis, 2007]. Moreover, the ontology behind supervised text classification highlights the need for enough data to train the classifier. In other words, problem solving cases dealing with commodity classification focuses on using commodity features to group similar items. Consequently, the research topic in this paper focuses on forming a corpus of documents with pre-determined structure required to form the classification model. Additionally, we present a methodology to address the question of classifying commodity descriptive features by dealing with parcel delivery records with commodity description in two languages: English and Arabic. We are proposing a three step framework to analyze and categorize products in 70 commodity description groups provided by the parcel delivery company:

1. Data cleansing, pre-processing and translation
2. Corpus creation and class labeling
3. Model selection and results verification

The main objectives of our study is to classify items in big datasets of delivered commodity descriptions including multiple languages and to normalize categories into a set of pre-defined items in company’s interest. Here, the second main approach in text mining which uses pre-defined dictionary of terms would also seem reasonable to perform. We will also illustrate the implementation of “Wordnet”, as the widely used lexicon-or dictionary, to categorize items and the shortcomings of such method in this case study. The rest of this paper is organized as follows. Section 2 will go through recent study and evidence applying text mining rules in text classification. Section 3 focuses on the proposed framework based on the structure of data in this study. In section 4, model selection is discussed in details. The results of the proposed framework have been displayed in section 5. Finally, section 6 outlines the conclusions.

2. Recent work

The state of the art in text mining mainly focuses on sentiment analysis. The goal in sentiment analysis is to extract the modality of user behavior expressed in comments towards a subject. Applications tend to classify the reviews in bipolar classes of negative or positive. Moreover, commodity classification has been increasingly explored from the quantity point of view which reports on the number of entities instead of classes themselves. Commodity description features refine the classification task in this study though. Underlying qualitative features like the use case, nature and the customer segment, categorize the commodity description.

Ghani et al [Ghani et al, 2006], have represented the idea of product identification based on series of attribute-value pairs. Thus, such a methodology allows feature vector extraction that supports the term frequency framework in text classification.

Popescu and Etzioni [Popescu, Etzioni, 2005] have built the model called OPINE based on the review scores and features extracted from customer reviews. A classification framework based on these features can be constructed. Furthermore, feature extraction in text classification is of great importance. Although different definitions of a feature might be found due to context, a feature in text mining is theoretically the presence or absence of each word in text. Distinctive features are to be scored higher in tf-idf method, a significant feature for grouping similar documents.

The concepts for tf-idf are separately discussed in [Luhn, 1957] and [Spärck Jones, 1957]. Recently the application of "Wordnet" in lexicon-based text classification has created opportunities in semantic mappings in product classification [Beneventano et al, 2004]. Although lexicon-based text mining approaches employ dictionaries of word trees, with broad synsets, which are synonyms for English words, a supervised classification approach determines manually categorized descriptions for item classes, which would result in higher precision and recall degrees for the classification model. On the other hand, the use of dictionaries on item class description requires clear description words, which in this case is limited and no specific study has undertaken such an approach so far.

3. Proposed methodology

Representation of the case study in this research focuses on the application of current text classification techniques on a common data classification problem in business context. Due to lack of matching research applications, we have been motivated to answer commercial production request for more information regarding classification of products based on their technical features, use cases, age and gender specifications. Customer segmentation and profiling are two major fields of study by categorizing parcel packages. Parcel delivery companies are not normally able to open packages to get insights into the contents. Commodity features are provided by the merchants that act as package senders. These types of information are represented using various languages. On the contrary to opinion mining problems dealing with the polarity of features dependent to an entity, the use of dictionaries is limited to text classification with few text features. In this perspective, we propose a supervised classification methodology for commodity classification based on commodity descriptions. In order to unravel the commodity classification case study, a data structure deliberation is presented and the need for our research methodology is introduced.

3.1 Data

Commodity classification task we have at hand deals with a dataset of hundred thousand items of commodities delivered to customers by a parcel delivery company. The parcel delivery company has invested in two major delivery systems:

The first one is an international e-commerce based shopping ship service that aims to connect major merchandising and shopping platforms like eBay and Amazon with potential customers in countries with less product delivery coverage. 250,000 customers have recently used this initiative.

The second service offered by the company is their domestic parcel delivery system that is generating loads of data regarding package features. Statistics showed that 3 million users have used this service. The parcel delivery company has focused on three main merchants with the highest market shares.

The original dataset forms the parcel delivery record including 56 parcel features. The feature types vary from size dimensions and weight to payment type and destination. The commodity description feature is very broad and the need for a general categorization into 70 pre-defined classes is inevitable by the parcel delivery company. Assigning a class to each commodity becomes challenging when the only feature to be used for classification is the commodity description as it provides information about the name, quantity and brand of the product delivered to the customer. However, classification based on size dimensions is not in the investigation of this study, considering that there is no record of classification based on the dimensions of the commodity. Still, description features provide a range of descriptive properties for each item that introduce the brand, technical specifications, model, etc. As commodity description shares various property names with the original classes, it is reasonable to take it into consideration as the feature extraction source.

The challenge arises when the description feature contains non-ascii characters (Table 1). Text preprocessing is to be done in order to normalize the text.

Commodity Description
#1/DKNY WOMEN WATCH MODEL NO NY8833
#1/براون سيلك ابيل لازاله الشعر و تحديد الحواجب BRAUN
#1/Aramis Brown for Men 110ml Eau de Toilette
#1/RGB Led Strip Waterproof 5M SMD 5050 300
#1/كيراتين مرطب لا يشطف 177مل
#1/Braun SE 5780 Silk Epil 5 Epilator

Table 1: a sample of data before preprocessing

3.2 Preprocessing advantages

The main obstacle in normalizing text in this dataset is the fact that in some cases, the commodity description is added in Arabic or it is a combination of Arabic and English. Such text inputs would drastically lower the classifier accuracy as wrong or no specific classes would be assigned to commodities with such descriptions. While dealing with big datasets, one approach could be to remove commodity records with descriptions in Arabic. Our text classification approach however, proposes to translate the descriptions into English. Translation function accuracy is assessed by native Arabic speakers in order to remove the outliers and less precise translations.

Furthermore, in some cases, description feature states names which represent distinctive classes. Except for size and weight, that might be able to provide more distinctive feature for each commodity description in these cases, it is not possible to figure out the actual class. As the classification is built on description only, these records are to be removed from the training and test set. Removing punctuation marks and English stop words are respectively the next steps in this dataset, as the focus in item classification is on proper nouns. Numbers and hash tags are as well removed as they will not create added value. Missing values from commodity description column are few but possible. Removing such records is necessary along with duplicate items. Further, we implemented stemming in order to normalize further proper commodity names. The role of stemmers in text preprocessing is to reduce the number of derived forms of words. Removing prefixes is the other advantage in this case. For large corpuses, stemmers have led to better results [Vijayarani et al, 2001]. In our case, all forms of "package", "packages" and "packaging" are relatives of the basic form "package". Similarly, stemming significantly impacts the multiple forms of verbs usage. Figure1 is visualizing the schematic of preprocessing steps which were applied in this study. In the next part, we will go through model selection.

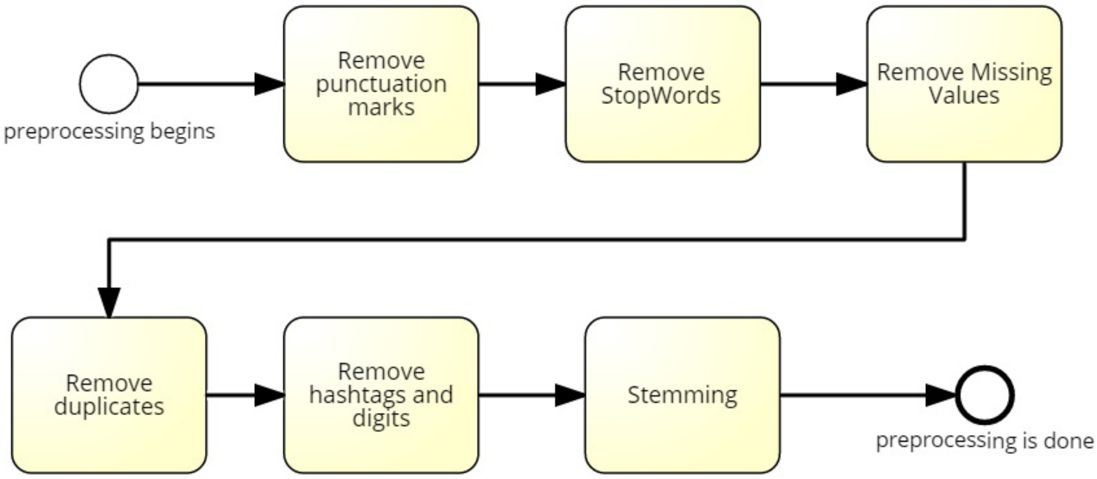

Figure 1: preprocessing steps implemented using Signavio Process Editor

4. Model selection based on text structure

Natural language processing tasks vary from part of text tokenization to sentiment detection and therefore, multiple probability distribution estimation techniques can be applied to text classification projects. Our desired task is to assign a class to a set of words using pre-labeled training data. 70 pre-defined classes imply the use of supervised text classification. The algorithm used to assign respective classes requires a set of features for each item as training data. A feature is primarily a conversion function to transform each input value to a set of input values and relative label given to it. The machine learning algorithm then models these feature sets and extracts a classification schematic that can be used for incoming unlabeled data. The transformation function will be implemented on new data and feature extraction is performed again and the predicted model will output results, depending on the classification algorithm used [Aggarwal et al, 2012].

Since classifying our feature description can be viewed from the unsupervised text classification point of view as well, the most noted lexicon, or dictionary to be used in such a methodology is “Wordnet”. Recent use of Wordnet has spread in academic text classification use cases due to comprehensive network of word trees, creating a . Initial use of Wordnet was a step toward building language model using treebanks of synonyms and hyponyms of words. The NLTK package for Python represents a sound implementation of Wordnet and has gathered tools to perform various NLP information extraction techniques. Despite huge capabilities of using it, the obstacle on our way is linked with the pre-defined 70 restricted classes. Our implementation of Wordnet was based on evaluating the commodity description word by word. For each word, we found all synonyms. Lookup function would then match if any of the words would match any synonyms found in Wordnet. If not, we moved one level deeper in the word tree and tried again. Such a method would try the same procedure on bi-grams on the next phase and would try the N-gram model for N=3 and N=4 as well. While using Wordnet, which is basically a dictionary of each word and its categorical mappings, is a type of unsupervised machine learning, it has proved to be accurate on document classification. Restricted classes we have in our study did not map correctly and precisely onto equivalent Wordnet categories, and that is why using supervised corpus based text classification is preferred in this study

Choosing the right model has been studied in this research among a set of classification algorithms. A comparative statistical analysis will show the state-of-the-art modeling capabilities of each classifier. Two popular classification algorithms in this context are Naïve Bayes and K-Nearest Neighbor, due to their predictive capabilities and precise performance. Naïve Bayes will be error-prone for small datasets while it is easy to implement in text categorization problems. However, with regard to classification accuracy it will not be preferred over Support Vector Machines [Rennie et al, 2003] [Zhang, Oles, 2011]. While Support Vector Machines (SVMs) has been broadly accepted for text classification tasks [Joachims, 1998] [Pang et al, 2002], we have developed models based on underrated classification algorithms. These include Boosting, GLMNET, MAXENT and SLDA.

[Kudo, Matsumoto, 2004] has introduced a framework based on Boosting algorithms to classify semi-structured sentences represented as a labeled ordered tree. The paper presents the subtrees as Features extracted from text. Boosting algorithms leverage the weak learners, classifiers which predict the right class only slightly better than random guessing, to become strong learners, which would correlate rigorously with the right classification. Each classifier will be trained based on the hardest instances to classify by previous classifier [Schapire, Singer, 2000]. Implementations for GLMNET, MAXENT and supervised LDA are available using R packages. MAXENT algorithm in our implementations has achieved the highest precision and recall measures. Maximum Entropy classifier (abbreviated in MAXENT), has proved the efficiency of the classifier in text categorization scope.

Della Pietra et al. [Della Pietra et al, 1997] state that given a set of classes C with

$$C : \{c_1, c_2, c_3, \dots c_N\}$$

A set of labeled documents with classes:

$$(d_1, c_1), (d_2, c_2), \dots (d_n, c_N)$$

For the document d and the pertinent class c, MAXENT representation of $P(c|d)$ is an exponential form of:

$$P(c|d) := \frac{1}{Z(d)}\, exp\, (\sum_{i} \lambda_{i,c}\, F_{i,c}\, (d,c))$$

here, $\lambda_{i,c}$ is the set of features and their weights. $F_{i,c}$ is the feature in this model and $Z(d)$ is the normalization function:

$$Z(d) = \sum_{c} exp\, (\sum_{i} \lambda_{i,c}\, F_{i,c}\, (d,c))$$

$F_{i,c}$ is a binary function and depending the relativity of the feature selection to context, the output for each instance to be classified is either zero or one [Chieu, Ng, 2002]. Theoretically speaking, features in Maximum Entropy model are a function of classes predicted and a binary context constraints. [Ratnaparkhi, 1996] has introduced the representation of such a function in his part-of-speech tagger. A feature in this explanation is a set of yes/no questions with a constraint applied to each word. Any word that satisfies the constraint would be a feature. Maximum entropy models state the fact that among all probability distributions available to model testable data, which in this study is the groupings of the commodities, the one model with the highest entropy is the true model. The main substantial feature in maximum entropy in this study is the ability to handle comprehensive features which will be the case here. Document term matrix or feature matrix is created using the package tm of R, and it is the nominal representation of the most distinctive terms used in each class for categorization. Additionally, a feature-cutoff will leave feature seen than a specific number of times. Figure 2 represents an overview of model generation procedure in this study.

We have used the RTextTools library for machine learning in text classification tasks. The R distribution of MAXENT, trains a maximum entropy model using a document term matrix and feature vector. Document term matrices are numerical representations of documents. Each document would be represented word by word in rows and the columns state the existence or absence of words in each document. A feature vector would as well represent the given labels. Consequently, the trained model will be tested with new unlabeled data. To train models, there are labelled datasets from headlines of New York Times and bills from United States Congress.

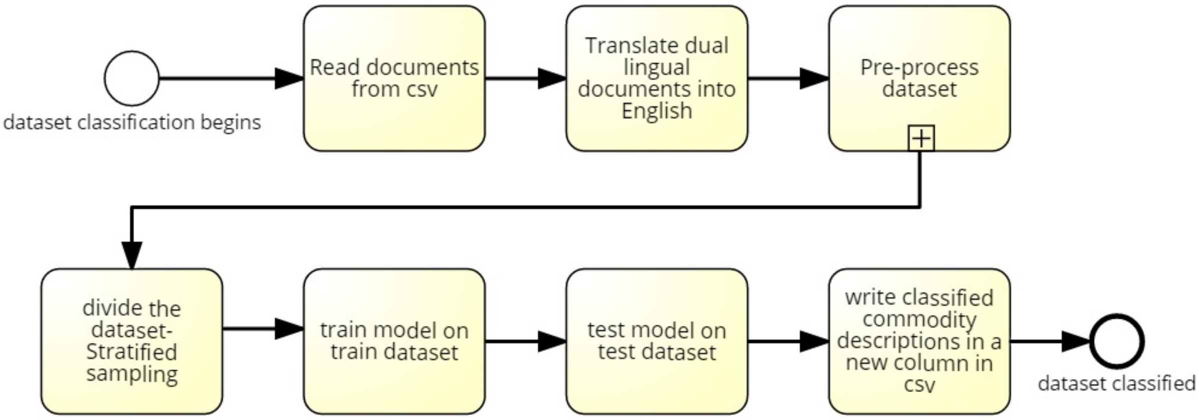
dataset classification begins
Read documents from csv
Translate dual lingual documents into English
Pre-process dataset
divide the dataset- Stratified sampling
train model on train dataset
test model on test dataset
write classified commodity descriptions in a new column in csv
dataset classified

Figure 2: Model generation steps implemented using Signavio Process Editor

5. Discussion and Results

Using stratified sampling and with respect to 70 classes at hand, we have assigned %80 of data to training data and the model trained on this portion will be applied to the rest %20. Data loaded will be used to create a document-term matrix. Preprocessing options are available in this step. In the next step a container will be created from the document-term matrix that acts as a list of objects for the machine learning algorithm. Training models will be initiated next and classified data based on the trained models will be provided consequently. Finally, the analytics for classification task will be provided and the results will be exported to the output file desired by the user [Jurka et al, 2013].

Applying our classification methodology to new unlabeled data consistently provided by our parcel delivery partner proved to be highly efficient both theoretically and empirically. The classification model achieved the precision and recall measures of 0.9365217 and 0.9144928, respectively. Precision measure introduces the number of right returned instances that are queried by the classifier.

$$Precision = \frac{true\ positive(t_p)}{true\ positive(t_p) + false\ positive(f_p)}$$

In this equation, the true positives and false positives are respectively the number of hits or true instances to be found and the number of instances that are selected as hit but belong to other groups. In other words, precision detects the percentage of relevant items selected. Recall is the other performance measure for classification tasks and it detects the percentage of selected items that are relevant.

$$Recall = \frac{true\ prositive(t_p)}{true\ prositive(t_p) + false\ negative(f_n)}$$

False negative in this equation indicate the case where an instance is rejected to be in the respective class, while it actually belongs to the class. Precision and recall are both indications of the relevancy of model. Table 2 shows precision and recall rates for the five classification algorithms applied on our dataset. As the table represents, Support Vector Machine algorithm shows lower precision and recall rates, compared to Max Entropy and the reason lies in the fact that text data is not a complete relevant input for SVM. Vector representation of text results in sparse matrices and these matrices will not lead to the highest ranked results in SVM.

Algorithm	**Precision**	**Recall**	**F-score**
GLMNET	0.5207246	0.4208696	0.4407246
MAXENT	0.9365217	0.9144928	0.9228986
BOOSTING	0.9502899	0.9165217	0.9260870
SLDA	0.8911594	0.8952174	0.8839130
SVM	0.9026087	0.8834783	0.8882609

Table 2: precision, recall and F.score for each model

Some of the models that have been tested in this study have showed close precision and recall measures. With respect to precision, Boosting, MAXENT, SVM and SLDA have gained the highest ranks respectively, outperforming GLMNET significantly with the precision roughly around %52. Recall measure for GLMNET is low as well compared to the rest of the models. Big number of classes in our case study lowers the results for GLMNENT classifier. Furthermore, GLMNET is a low-memory classifier and the size of case study dataset is much bigger than the 30,000 text documents it can handle. MAXENT and BOOSTING gain very close recall measures, followed by SLDA which would outperform SVM by around %1. Figure 3 and figure 4 represent the visualization for precision and recall achievements for all models tested in this study.

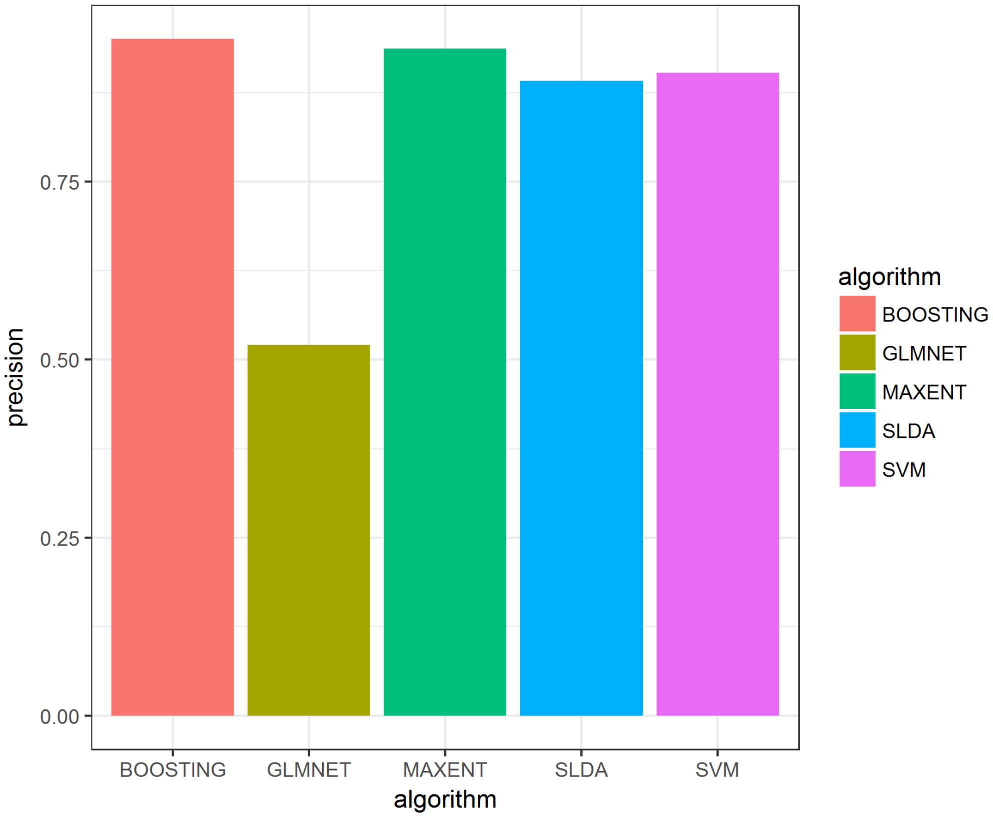

Figure 3: precision comparison charts for all models

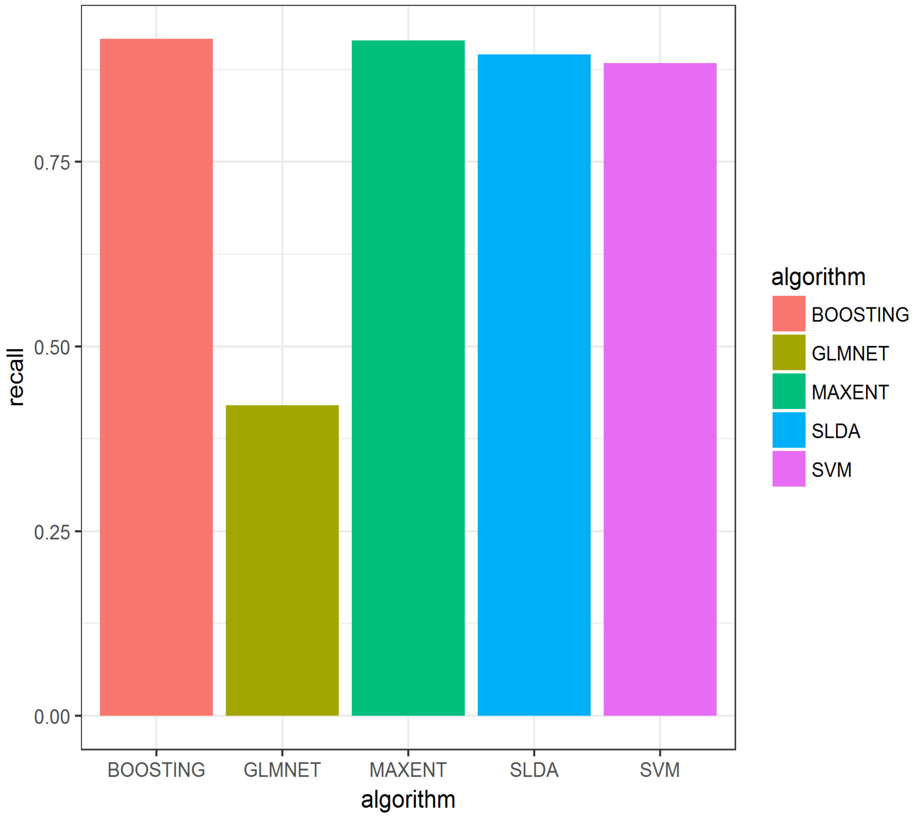

Figure 4: recall comparison charts for all models

The F-score for each classification algorithm is illustrated in table 2. The F-score is simply the harmonic mean of precision and recall. For the case where precision and recall are both one, we multiple the measure by 2.

$$F = \frac{precision * recall}{precision + recall}$$

The highest F-score is achieved by BOOSTING and MAXENT respectively, followed by SV and as expected, the F-score for GLMNET is considerably lower, around %44. Figure 5 represents the comparison of F-score achieved by all models in this study.

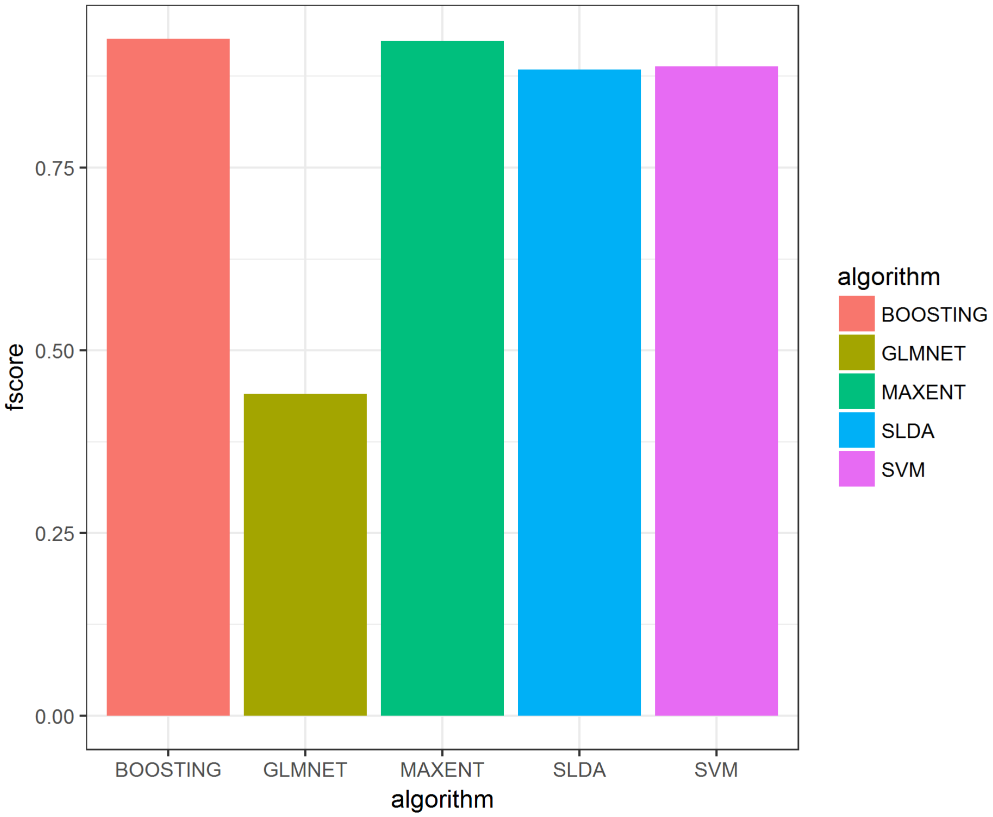

Figure 5: F-score comparison results for all models

Our implementation in R would read the unlabeled data from a csv file and the output would be written to a new column on the csv file, making it easy to transform complicated descriptions into a definite set of 70 classes. Low confidence labels are put in separate category which is called “Others”. These instances are not to receive a class tag or have loose bonds with any of the classes. In other words, the feature set extracted from these description items is by no means clearly showing pointing to any of the 70 classes defined.

The implementation of MAXENT classifier in R has proved to be memory friendly. Models for MAXENT and GLMNET can run on local hosts while the rest of the models can run out of memory even on our access point to Flemish Supercomputer that is a powerful computing node, which in fact originates from the vast span features and high dimensional vector matrices.

6. Conclusion and future work

Supervised machine learning has proved to be significantly comprehensive in natural language processing tasks. In one particular case, we have proposed a text classification methodology, specifically used in commodity features descriptive datasets. Text including dual language characters makes it challenging to get insights into text classes. Our proposed framework employs a single language translation function to convert every item in a combination of Arabic and English strings in our dataset into English. Several preprocessing and data cleansing techniques are implemented to prepare the data to be fed into a model based on a couple of supervised machine learning algorithms. The case study introduced in this research is a typical data analytics application that many companies in production and service industry face daily. As long as descriptive data stream is loaded in databases, representing wide features of products, we are able to handle dual-lingual text, remove noise and classify each item respectively. We have addressed the problem for the parcel delivery company to categorize items and answer socio-demographics qualitative questions concerning customers and how their parcel delivery records can be monetized as valuable data source for merchandising rivals. We have gained notable precision and recall measures, proving our methodology to be responsive and accurate.

From a practical point of view, the business partner is able to monetize such a categorization function with respect to the highly desirable items specifically in online shopping. In case the vendor, the value and the frequency of sale for each commodity is available, online markets will gain more insights into sale and marketing. Other example applications would be in geo tagging parcel delivery service selection with respect to category and value of parcel.

While our solution to the commodity classification problem is based on supervised machine learning, further insight into customizations based on "Wordnet" to make it proper for similar cases and implementing the N-gram model can be further studied and the results could be interesting.

Bibliography

[Feldman, Sanger, 2006] R. Feldman. and J. Sanger. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006.

[Shankar, Lin, 2016] S. Shankar and I. Lin. Applying Machine Learning to Product Categorization. Stanford CS229, 2016.

[Kotsiantis, 2007] S.B. Kotsiantis. Supervised Machine Learning: A Review of Classification Technique. Informatica 31, 2007, 249-268 p.

[Ghani et al, 2006] Ghani et al. Text Mining for Product Attribute Extraction, SIGKDD Explorations, Volume 8, Issue 1, 41-48 p

[Popescu, Etzioni, 2005] A.M. Popescu, and O. Etzioni. Extracting Product Features and Opinions from Review. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (2005), 339-346 p.

[Luhn, 1957] H.P. Luhn,. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of research and development, IBM. 1 (4) ,1957, 315p.

[Spärck Jones, 1957] K. Spärck Jones. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation 28, 1957, 11-21 p.

[Beneventano et al, 2004] Beneventano et al. A web service based framework for the semantic mapping amongst product classification schemas. Journal of Electronic Commerce Research, VOL. 5, NO.2, 2004

[Vijayarani et al, 2001] Vijayarani et al. Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks, Vol 5(1), 2001, 7-16 p.

[Aggarwal et al, 2012] Aggarwal et al. A survey of Text Classification Algorithms. Mining Text Data 2012, 163-222 p.

[Rennie et al, 2003] Rennie et al. Tackling poor assumptions of naïve Bayes classifiers. Proceedings of 20th international conference on machine learning, 2003, 616-623 p.

[Zhang, Oles, 2011] T. Zhang. and F.J. Oles Text categorization based on regularized linear classification methods. Information retrieval 4, 2011, 5-31 p.

[Joachims, 1998] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. Journal of Machine Learning: ECML-98,1998), 137-142

[Pang et al, 2002] Pang et al. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing , Volume 10, 2002, 79-86 p.

[Kudo, Matsumoto, 2004] T. Kudo and Y. Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. Conference on Empirical Methods in Natural Language Processing 2004.

[Schapire, Singer, 2000] R. Schapire and Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. Machine Learning 39(2/3), 2000, 135-168 p.

[Della Pietra et al, 1997] Della Pietra et al. Inducing Features of Random Fields. IEEE Transactions Pattern Analysis and Machine Intelligence, Vol. 19, NO. 4, 1997.

[Chieu, Ng, 2002] H.L. Chieu and H.T. Ng: Named entity recognition: a maximum entropy approach using global information. Proceedings of the 19th international conference on Computational linguistics, Volume 1, 2002, 1-7 p.

[Ratnaparkhi, 1996] A. Ratnaparkhi A Maximum Entropy Model for Part-Of-Speech Tagging. Conference on Empirical Methods in Natural Language Processing 1996.

[Jurka et al, 2013] RTextTools: A Supervised Learning Package for Text Classification, The R Journal Vol. 5/1 June 2013.

Authors' Information

Aziz Yarahmadi *– PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University,B2a,Campus Diepenbeek, B-3590 Diepenbeek, Limburg, Belgium*
e-mail: aziz.yarahmadi@uhasselt.be

Major Fields of Scientific Research: Social Networks Mining, Text Mining Applications in Business, Business Process Management and Improvement

Mathijs Creemers *– PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B56, Campus Diepenbeek, B-3590 Diepenbeek, Limburg, Belgium*
e-mail: mathijs.creemers@uhasselt.be

Major Fields of Scientific Research: Process Mining, Knowledge Management, Resource Metrics

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B2a, Campus Diepenbeek, B-3590 Diepenbeek,Limburg,Belgium*
e-mail: hamzah.qabbaah@uhasselt.be

Major Fields of Scientific Research: Data mining, E-business, Social Network Analysis, Knowledge Management

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek,Limburg,Belgium*
e-mail: koen.vanhoof@uhasselt.be
Major Fields of Scientific Research: data mining, knowledge retrieval