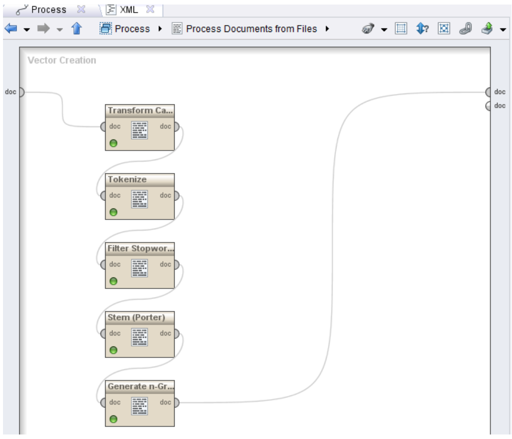

Fig. 7. Tokenization Processes

The upper level parameters for the “Process Documents From File” operator have been set to generate a document victor matrix that identify the important tokens in these clinical documents by calculating their TF-IDF and to optimize the victor document using an absolute pruning method (see Figure 2). The lower level process contains operators like the “Tokenize” where it have variety of options including to use no letters separators, regular expressions, linguistic sentences, linguistic tokens or the use of specific characters as tokens separators. We initially started by using no letter option for the tokenizer. The other operators used in the tokenization are to eliminate the Standard English stop words and to convert the tokens to their stem using the porter dictionary.

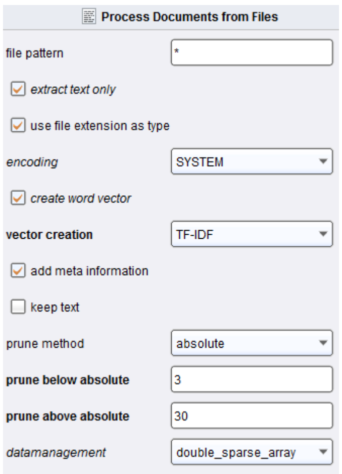

Fig. 2. Parameters used in Converting the Clinical Narratives in Document Matrix .

Once we completed the tokenization process then we can start to experiment with using different classifiers for categorizing clinical documents into the eight selected categories. For this purpose we need to have processes for training the classifiers and for testing the trained classifier on new provided documents that we need to know its category. Figure 3 illustrates the upper level for the operators required to be dragged onto the design stage. In the training part, we need to store the vectored documents and to use a classification operator container (i.e. the Validation Operator) and to store the resulted model of training into a store. However, the categorization process starts by reading both the vectored documents and the stored model of classification and apply this model on the newly provided clinical document (this requires the use of a “Process Document from File” operator) to predict its category.

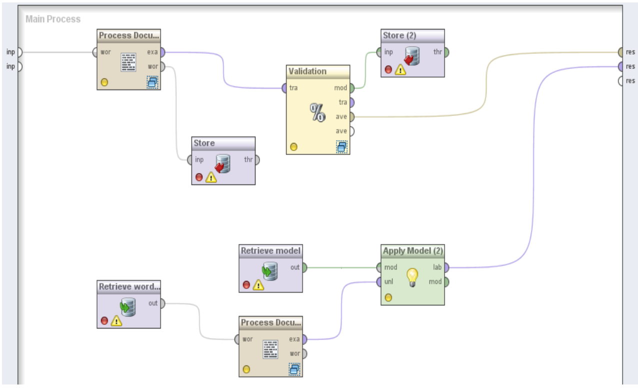

Fig. 3. Training Classifiers and Categorizing New Documents.

The validation operator generates new stage with two windows, one for the including the classifier and the other for applying the classifier on the new provided documents and measure the performance statistics. RapidMiner provides huge number of classifiers (e.g. Naive Bayes, SVM, J48, JRip, ZeroR, Random Tree, K-NN) as well as wide range of performance measures (e.g. Accuracy, Kappa Statistics, Recall, Precision, Absolute Error, Cross Entropy and Correlation). Figure 4 illustrates the classification and categorization container where the K-NN has been used as the classifier.

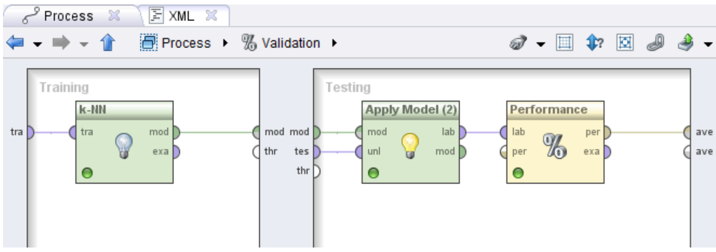

Fig. 4. The Classification and Categorization Process of Clinical Narratives.

Based on this container for classification and categorization template, we selected seven notable classifiers and run them against a sample of 40 clinical narratives for each the eight different clinical categories where the test data represent 24 clinical narratives that we need to predict its category (e.g. Autopsy, Diet, Discharge

Summaries, Chiropractic, Cosmetic, Dental, ENT and Radiology). Figure 4 illustrates the overall accuracy of the seven selected classifiers.

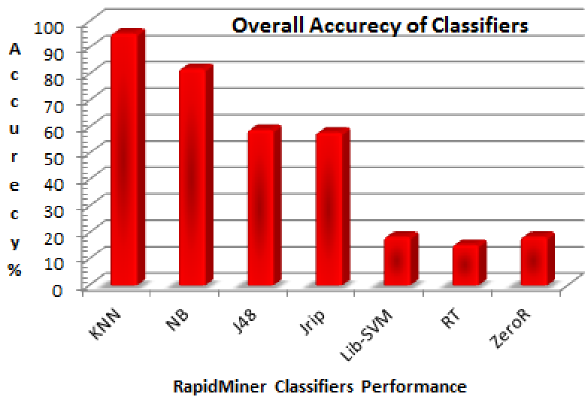

Fig. 4. The Accuracy of the Seven Selected Classifiers.

K-NN proves to provide the highest accuracy (95.5%) compared to the other classifiers. Figure 5 illustrates the outcome of predicting the category of the 24 test sample.

Row N...	label	metadata_file	prediction(l...
1	Test	Autopsy_R6.txt	Autopsy
2	Test	Autopsy_R7.txt	Autopsy
3	Test	Autopsy_R8.txt	Autopsy
4	Test	ChriopracticR6.txt	Chiro
5	Test	ChriopracticR7.txt	Chiro
6	Test	ChriopracticR8.txt	Chiro
7	Test	CosmeticR6.txt	Cosmetic
8	Test	CosmeticR7.txt	Cosmetic
9	Test	CosmeticR8.txt	Autopsy
10	Test	DentalR6.txt	Dental
11	Test	DentalR7.txt	Dental
12	Test	DentalR8.txt	Dental
13	Test	Diet_R6.txt	Diet
14	Test	Diet_R7.txt	Diet
15	Test	Diet_R8.txt	Diet
16	Test	DSR6	DS
17	Test	DSR7	DS
18	Test	DSR8	DS
19	Test	ENTR6.txt	ENT
20	Test	ENTR7.txt	ENT
21	Test	ENTR8.txt	ENT
22	Test	RadiologyR6.txt	Radio
23	Test	RadiologyR7.txt	Radio
24	Test	RadiologyR8.txt	Radio

Fig. 5. Running the K-NN on the 24 Test Data.

The K-NN classifier miss categorized the case test 9 in which the predicted category was Autopsy instead of the real category of Cosmetic. This miss categorization case may be caused by using ill sensitive tokenization as the no letters separators were used for the purpose of tokenization. However, after using more sensitive tokenization such as identifying tokens based on the linguistic features, the accuracy have been raised to 97.5%. Moreover, one can enhance the accuracy further by choosing more careful document vector pruning such as the absolute pruning instead of the traditional perceptual pruning. The change has raised the accuracy to be 100%. It is also interesting to note the precision of categorizing each class of clinical narrative. Figures 6 and 7 illustrate the precision of categorizing two classes from the eight classes of the clinical narratives (Discharge Summaries and Chiropractic Reports).

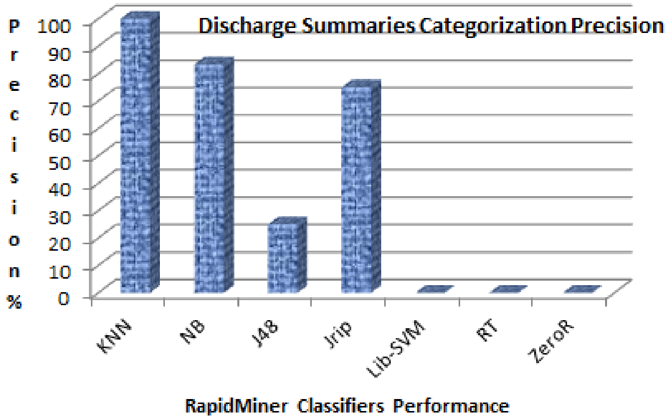

Fig. 6. The Categorization Precision of Distinguishing Discharge Summaries from the rest of Clinical Narratives.

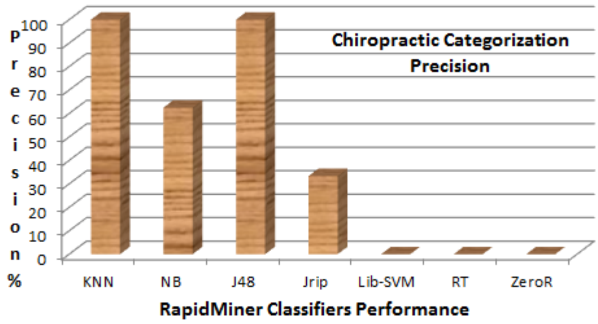

Fig. 7. The Categorization Precision of Distinguishing Chiropractic Reports from the rest of Clinical Narratives.

Moreover, the class recall is another classifier performance measure that provides information on the goodness of a classifier. Figures 8 and 9 illustrated the class recall of two classes (Discharge Summaries and Chiropractic Reports).

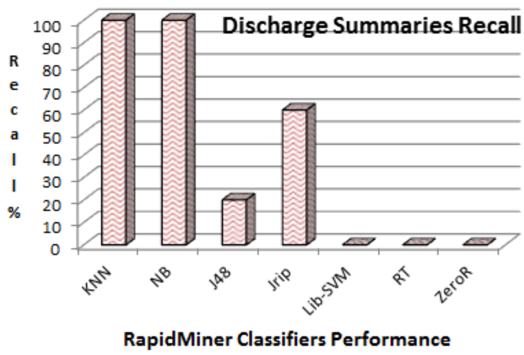

Fig. 8. The Discharge Summary Recall Measure.

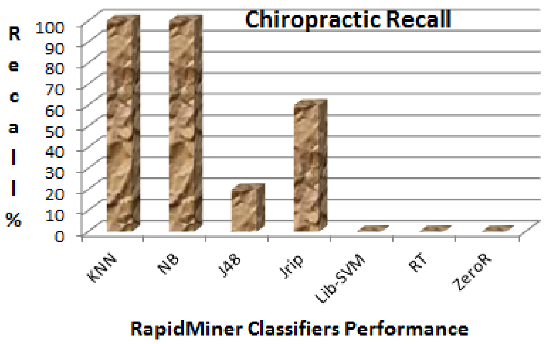

Fig. 9. Chiropractic Recall Measure.

Both the precision and recall identified K-NN to be the best compared to the other classifiers. This is an encouraging result for categorizing clinical narratives. However, one may argue that the classifiers did perform well because there are major differences between the token varieties used by each of the eight different clinical classes. This might be quite true and for this purpose we decided to use the most successful classifier like the K-NN and test its categorization ability when we use rather closely related clinical documents.

4 Validating the Categorizing Ability of the K-NN

In order to validate the ability of any categorization classifier we need to use sound dataset that can be compared to the achievements of other attempts. For this purpose, we used the i2b2 smoking dataset[8] which provides clinical narratives in five different classes as judged by a human expert (Current Smoker, Smoker, Past Smoker,

[8] https://www.i2b2.org/NLP/DataSets/Main.php

Non_smoker, Unknown) [9]. The clinical narratives in this dataset share many similar token sets which make it hard for an automatic categorization system to predict the correct category of test data. For simplicity we focused on two categories (Current Smoker and Non-Smoker) and extracted 48 narratives for each of the two categories (see Table 1) as well as 13 narratives test cases (see Figure 10). After running the K-NN classifier on this selected dataset, the performance measures were as follows:

Accuracy: 80.36%
Classification Error: 19.69%
Kappa: 0.616
Average Class Precision: 85.39%
Average Class recall: 81.25%
Absolute Error: 0.196
Relative Error: 19.69
Correlation: 0.661

Table 1: i2b2 Training Smoking Dataset Sample.

Nonsmoker	Current Smoker	Nonsmoker	Current Smoker
696	641	761	130
710	681	764	223
714	704	766	236
716	757	777	241
718	786	794	260
742	872	799	265
759	874	823	284
839	535	552	346
862	540	570	352
212	543	571	370
249	563	573	85
879	564	577	130
888	565	586	845
896	585	600	515
899	602	603	562
907	626	614	633
913	643	617	906
519	681	627	109
530	25	628	1
542	133	629	220
547	328	630	151
551	406	639	202
640	31	9	214
27	43	36	73

Figure 10 illustrates how the K-NN performs in predicting the 13 unknown cases.

Row No.	label	metadata_file	prediction(l...
1	STest	R643_S.txt	Smoker
2	STest	R704_S.txt	Smoker
3	STest	R757_S.txt	Smoker
4	STest	R865_N.txt	Nonsmoking
5	STest	R868_N.txt	Nonsmoking
6	STest	R872_S.txt	Nonsmoking
7	STest	R874_S.txt	Nonsmoking
8	STest	R888_N.txt	Nonsmoking
9	STest	R896_N.txt	Nonsmoking
10	STest	R899_N.txt	Nonsmoking
11	STest	R906_S.txt	Smoker
12	STest	R907_N.txt	Nonsmoking
13	STest	R913_N.txt	Nonsmoking

Fig. 10. Categorizing the Class of the 13 Cases.

Only two cases were miss categorized by our K-NN. This result represent a good one although it was weaker than the results on the eight clinical categories dataset. However, since the i2b2 smoking dataset is a public one, there are many attempts to use classifiers for categorizing clinical documents for the categories related to smoking. Ozlem Uzuner [10] published these attempts and their accuracy measures. Figure 11 illustrates the comparison of the average precision and recall in categorizing two different classes (Current Smoker vs Non_Smoker) using our K-NN method and 11 other attempts. Interestingly, our K-NN categorization method showed higher precision and recall than any other approach.

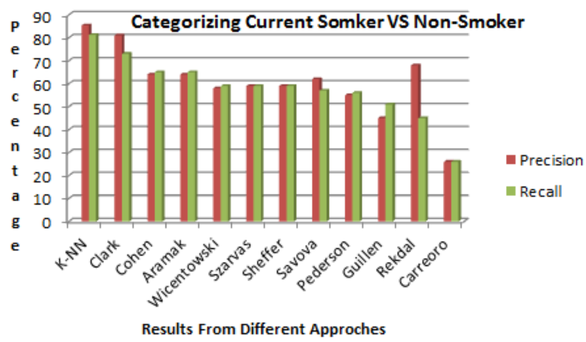

Fig. 11. Comparing K-NN with Other Notable Approaches in Categorizing Clinical Narratives for Current Smoker and Non-Smoker.

5 Discussion and Conclusion

This article demonstrates how clinicians can use visual programming tool like the RapidMiner to tokenize and categorize clinical narratives. Clinicians can flexibly drug and past variety of visual operators and change their behaviors' via parameterizations even for complex processes like tokenization and classifications. Several experiments have been conducted for this purpose that reveal major findings for categorization of clinical narratives. For example K-NN classifier outperform other classifiers in categorizing diverse clinical reports including those narratives that include high degree of similarity like the smoking vs nonsmoking discharge summaries. However, this work is only our initial attempt as we are intending to enrich the categorization process with higher sense of context-awareness. The next step that we are currently experimenting with is to enable clinicians through RapidMiner to incorporate ontologies in the process of tokenizing and categorizing clinical narratives. This extension is quite possible with the RMonto plugin for RapidMiner. Clinicians can develop their own ontologies as well as to use an existing one. This work is left to our next research work.

References

1. Illés Solt, Domonkos Tikk, Viktor Gál and Zsolt T. Kardkovács, Semantic Classification of Diseases in Discharge Summaries Using a Context-aware Rule-based Classifier, J Am Med Inform Assoc. 2009 Jul-Aug; 16(4): 580–584.
2. Jing Jiang, ChengXiang Zhai, An empirical study of tokenization strategies for biomedical information retrieval, Information Retrieval Journal, October 2007, Volume 10, Issue 4-5, pp 341-363
3. Hassler M, Fliedl G: Text preparation through extended tokenization. In Data Mining VII: Data, Text and Web Mining and Their Business Applications. Volume 37. Edited by Zanasi, A and Brebbia, CA and Ebecken, NFF. WIT Press/Computational Mechanics Publications; 2006::13-2
4. Alexandre Manuel Fajardo Vicente, LexMan: a Tokenizer and Morphological Analyzer with Transducers, https://fenix.ist.utl.pt/downloadFile/395145514800
5. K. Tomanek, J. Wermter, and U. Hahn. Sentence and token splitting based on conditional random fields. In PACLING 2007 – Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 49–57. Melbourne, Australia, September 19-21, 2007. Melbourne: Pacific Association for Computational Linguistics, 2007.
6. Linlin Li, Benjamin Roth and Caroline Sporleder, Topic models for word sense disambiguation and token-based idiom detection, Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Pages 1138-1147, PA, USA 2010.
7. Patrick Ruch et al,Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models, Proceedings o/CoNLL-2000 and LLL-2000, pages 111-114, Lisbon, Portugal, 2000.
8. Joshi, A. K., & Srinivas, B. (1994). Disambiguation of super parts of speech (or supertags): almost parsing. In Proceedings of the 15th Conference on Computational Linguistics - volume 1 (pp. 154–160). Kyoto, Japan.
9. Deryle Lonsdale, Rebecca Madsen, Unifying language modeling capabilities for flexible interaction, 9th International Conference on Spoken Language Processing (ICSLP 2006) Sep 17-21, 2006, Pittsburgh, Pennsylvania.
10. Uzuner Ö., Goldstein I, Luo Y, Kohane I. "Identifying patient smoking status from medical discharge records". J Am Med Inform Assoc. 2008; 15(1)15-24.

