Investigating the Role of Preprocessing and Attribute Selection Methods Towards the Performance of Classification Algorithms on News Dataset

Wateen A. Aliady

***Abstract*—** This paper uses two datasets Reuters and 20NewsGroup to analyze the impact of text preprocessing steps like tokenization, stemming and stopwords removal on classification results. In addition, it studies the effect of unigram, bigram and trigram attribute on classification results. Furthermore, it studies the impact of attribute selection methods on the generated number of attributes and classification accuracy. The paper analyzes the effect of the previous based on six classification algorithms. The results show that there is a positive impact of text preprocessing techniques on the used datasets on terms of classification performance accuracy. In addition, the unigram achieved the best results because there was an associated stop words removal list unlike the bigram and trigram. Furthermore, attribute selection methods can have positive impact on the performance of text classification algorithms but choosing the best attribute selection algorithm is dependent on the dataset used.

***Keywords*—** Correlation Based Feature Selection, Chi-squared, Naive Bayes, Support Vector Machine, Sequential minimal optimization, Decision Tree, and HyperPipes.

I. INTRODUCTION

One of the vital jobs done by data mining experts is classification. Classification is a process of grouping instances into a certain class for a general understanding [1]. Several classification algorithms have been proposed by researchers. It is a known fact that there is no classification algorithm that is suitable for all types of data [2].

Text classification is performed on the basis of several steps [3]. According to [3] the first step in text categorization is the collection of text documents in different formats. The following step is about conversion of these different text files (html, sgml, txt etc) into single acceptable format. Afterwards, these files are indexed into unified documents. After that, the selection of features is performed that has a great effect on the classification results. It should be noted that there are several feature selection algorithms or methods. After that, the classification algorithms are applied. The final stage is to evaluate the performance of algorithms using different evaluation measures.

The internet is the main source of data production and these data is needed in numerous businesses. For example, News website update their website with news on a minute basis. Some businesses need this information to be classified into different categories i.e. sports, entertainment or politics. Data produced on social networks like Facebook, Twitter, Instagram and Linkedin has a great importance. Reviews on products and comments on social media posts always carry significance for marketing and even political purpose. Text classification is

Wateen Aliady , Riyadh, SA (e-mail: wateen.aliady@gmail.com).

also used for email detection like if the email is spam or not. Human beings can understand linguistic structures and their meanings easily, but machines are not successful enough on natural language comprehension.

There are two types of text classifiers: the first type is the supervised classifier that splits the data into two set: training set and testing set. The second type is the unsupervised classifier that do not need any training data where there is no labeled data. Each of these need attributes or features on the basis of which they classify the text documents. These attributes have an impact on the classification results accuracy. Therefore, picking the best features, or attributes that provides more information is essential [4].

The motivation an aim for this research is to study:

- The impact of text preprocessing techniques on the classification algorithms performance in terms of accuracy.
- The impact of unigram, bigram and trigram attributes on the results of text classification algorithms.
- How far the attribute selection algorithms are useful in achieving high classification accuracy?

II. Dataset and Methods

A. *Dataset*

In this research two popular datasets are used: 20Newsgroups and Reuters 21578 for classification. The 20 Newsgroups was collected by Ken Lang [5]. The documents in the Reuters-21578 collection was presented by Reuters Ltd. in 1987[6]. In total more than 2000 text documents are used in these experiments, where each dataset contains more than 1000 text documents. Each dataset has been treated in 12 different ways which means 12 versions have been created for each dataset. In total 24 data versions have been used in this experiment. Table I and II presents the details of each dataset.

B. *Tool*

WEKA is a famous tool which has built in implementations of data mining and machine learning algorithms. It is one of the mostly used machine learning tool by researchers. Weka stands for Waikato Environment for knowledge analysis developed by University of Waikato, Newlands. [7]. Weka is free available tool for text classification and machine learning purpose. It can be used in two ways, command line and Graphical user interface.

C. *Text Preprocessing*

Different text preprocessing methods has been used to prepare that data sets. These preprocessing steps has great impact on the results of classifiers. Figure 1 presents the preprocessing steps applied in this work. It consists of several steps which help in purifying the data and get it ready to be used for classification.

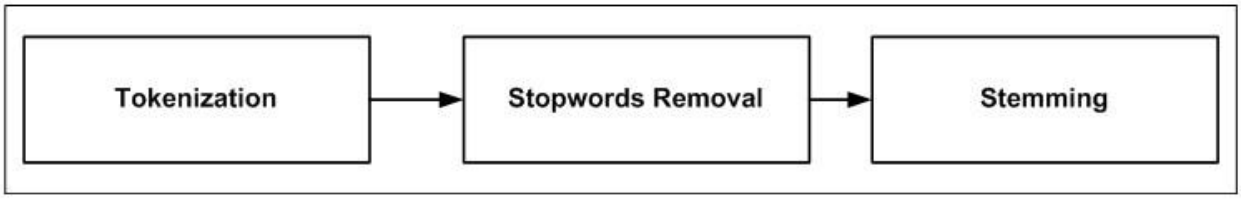


Figure 1: Text Preprocessing Steps

Tokenization is the process of splitting down the array of text in smaller chunks, i.e. words or phrases. The text in the used dataset for this work consists of news articles and they contain impurities that is the machine learning algorithms cannot understand the whole sentence like we human can understand. In order to make them ready to be used in machine learning algorithms, the sentences must be broken into single words or phrases [8].

As stated in the beginning that this paper studies the impact of different types of attributes, and therefore the data is treated within the following attribute types: the unigram type that is those attributes that consists of a single word or term. The only problem with unigram attributes is that when a document is divided in to single word it generates a huge vocabulary size. The bigram is those attributes which consists of two terms. Although it resolves the problem of the unigram attribute of huge vocabulary size and thus resolve the time and space complexity, it has its own problem. It is time consuming to develop a stop words as there is no stopwords list available for bigrams in Weka. The trigram is when the attribute contains three terms. The trigram further decreases the number of attributes but again we have no stopwords list for trigrams [9].

Stop words are those words which carry no special meaning by itself unless added to another word which provide sense to the sentence [8]. In order to understand their role, different datasets are created which is called data versions, where in some of these versions stops words were removed while in others were not. In this paper, no stopwords removal list is designed, but the build in list was used.

Stemming is another important preprocessing step in which the term is reduced to its root. This will also help in reducing the size of attributes. Different algorithms have been proposed by number of researches like Potter Stemmer and Lovin stemmer. In this study the Lovin stemmer is used [10].

Data version creation is an important step to understand the behavior of text preprocessing techniques in detail. The idea comes from the fact that it helps in observing the preprocessing impact on each data version separately. The total number of attributes are presented in the last two columns for the two-news dataset in table 1.

TABLE I
DATA VERSIONS AND NUMBER OF ATTRIBUTES

Dataset	Stemming	Stopwords removal	Tokenization	No: of attributes 20NesGroups	No: of attributes Reuters21578
D1	✓	✓	Unigram	1479	1692
D2	✓	✓	Bigram	1763	2071
D3	✓	✓	Trigram	2207	2361
D4	×	×	Unigram	1667	1854
D5	×	×	Bigram	1746	2077
D6	×	×	Trigram	2083	2184
D7	✓	×	Unigram	1565	1693
D8	✓	×	Bigram	1763	2071
D9	✓	×	Trigram	2207	2361
D10	×	✓	Unigram	1551	1689
D11	×	✓	Bigram	1746	2077
D12	×	✓	Trigram	2083	2184

D. *Attribute Selection Methods*

In this study, three attribute selection methods are used. This first attribute selection method**,** **Correlation Based Feature Selection (CFS)**, is based on feature correlation introduced by Hall [11]. This attribute selection algorithm selects those classification feature which have high correlation with the class while they have uncorrelated with each other. This algorithm is tested on different datasets and it shows that it eliminates irrelevant, redundant and noisy feature. It may degrade the performance of classifier when deleting the useful attribute. But in this study, it is explored to study its performance on text dataset.

The second attribute selection method is called the **Chi-squared**. It is a probabilistic model for selecting an appropriate set of features for classification purpose. It sees the relationship between the attribute and class. It is also called a statistical model. It was proposed by Liu and Setiono [12].

The third attribute selection method called **FilteredSubsetEval** is filter subset of attribute are evaluated. This algorithm is implemented in Weka and is used from there.

The impact of these attribute selection methods can be observed from the value presented in the Table II, presenting 20NewsGrousp and Reuters 21578 respectively. It is clear that the number of attributes are reduced dramatically when using these attribute selection methods.

TABLE II
NUMBER OF ATTRIBUTES AFTER APPLYING ATTRIBUTE SELECTION METHODS

Dataset Properties		**Total number of attributes after applying attribute selection methods (20 NewsGroups)**			**Total number of attributes after applying attribute selection methods (Reuters)**		
Dataset	**Total Attributes**	**Chisquared**	**CFS**	**FSE**	**Chisquared**	**CFS**	**FSE**
D1	1479	859	52	15	1177	45	22
D2	1763	1127	56	59	1725	54	33
D3	2207	1845	53	139	2156	58	40
D4	1667	1102	56	30	1300	46	43
D5	1746	1150	51	66	1765	50	39
D6	2083	1762	51	117	2002	56	62
D7	1565	940	53	15	1163	45	22
D8	1763	1127	56	59	1725	54	33
D9	2207	1845	53	139	2156	58	40
D10	1551	923	54	41	1249	51	10
D11	1746	1150	51	66	1765	50	39
D12	2083	1762	51	117	2002	56	62

E. *Classifiers*

The first classifier used is **Bayes net** provide a graphical structure, showing the dependencies among different variable. Each node in the graph present a variable while the connection/arcs shows the relationship among the variables. This classifier follows the probabilistic model that shows all the possible states of domain.

The second classifier used is **Naive Bayes** classifier that is one of the simple classifiers that belongs to Bayesian classifier family [13], based on Bays Theorem. This classifier represents the data as vector attribute values, whereas the labels of class are pinched from finite set of data. This method is used for labeling the dataset instances.

The third classifier used is **Support Vector Machine (SVM)** is among the most widely used algorithms for text classification and machine learning purpose. This categorization was introduced by two data scientists Vapnik [14] and Joachims [15]. For the first time it was used for text classification purpose.

The fourth classifier used is **Sequential minimal optimization (SMO)**, developed by John Platt [16] at Microsoft Research. This classifier was invented for improving the SVM classification algorithm. The implementation of this algorithm is in Libsvm and also used for svm training.

The fifth classifier used is **Decision Tree (J48)** classifier that is one of the most famous and effective decision tree classification algorithms. It was developed by Quinlan [17]. Furthermore, it works on information gain and the attribute with high information gain appears on the top of tree by recursively dividing the attributes into subsets by the normalized information gain.

The sixth classifier used is **HyperPipes** classifier is among the fastest and simplest classification algorithms [18]. It is a straightforward classifier and make it ensures consistency for each attribute. The HyperPipes also contain the bounds for the values of attribute [19].

III. Experiment And Results

There is a positive impact of text preprocessing techniques on the classification algorithms performance in terms of accuracy. The text preprocessing has positive effect in many cases while in some cases it has negative effect. The main reason for construction different versions of datasets is to know the impact of these text preprocessing techniques individually. Here in this study, three most popular techniques are used i.e. tokenization, stemming and stopwords removal

As for the tokenization process the results were that using stemming and stopwords removal on 20Newsgroups dataset has positive effect on the performance of classifiers in terms of accuracy. Most of the algorithms that is four out of six achieve high performance in terms of accuracy when treated with stemming and stopwords removal. On the other hand, it has been observed that the Reuters dataset, algorithms performs differently. The behavior shows that stopwords removal has negative impact on the results of classifiers whereas stemming proved itself positive when it comes to enhance the performance of categorization algorithms.

Figure 2 presents the difference between applying preprocessing technique and without applying preprocessing techniques on text. It shows the accuracy results of all six classification algorithms with total number of attributes, which mean that no attribute selection methods have been used. It can be seen from the below graph except J48 that all the algorithms perform well on unigram while trigram performs worst.

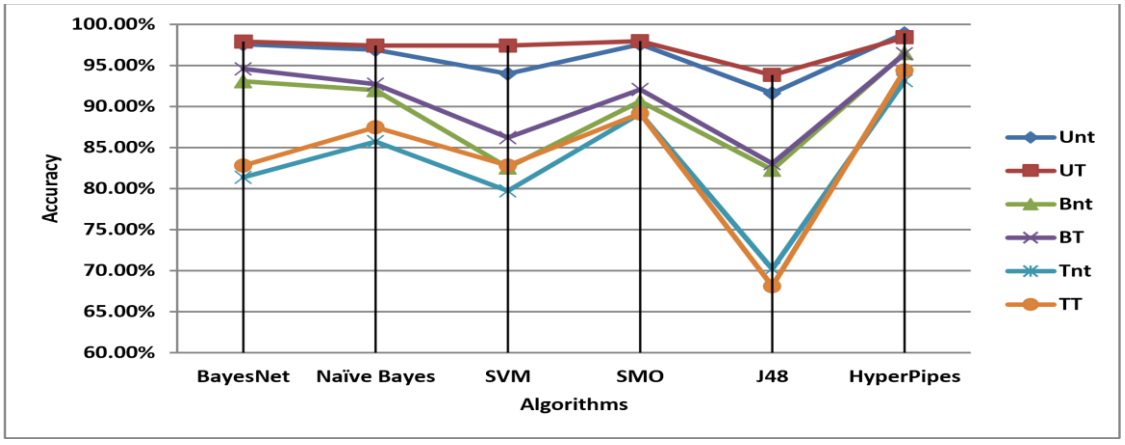


Figure 2: Difference between applying preprocessing methods and without using preprocessing methods on 20NewsGroups dataset

In the above graph on the y-axis accuracy is presented while the x-axis shows text classifiers. Other abbreviations in the graph are as in Table III.

TABLE III
SYMBOLS FOR RESULTS GRAPHS

Symbol	Stand For
Unt	Unigram attribute without text preprocessing
UT	Unigram attribute with text preprocessing
Bnt	Bigram attribute without text preprocessing
BT	Bigram attribute with text preprocessing
Tnt	Trigram attribute without text preprocessing
TT	Trigram attribute with text preprocessing

.

The impact of unigram, bigram and trigram attributes on the results of text classification algorithms can be shown in figure 2 above. Unigram attribute proved to be efficient because it increases the accuracy of classifiers. It also illustrates that unigram feature achieved high accuracy. One reason for this result is that there is no stopwords list for bigram and trigram. This can cause increase in the attribute list which according to some research degrade the performance of classifiers as shown in figure 3.

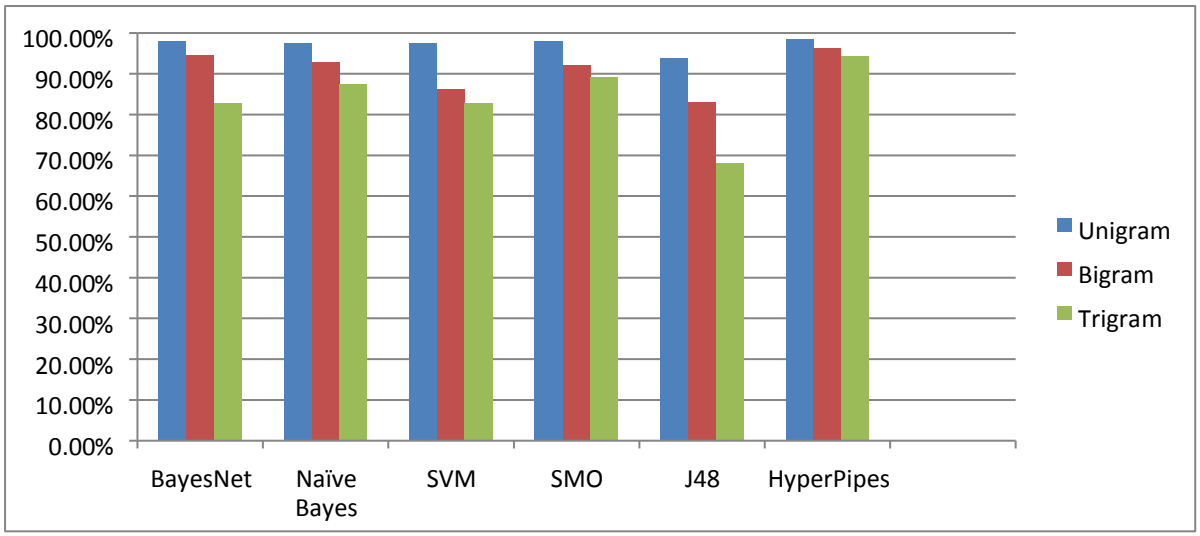


Figure 3: Classifiers accuracy level when using unigram, bigram and trigram attributes

The effect of attribute selection algorithms in achieving high classification accuracy could be splinted into two parts. The first is the impact of feature selection algorithms on the number of attributes presented in figure 4. The second part of this question is to investigate the impact on performance of classifiers, it can be shown in Table IV below.

Figure 4 below shows that attributes decrease dramatically when using CFS and FSE feature selection methods while Chi-square do not reduce too much attribute from the attribute list. The blue line presents the total number of attributes before applying the attribute selection methods.

Chi-squared enhances the performance of Baysnet, Naivebayes and HyperPipes performance for 20NewsGroups dataset and Chi-Squared also achieve high accuracy for HyperPipes using Reuters dataset. SVM and SMO perform well on total number of attribute when using 20Newsgroups dataset. CFS attribute selection has a positive impact on the performance of J48 decision tree, Baysnet, and SVM when using Reuters Dataset. FSE attribute selection method achieve high accuracy on Naive Bayes, and SMO when using Reuters dataset. Hence it proves that attribute selection methods can have positive impact on the performance of text categorization algorithms.

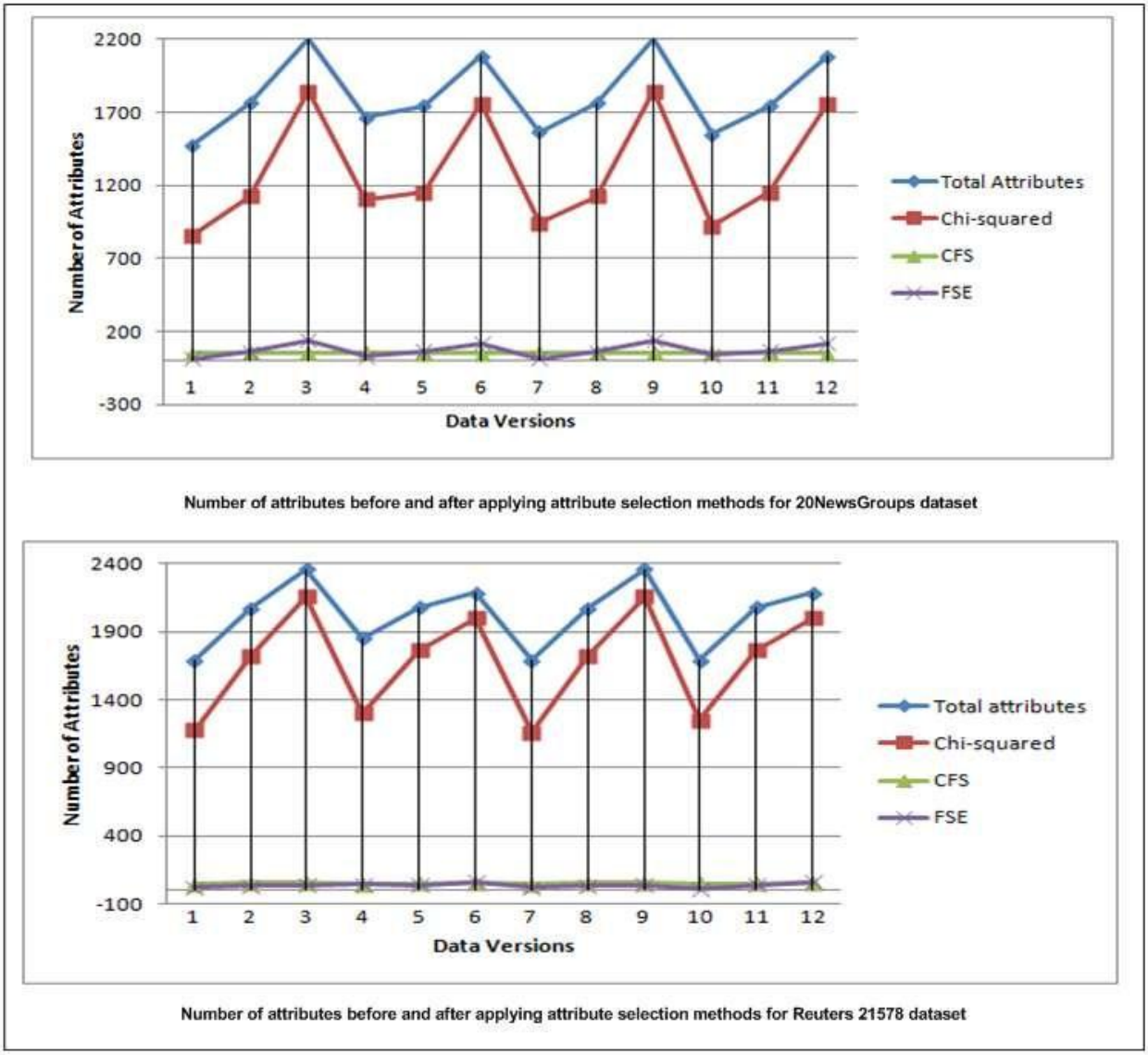


Figure 4: Different between number of attributes before and after applying attribute selection methods

TABLE IV
Data Versions and Attribute Selection Method having higher accuracy score

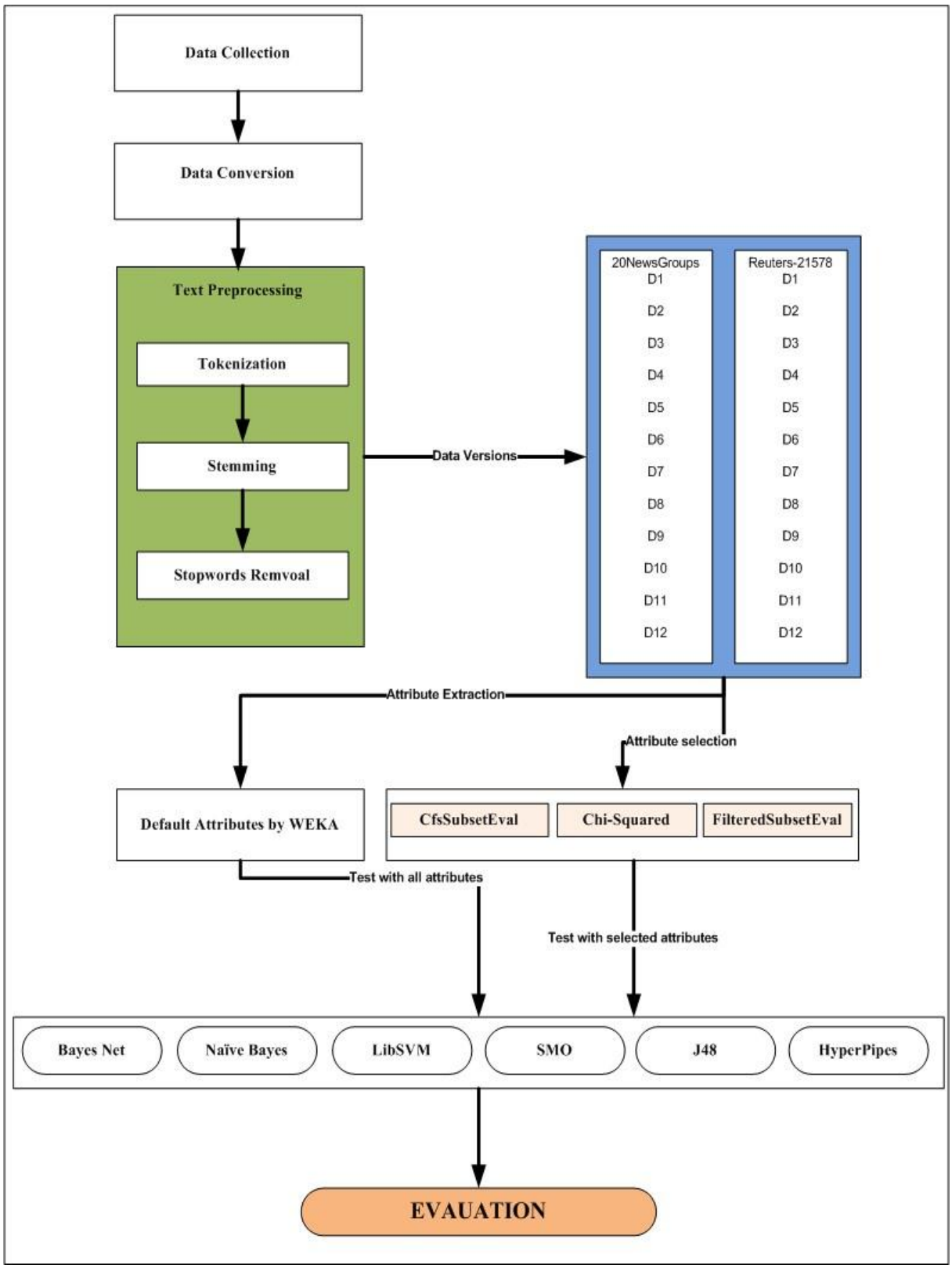
Classifier	20NewsGroups		Reuters21578	
	Attribute Selection Method	**Data Version**	**Attribute Selection Method**	**Data Version**
Baysnet	Chi-Squared	D1	CFS	D7
Naive Bayes	Chi-Squared	D1	FSE	D7
SVM	Total Attributes	D1	CFS	D7
SMO	Total Attributes	D7	FSE	D7
J48	CFS	D7	CFS	D7
HyperPipes	Chi-Squared	D4	Chi-Squared	D8

IV. CONCLUSION

This paper aims to study the impact of text preprocessing techniques on the performance of classification algorithms in terms of accuracy. Furthermore, it analyzes the impact of unigram, bigram and trigram attributes on the classification result values. Nonetheless, it studies the impact of attribute selection algorithms on classification accuracy. It uses two datasets for this purpose: 20Newsgroup and Reuters-21578 datasets. Figure 5 sums up all the work implemented work in Weka tool, where it presents the sequential order of this work as it goes through text preprocessing to generate 12 data versions for each data set. Then, attribute selection and classification are performed.

To conclude, there was a positive impact of text preprocessing techniques on the used datasets on terms of classification performance accuracy. In addition, the unigram achieved the best results because there was an associated stop words removal list unlike the bigram and trigram. Furthermore, attribute selection methods can have positive impact on the performance of text classification algorithms but choosing the best attribute selection algorithm is dependent on the dataset used.

Figure 5: Methodology



REFERENCES

[1] K. Nalini and L. Jaba Sheela, "Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering, vol. 1, no. 6, 2014. [Accessed 10 March 2019].

[2] A. Gupte, S. Joshi, P. Gadgul and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologie, vol. 5, no. 5, 2019. [Accessed 10 March 2019].

[3] J. Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification," vol. 6, no. 2, pp. 126–129, 2016.

[4] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 2, 2013. Available: 10.14569/ijarai.2013.020206.

[5] K. Lang, "Newsweeder: Learning to filter Netnews," in Proceedings of the 12th international conference on machine learning, 1995, pp. 331–339

[6] D. D. Lewis. Reuters-21578 text Categorization test collection. Distribution 1.0. README file (version 1.2). Manuscript, September 26, 1997

[7] WEKA, "WEKA: Data mining and Machine learning tool," 2013.

[8] V. Gurusamy, Vairaprakash & S.Kannan, Subbu. "Preprocessing Techniques for Text Mining", Proc. RTRICS,2014

[9] S. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson and S. Lin, "Application of clinical text data for phenome-wide association studies (PheWASs)", Bioinformatics, vol. 31, no. 12, pp. 1981-1987, 2015. Available: 10.1093/bioinformatics/btv076.

[10] S. Vijayarani, M. Nithya and J. Ilamathi, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, vol. 5, no. 1, 2019. [Accessed 10 March 2019].

[11] M. a Hall, "Correlation-based Feature Selection for Machine Learning," Methodology, vol. 21i195-i20, no. April, pp. 1–5, 1999.

[12] R. Setiono, "Chi2: feature selection and discretization of numeric attributes," Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. pp. 388–391, 1995.

[13] M. Friedman, Nir and Geiger, Dan and Goldszmidt, "Bayesian network classifiers," Mach. Learn., vol. 29, no. 29, pp. 131--163, 1997.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.

[15] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in Proceedings of the 10th European Conference on Machine Learning, 1998, pp. 137–142.

[16] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," Adv. kernel methods, pp. 185--208, 1999.

[17] J. Quinlan, C4. 5: programs for machine learning, vol. 240. Elsevier, 1993.

[18] E. Witten, Ian H and Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[19] S. Alvestad, "Early warnings of critical diagnoses," Institutt for datateknikk og informasjonsvitenskap, 2009.

APPENDIX

A. *Bayes Net Classifier*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	**97.91%**	**97.91%**	97.49%	95.82%	88.14%	88.41%	98.42%	97.86%
DV2	94.57%	94.57%	91.06%	92.23%	91.56%	91.56%	95.55%	94.53%
DV3	82.80%	82.80%	79.88%	83.55%	91.38%	91.38%	89.52%	88.41%
DV4	97.57%	97.57%	97.32%	95.99%	**95.18%**	**95.18%**	97.12%	**98.33%**
DV5	93.07%	93.01%	89.06%	91.90%	91.84%	91.84%	94.25%	93.32%
DV6	81.38%	81.38%	79.63%	81.96%	89.24%	89.24%	94.06%	94.71%
DV7	97.82%	97.82%	97.49%	95.82%	92.30%	92.30%	**98.70%**	98.23%
DV8	94.57%	94.57%	91.06%	92.23%	91.56%	91.56%	95.55%	94.53%
DV9	82.80%	82.80%	79.88%	83.05%	91.38%	91.38%	89.52%	88.41%
DV10	97.49%	97.74%	**97.57%**	**97.32%**	88.22%	88.22%	97.47%	94.06%
DV11	93.07%	93.07%	89.06%	91.90%	91.84%	91.84%	94.25%	93.32%
DV12	81.38%	81.38%	79.63%	81.96%	89.24%	89.24%	94.06%	94.71%

B. *Naive Bayes*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	**97.41%**	**97.57%**	97.27%	95.82%	88.14%	88.78%	98.70%	98.51%
DV2	92.73%	94.57%	90.15%	91.73%	91.56%	92.21%	94.71%	94.43%
DV3	87.49%	88.48%	78.71%	82.22%	91.38%	**94.62%**	90.26%	88.04%
DV4	96.91%	97.16%	96.99%	95.90%	**95.18%**	95.73%	96.94%	97.96%
DV5	91.98%	93.82%	87.98%	90.98%	91.84%	92.02%	96.29%	93.24%
DV6	85.72%	87.06%	77.37%	79.54%	89.24%	94.06%	93.41%	93.97%
DV7	97.07%	97.41%	**97.32%**	95.82%	92.30%	93.04%	**98.88%**	**98.98%**
DV8	92.73%	94.57%	90.15%	91.73%	91.56%	92.21%	94.71%	94.43%
DV9	87.47%	88.48%	78.71%	82.22%	91.38%	**94.62%**	90.26%	88.04%
DV10	96.66%	97.07%	97.16%	**96.82%**	88.22%	89.06%	96.94%	94.62%
DV11	91.98%	93.82%	87.98%	90.98%	91.84%	92.02%	96.29%	93.23%
DV12	85.72%	87.06%	77.37%	79.54%	89.24%	94.06%	93.41%	93.97%

C. *SVM Classifier*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	97.41%	97.32%	97.32%	95.74%	88.14%	98.79%	99.07%	98.51%
DV2	86.22%	88.06%	83.30%	86.14%	91.56%	91.10%	97.59%	94.43%
DV3	82.80%	83.72%	66.77%	72.53%	91.38%	70.62%	94.34%	88.04%
DV4	93.99%	94.90%	96.74%	94.99%	95.18%	96.94%	98.05%	97.96%
DV5	82.55%	83.97%	85.80%	86.97%	91.84%	91.47%	96.10%	93.24%
DV6	79.71%	77.79%	70.45%	70.78%	89.24%	96.23%	91.84%	93.97%
DV7	96.32%	96.74%	97.07%	95.74%	92.30%	98.98%	99.25%	98.88%
DV8	86.27%	88.06%	83.30%	86.14%	91.56%	91.10%	97.57%	94.43%
DV9	82.80%	83.72%	70.78%	72.53%	91.38%	70.62%	94.34%	88.04%
DV10	96.57%	96.91%	96.82%	96.07%	88.22%	98.23%	97.03%	94.62%
DV11	82.55%	83.97%	85.80%	86.97%	91.84%	91.47%	96.10%	93.23%
DV12	79.71%	77.79%	70.45%	70.78%	89.24%	69.23%	91.84%	93.97%

D. *SMO Classifier*

Dataset	**20NewsGroups Cross Validation @10 Fold Accuracy**				**Reuters-21578 Cross Validation @10 Fold Accuracy**			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	97.97%	97.41%	97.07%	95.65%	88.14%	99.35%	99.44%	99.19%
DV2	92.07%	94.49%	90.06%	90.98%	91.56%	98.79%	97.77%	91.75%
DV3	89.14%	91.37%	83.72%	90.06%	91.38%	95.05%	93.97%	91.75%
DV4	97.57%	97.41%	97.74%	95.74%	**95.18%**	99.16%	97.49%	95.60%
DV5	90.65%	92.15%	88.89%	91.40%	91.84%	98.60%	95.92%	96.20%
DV6	89.14%	90.31%	81.88%	87.72%	89.24%	97.40%	95.77%	93.60%
DV7	**97.99%**	**97.91%**	97.41%	95.65%	92.30%	**99.53%**	**99.44%**	99.25%
DV8	92.07%	94.49%	90.06%	90.98%	91.56%	98.79%	97.77%	**99.77%**
DV9	89.14%	91.73%	83.72%	90.06%	91.38%	97.33%	93.97%	91.75%
DV10	96.49%	97.24%	**97.82%**	**96.82%**	88.22%	99.07%	98.51%	94.71%
DV11	90.65%	92.15%	88.89%	91.40%	91.84%	98.60%	95.92%	96.20%
DV12	89.14%	90.31%	81.88%	87.72%	89.24%	97.40%	95.27%	93.60%

E. *J48*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	**93.82%**	**95.15%**	94.74%	**94.65%**	88.14%	96.01%	96.84%	96.38%
DV2	83.05%	84.64%	77.71%	77.39%	91.56%	96.75%	95.55%	95.92%
DV3	68.11%	68.61%	78.71%	66.77%	**91.38%**	94.25%	84.61%	84.80%
DV4	91.62%	93.23%	93.07%	92.90%	95.18%	96.38%	95.27%	95.82%
DV5	82.30%	84.22%	79.38%	79.21%	91.84%	95.18%	94.06%	94.25%
DV6	70.20%	70.11%	65.77%	65.94%	89.24%	93.88%	87.95%	89.62%
DV7	**93.82%**	94.24%	**95.49%**	**94.65%**	92.30%	96.38%	**97.49%**	**97.03%**
DV8	83.05%	84.64%	77.71%	77.37%	91.56%	**96.75%**	95.55%	95.92%
DV9	68.11%	68.61%	66.77%	66.77%	91.38%	94.25%	84.61%	84.80%
DV10	92.57%	92.07%	93.65%	94.32%	88.22%	95.73%	96.20%	93.69%
DV11	82.30%	84.22%	79.38%	79.21%	91.84%	95.18%	94.06%	94.25%
DV12	70.20%	70.11%	65.77%	65.94%	89.24%	93.88%	87.95%	89.62%

F. *Hyperpipes*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	98.41%	98.49%	94.24%	86.89%	88.14%	96.10%	76.27%	75.16%
DV2	96.41%	96.57%	86.47%	89.64%	91.56%	98.70%	90.36%	84.98%
DV3	94.40%	94.40%	81.55%	90.40%	91.38%	97.77%	89.24%	84.70%
DV4	**98.91%**	**98.99%**	**95.40%**	92.32%	**95.18%**	98.33%	87.95%	**89.06%**
DV5	96.49%	96.66%	86.64%	89.64%	91.84%	96.94%	87.58%	82.11%
DV6	93.15%	93.23%	79.46%	87.81%	89.24%	96.20%	86.56%	87.85%
DV7	98.33%	98.41%	94.65%	86.89%	92.30%	96.38%	78.03%	76.16%
DV8	96.41%	96.57%	86.47%	89.64%	91.56%	**98.70%**	**90.36%**	84.98%
DV9	94.40%	94.40%	81.55%	90.64%	91.38%	97.77%	89.24%	84.70%
DV10	98.66%	98.74%	95.07%	**94.49%**	88.22%	98.14%	80.74%	50.13%
DV11	96.40%	96.66%	86.64%	89.64%	91.84%	96.94%	87.58%	82.11%
DV12	93.15%	93.23%	79.46%	87.81%	89.24%	96.20%	86.56%	87.85%

.