AMHARIC TEXT NEWS CLASSIFICATION

[1]Misrak Assefa [2]Vishal Goyal
[1]MTech Computer Science, [2]Associate Professor Computer Science
Computer Science Department
Punjabi University Patiala, India

Abstract : Amharic language is the second most widely spoken Semitic language in the world. There is a number of news overloaded on the web. Searching some useful documents from the web on a specific topic which is written in Amharic language is a challenging task. Hence, document categorization is required for managing and filtering important information. Different researchers have done on Amharic document classification using machine learning approaches. However, there is still a gap in the domain of information that needs to be launch. This study attempts to design an automatic Amharic news classification using supervised learning mechanism on four un-touch classes. To achieve this research 4,182 news articles were used. Naïve Bayes (NB) and Decision tree (j48) algorithms were used to classify the given Amharic dataset. In this paper, k-fold cross-validation is used to estimate the accuracy of classifier. As the result shows those algorithms can be applicable in Amharic news categorization. The best average accuracy result is achieved by j48 decision tree and naïve Bayes is 95.2345 %, and 94.6245 % respectively using three categories. This research indicated that typical decision tree algorithm is more applicable to Amharic news categorization.

Keyword: Text categorization, supervised machine learning, Naïve Bayes, Decision tree.

I. INTRODUCTION

In the today's world, text is the main form of communication that accelerate communication parties to exchange news or idea in business, social, family and personal issue through email, ping, tweeter, etc. as the growth of technology becomes greater the amount of text(document) needed by the people also increases [1]. This led to the difficulties for the users to access required information from unstructured document [1]. Without formal organization retrieval of information is tedious and time killing. However, automatic text classification which is also known as text categorization is a solution to make documents structured and easy to execute.

Text classification is used to classify document to the various predefined class. It is a process of assigning text into one or more classes [2]. Text Classification is an active research area in text mining with increasing demand of managing large textual databases. Automatic text classification used two views in different literature. The first view of automatic text classification is to automatically identify a set of domains and classify documents under them. This is referred to as text clustering. The second view refers to the automatic assignment of documents to a predefined (pre-labelled) set of categories, which is referred to as text classification [3]. There are various applications in text classification such as sentiment analysis, degree readability, email routing, authorship attribution, automated survey coding, text filtering and word sense disambiguation. The main objective of this study is to travel through automatic Amharic text new classification based on un-touch domain. This research used NB and typical c4.5 decision tree classifier in Amharic news classification.

II. AMHARIC LANGUAGE

Amharic is the native language of people lives in the north-central part of Ethiopia [4]. It is the second most spoken Semitic language in the world (after Arabic) and morphologically rich national language of Ethiopia [5]. It has its own writing, punctuation and numbering system that developed from the script of Ethiopia's classical language, Ge'ez. In addition, it consists core of thirty-three characters each occurs in seven other forms called "Fidel" (□□□). Totally 231 characters used to represent Amharic writing system.

III. DOCUMENT PRE-PROCESSING

Data preparation is a process of making data suitable for the machine learning before classification performed. It is an important process to improve accuracy, efficiency, and scalability of classification process to get better experiment result. The first data preparation step is changing documents into text form. There are language dependent data per-processes i.e. stemming, and normalization which needs to know morphological principle of the language [1]. In this study tokenization, stop-word removal, punctuation and number removal, compound statement, stemming, normalization, and transliteration data preparation were performed using python programming language (python 2.7.1.5).

3.1 Feature Selection

Feature selection applied by keeping the most relevant variables from the original dataset based on term frequency. In this research experts have used Term Frequency-Inverse Document Frequency (TF-IDF) feature selection technique.

TF-IDF is a product of term weight and inverse document frequency which used to evaluate term weight across the dataset.

3.2 Term Frequency *(TF)*

In terms frequency approaches the computation of term weighting in each document is equal to the number of times term appears in the document. Often the term which has high-frequency value could not discriminate documents from the other.

Long document will have a higher advantage over the short document because it has repeated terms. To remove such problem normalization is used by the inverse document frequency method. TF is:

$$tfij = fij / max\{fij\} \quad (1)$$

where:

f_{ij} = frequency of term i in document j

tf_{ij} = term frequency in i^{th} and j^{th}

3.2.1 Inverse Document Frequency *(IDF)*

IDF reduces the weight of terms that occur very frequently in the dataset and increases the weight of terms that occur rarely. It computed as:

$$idfi = log2 (N/ df\ i) \quad (2)$$

where:

idf_i = inverse document frequency of term i,

N is the total number of documents

Finally, TF ×IDF computed as:

$$TF \times IDF = tfij * idfi \quad (3)$$

IV. RELATED WORK

Considering the evolution of information transfer and benefit of automatic news classification plenty of researches have been deployed on document classification. However, only few researches have been performed on Amharic document categorization and those research papers are summarized as follows:

Surafel (2003), he has tried to conduct supervised learning algorithm for Automatic Amharic text news classification using naïve bay's and kNearest Neighbour (kNN) algorithms. The researcher has performed categorization on sixteen domains. Namely, Accidents, International relations, Weather condition, List of Events, Health, Education, Sport, Law and Justice, Politics, Defence and Security, Other classification, Science & Technology, Social, Agriculture, Economy and Cultural affairs. He performed three experiments for model construction and model usage. among them, the best result obtained was on the first experiment that he performed on three categories (Social, Sport, and Education). The result was 95.80% and 89.61% by the naïve Bayes and kNN classifiers respectively. And the least outcome was seen in the third experiment on sixteen domains. The result has been 78.48% and 64.50% for naïve Bayes and kNN classifier. He included standard stop-word, spelling checker, stemmer and corpus as the future work on his paper. In addition, he recommended further research on Amharic text classification [6].

Lars Asker et al. (2009), have proposed Self-Organizing Map (SOM) model. Beyond the classification of Amharic web news, they had an experiment on the effects of stemmer on classification. The research conducted classification on Amharic web news in ten predefined domains. The domains are Sport, Hot news, Editorials, Politics, Business & Economy, Social, Culture, Science & Technology, Health, and Art. from the three experiments they have done, two of them where organized by SOM. Which are SOM predefine and query label which are used for the realization of first and second experiment. And the last experiment used rulebased classifier such as decision tree. The researcher designed the third experiment on decision tree classifier based on full, noun and stemmed data. The result obtained by SOM predefine was 72.9% and using SOM on query label was 69.5%. The experiment that had been employed by Decision trees algorithm was 69.4% in full, 68.9% nouns and 68.1% of stemmed data. In the remediation of the research, they highlighted the absence of standard corpus and miss match of the available stop-word list [7].

Animut (2012), he has designed and implemented semi-supervised approach in Amharic language. in this Amharic news classification research, ten domains are used. Namely economy, politics, sport, health, culture and tourism, science, social, education, law, and accident were comprised. In this paper, some improvements were seen in document category than the research done on automatic categorization of Amharic news text: a machine learning approach. However, there are still a lot of domains that need to be done. In this paper, three experiments were performed. The first experiment used four classes or domains.

The second experiment was performed on seven classes and the third experiment carried out with the whole(ten) classes using naive Bayes, Hyperpipes, and RBF Network algorithm. Moreover, the experiment addressed the comparison between supervised and semi-supervised approach. The best result confirmed by the supervised approach using four domains was 76.48%, 74%, 72.16% respectively. And the best result in semi-supervised was 83.44%, 82.8%, 82.4% respectively. The researcher recommended to explore document classification and clustering in research publication and email. In addition to this, the paper indicated further work to be done using another classifier algorithm that has a better performance with the minimum cost and high accuracy [1].

Worku (2013), he has proposed A neural network approach for text representation and categorization. This research used Learning Vector Quantization (LVQ) algorithm to classify the given document into nine categories. the domains are Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work. In this research, the TF and IDF of the feature extraction metrics were used for the training data. The

researcher conducted three experiments and the average result obtained by TF and IDF was 75.5% and 71.96% respectively. the research paper limited on standard stemmer, corpus, spelling checker, dimension reduction, and feature preparation [8].

Kamal et al. (2017), and Ramesh Babu (n.d.) they have proposed an approach for Affan Oromo language. The language is the most widely spoken regional language of Ethiopia. The researcher implemented news categorization on six domains those are agriculture, sport, education, politics, health, and business. This research used decision tree classifier and support vector machine algorithm.

This paper tried to select a suitable method for Affan Oromo text classification by comparing the above machine learning algorithms. According to the research result that obtained by decision tree was 96.58% and result by support vector machine algorithm was 84.93%. the researchers observed that the classification of Affan Oromo news is possible without using the highly complex feature reduction techniques [9].

4.1 Summary of Reviews

In the automatic text classification, different algorithms can be applied in different languages. However, during pre-processing and feature extraction it is required to know about language inflection or morphological properties. Because the technique that we used during document pre-processing and feature selection have an impact on the enhancements of performance. Furthermore, a number of domains and amount of data used in the training corpus also have an influence on accuracy. Moreover, every text classification system needs to classify the given documents to the provided classes.

Based on literature review it can be seen that researches have not been conducted decision tree classifier for Amharic document classification. Accordingly, we have used new methodology for unseen Amharic news category.

Table 1: Summary of Reviews in Categorization

No.	Author & year	Classification approach	Domains	Language	Accuracy	
1	Surafel (2003)	NB	Accidents, International relations, Weather condition, List of Events, Health, Education, Sport, Law and Justice, Politics, Defence and Security, Other classification, Science & Technology, Social, Agriculture, Economy and Cultural affairs	Amharic	78.48%	
		k-nearest neighour (kNN)			64.50%	
2	Lars Asker et al. (2009)	Self-organizing map	Sport, Hot news, Editorials, Politics, Business & Economy, Social, Culture, Science & Technology, Health and Art	Amharic	72.9%	
		Rulebased(decision)			69.4%	
3	Animut (2012)	Naïve Bayes	economy, politics, sport, health, culture and tourism, science, social, education, law, and accident	Amharic	supervised	Semi Supervised
					69.70%	55.42%
		Hyperpipe			46.53%	57.26 %
		RBF network			59.26%	51.90%
4	Worku (2013)	Learning vector quantization	Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work	Amharic	75.5% using TF and 71.96% using IDF	
5	Kamal et al. (2017)	Decision tree	Agriculture, sport, education, politics, health, and business	Affan Oromo	96.58%	

Table 2: Dataset

No.	Domains	Number of data
1	Award	892
2	Telecommunication	988
3	Micro-credit industry	743
4	Entertainment	1559

V. DATASET

To investigate the performance of proposed system the dataset collected from a different source that was written in Ge'ez font we call it Fidel [3][10]. Namely FBC (Fana Broadcasting Corporate), Addis Admass Newspaper, ENA (Ethiopia News Agency), Walta Information Centre, Reporter Newspaper, Horn Africa Website, Golgul Website Newspaper, and DW (Deutsche Welle). Dataset consists of 4,182 Amharic news item belongs to four un-touch domains.

VI. K-FOLD CROSS-VALIDATION

Due to its relatively low bias and variance k-fold cross-validation is used in this research to estimate the accuracy of NB and c4.5 decision tree classifier. In k-fold cross validation initially, the dataset partitioned approximately equal size (k). e.g. d1, d2, d3……., d_k the value of k depends on the dataset. For each k experiment k-1 fold reserved for the training set (model construction) and the remaining one-fold were used for testing.

VII. EXPERIMENT

Four experimentations were performed on 4,182 Amharic news using Decision Tree (j48) and Naïve Bayes and labelled in their pre-defined classes. To test the performance of both classifiers the researcher used by increasing the number of class and news with the different predefined number of classes and its corresponding news.

7.1 Naïve Bayes

Naïve Bayes (NB) is one of the simplest algorithms to implement using a small amount of training set in machine learning. It can be defined as independent of feature model (each feature word independent of another feature word). Furthermore, NB work based on the Bayes theorem with strong independence assumptions (Animut, 2012). The result of the classifier shown in Figure 1:

7.1.1 Experimentation on Three Classes

During this experimentation, three domains were used. Award (□□□□), Telecommunication (□□ □□□□□□), and Microcredit industry (□□□□□ □□□□□) with somehow different number of news. The accuracy of the test shown in Figure 1:

===Stratified cross-validation === === Summary ===

```
Test mode: 10-fold cross validation
Total instance 2623
Correctly Classified Instance          2482
                                       94.6245%
Incorrectly Classified Instance        141
                                       5.3755%
=== Confusion Matrix ===
a      b      c   <-- classified as
 841    41    10     a = Award
20     722   1      b = Micro-credit industry
47     22    919    c = Telecommunication
```

Figure 1: Confusion Matrix for Three Domains using Naïve Bayes

A confusion matrix constructed from rows and columns where the rows indicate the actual categories and the columns is predicate number of news classified to the corresponding classes. As the confusion matrix Figure shows, the first row 841 news represent a correctly classified number of instance in the category of Award. The rest 51 award news are incorrectly classified to microcredit industry and telecommunication category respectively. In the second row 21 news of microcredit industry incorrectly classified in award and telecommunication category. Including 722 instances of micro-credit industry news classified correctly. The last row indicates 69 news of telecommunication incorrectly classified in award and micro-credit industry category. The last row with the last column indicates 919 instances of telecommunication correctly classified to its class.

The highest incorrect instance is 47 (confusion) score between telecommunication and award which indicates there is a common word. In this experiment averagely 94.6245 % of the news correctly classified and the rest 5.3755 % are incorrectly classified.

===Stratified cross-validation === ==Summary ===

```
Test mode: 10   -fold cross validation
Total instance 4182
Correctly Classified Instance        3804
                                      90.8417%
Incorrectly Classified Instance      378
                                      9.1583%
===   Confusion Matrix ===
a      b       c      d     <       -- classified as
792    52      39     9     a = Award
139    1368    20     32    b = Entertainment
19     1       722    1     c = Micro industry
48     3       24     913   d= Telecommunication
```

Figure 2: Confusion Matrix for Four Domains using Naïve Bayes.

7.1.2 Experiment on four Classes

During this experimentation, all the dataset was used to construct and usage of the model based on the common evaluation 10-fold cross-validation. Entertainment (□□□□□), Award (□□□□□), Telecommunication (□□ □□□□□□), Microcredit industry (□□□□□□□□□□□) are the category of the news. The accuracy of the test shown in Figure 2:

As we can see from the above experiment, average result 90.8417 % of the documents were correctly classified, and 9.1583 % of the documents were incorrectly classified. The highest incorrect instance is 139 in the above confusion matrix which arises between Entertainment and Award. This indicates these two classes are more related.

7.2 C4.5 Decision Tree

Decision Tree classifier is also most widely used inductive learning methods. A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label [11]. The result of the classifier presents as follow.

7.2.1 Experimentation on Three Classes

Similarly, in this experiment, three classes were used to train and test the given document using typical j48 decision tree algorithm. The accuracy of the test shown in Figure 3:

===Stratified cross-validation === ==Summary ===

```
Test mode: 10-fold cross validation
Total instance 2623
Correctly Classified Instance        2498
                                      95.2345%
Incorrectly Classified Instance      125
                                      4.7655%
=== Confusion Matrix ===
a      b      c    <-- classified as
851    9      32         a = Award
15     709    19         b = Micro-credit industry
39     11     938        c = Telecommunication
```

Figure 3: Confusion Matrix for Three Domains using j48 Decision Tree

As we can see from the above experiment average result 95.2345 % of the documents correctly classified and 4.7655 % of the document incorrectly classified using j48 decision tree algorithm. The highest incorrect instance is 39 in the above confusion matrix which arises between Entertainment and Telecommunication and indicates there have a common word.

7.2.2 Experimentation on Four Classes

Here, the researcher used four categories.

```
===Stratified cross-validation === ==Summary ===
Test mode: 10-fold cross validation
Total instance 4182
Correctly Classified Instance        3804
                                     90.9613%
Incorrectly Classified Instance      378
                                     9.0387%
=== Confusion Matrix ===
a      b      c     d     <-- classified as
811    60     13    8     a = Award
69     1384   26    80    b = Entertainment
11     20     710   2     c = Micro industry
16     65     8     899   d= Telecommunication
```

Figure 4: Confusion Matrix for Four Domains using j48 Decision Tree

Table 3: Algorithms Performance Evaluation in Different Class

Algorithm	**Class**	**Average Accuracy Result**
Naïve Bayes	Four domains	90.8417 %
	Three domains	94.6245 %
C4.5 Decision tree	Four domains	90. 9613 %
	Three domains	95.2345 %

From the above experiment result, we can see that the algorithm classified averagely 90.9613 % of the documents correctly and 9.0387 % of the document incorrectly. From the confusion matrix, the highest confusion is 69 happened between Entertainment and Award. This shows that these classes are more related. All the above news classification experiments are implemented using WEKA package. Based on the above experiment the best average result is 95.2345 % achieved by the typical j48 Decision tree classifier on three domains and the lowest accuracy average score is 90.8417% on four domain using NB. As we can see from the experiment when the number of domain becomes larger the accuracy of the classifier will be slightly decreases. This is because the new added domain causes higher confusion during the experiment. Confusion arise between two classes and it indicate these two classes are more related.

VIII. CONCLUSION AND RECOMMENDATION

This research was performed on Automatic Amharic News Categorization using NB and typical j48 decision tree classifier. Accordingly, four experiments were performed in two different number of classes and algorithms which helps to achieve applicable machine learning technique for Amharic news text categorization. In the experimentation process the system used the dataset described earlier, the documents belong to 4 uncover categories. The best result achieved by j48 decision tree and NB is 95.2345 %, and 94.6245 % respectively on three categories and the least performance was 90.9613 %, and 90.8417 % respectively on four categories. To nutshell, from the experiment, this research concludes typical j48 decision tree classifier algorithm is more applicable to Amharic news categorization than the Naïve Bayes (NB). In addition,

when the number of domain becomes larger the accuracy of the classifier will be slightly decreased. In future, the researcher recommends to improve the performance of Amharic News classification, to explore another algorithms and applications of supervised and unsupervised machine learning approach. However, there are also some more additional tasks to be addressed and employed in the future work. These includes: To explore other algorithms and approach with less cost and better accuracy, extend this work by adding a new un-touch category in Ontology-based. Moreover, in future, the researchers aim to explore the approaches related to clustering and association in Data Mining and Natural Language Processing area.

REFERENCES

[1] Animut Belay Asres (2012). a semi- supervised approach for Amharic news classification, (Master's Thesis, Department of Information Science Addis Ababa University, Ethiopia).
[2] Bijal Dalwadi, Vishal Polara and Chintan Mahant (2015). A Review: Text Categorization for Indian Language, International Journal of Engineering Technology, Management and Applied Sciences, Volume 3 Issue 3, ISSN 2349-4476, pp. 95-97.
[3] Nidhi and Vishal (2012). Algorithm for Punjabi Text Classification, International Journal of Computer Applications (0975 – 8887) Volume 37– No.11, pp. 30-32.
[4] Alemu Kumilachew Tegegnie (2010). Hierarchical Amharic news text classification, (Master's Thesis, Department of Information Science Addis Ababa University, Ethiopia), IJASCSE volume 6 issue 02, pp. 1-19.
[5] Lars Asker, Atelach Alemu, Björn Gambäck and Magnus Sahlgren (2009). Applying Machine Learning to Amharic Text Classification, Springer Science+Business Media.
[6] Surafel Teklu Weldesellassie (2003). automatic categorization of Amharic news text: a machine learning approach, (Master's Thesis, Department of Information Science Addis Ababa University, Ethiopia).
[7] Lars Asker, Atelach Alemu, Bjo¨rn Gamba¨ck, Samuel Eyassu, and Lemma Nigussie (2009). Classifying Amharic webnews, Inf Retrieval, pp.416-432.
[8] Worku Kelemework (2013). Automatic Amharic text news classification: A neural networks approach, Ethiop. J. Sci. & Technol. 6(2) 127-137, pp. 127-135.
[9] Kamal Mohammed Jimalo, Ramesh Babu P and Yaregal Assabie (2017). Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach, International Journal of Computer Trends and Technology (IJCTT) – Volume 47 Number 1, pp. 5-11.
[10] Biruk Mengistu (2014). Restoration and Retrieval of Historical Amharic Document Images, (Master's Thesis, Department of Information Science Addis Ababa University, Ethiopia).
[11] Kabita Thaoroijam (2014). A Study on Document Classification using Machine Learning Techniques, IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 2, No 1, pp. 217-220.
[12] Nidhi and Vishal (2011). Recent Trends in Text Classification Techniques, International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, pp. 45-49.
[13] Xiaoli Guo, Huiyu Sun, Tiehua Zhou, Ling Wang, Zhaoyang Qu, and Jiannan Zang (2015). SAW Classification Algorithm for Chinese Text Classification, Sustainability, ISSN 2071-1050, pp. 2339-2350.
[14] Hanumanthappa and Narayana Swamy M (2016). Indian Language Text Documents Categorization and Keyword Extraction, International Science Press, pp. 37-45.
[15] Ubeeka Jain and Kavita Saini (2015). A Review on the Punjabi Text Classification using Natural Language Processing, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, pp. 3777-3779.
[16] Fabrizio Sebastiani (2002). Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, pp. 2-6.
[17] R. Balamurugan1 and S. Pushpa (2015). A review on various text mining techniques and algorithms, JNU convention centre Jawaharlal Nehru University New Delhi, pp. 838-844.
[18] Anuj Sharma and Shubhamoy Dey (2012). Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis, Special Issue of International Journal of Computer Applications (0975 – 8887), pp. 15-18.
[19] Rasha Elhassan and Mahmoud Ahmed (2015). Arabic Text Classification on Full Word, International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 5, pp. 114-119.
[20] Motaz K. Saad and Wesam Ashour (2010). Arabic Text Classification Using Decision Tree, Proceedings of the 12th international workshop on computer science and information technologies CSIT', pp. 77-78.
[21] Ashis Kumar and Rikta Sen (2014). Supervised Learning Methods for Bangla Web Document Categorization, International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 5, No. 5, pp. 93-103.
[22] Rajnish M. Rakholia and Jatinderkumar R. Saini (2017). Classification of Gujarati Documents using Naïve Bayes Classifier, Indian Journal of Science and Technology, Vol 10(5), ISSN (Print): 0974-6846, pp.1-8.
[23] Rehab Duwairi (2007). Arabic text categorization, The International Arab Journal of Information and Technology, vol, 4, No. 2, pp.125-129.
[24] Adel Hamdan, Omar Al-Momani and Tariq Alwada (2016). Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier, International Journal of Current Engineering and Technology, Vol.6, No.2, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161, pp.478-480.
[25] Shruti Bajaj and Vishal Goyal (2014). Text News Classification System using Naïve Bayes Classifier, International Journal of Engineering Sciences, Vol. 3, ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online), pp. 209-212.
[26] Mohammed J. Bawaneh, Mahmud S. Alkoffash and Adnan I. Al Rabea (2008). Arabic Text Classification using K-NN and Naive Bayes, Journal of Computer Science 4 (7): 600-605, pp. 600-604.
[27] Foram P. Shah and Vibha Patel (2016). A Review on Feature Selection and Feature Extraction for Text Classification, IEEE WiSPNET, pp. 2264-2267.
[28] Fouzi Harrag, Eyas El-Qawasmeh, and Pit Pichappan (2009). Improving Arabic Text categorization using Decision Tree, IEEE, pp.110-114.
[29] Dr.S.Kannan, and Vairaprakash Gurusamy (2014). Preprocessing Techniques for Text Mining, Conference: RTRICS.
[30] Fragkiskos D. Malliaros and Konstantinos Skianis (2015). Graph-Based Term Weighting for Text Categorization, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ISBN 978-1-4503-3854-7/15/08, pp. 1473-1475.

[31] C.O.S. Sorzano, J. Vargas, and A. Pascual-MontanoA (2014). survey of dimensionality reduction techniques, C/Darwin, 3. Campus Univ. Autónoma, 28049 Cantoblanco, Madrid, Spain, pp. 1-2.
[32] Marina Sokolova and Guy Lapalme (2017). Performance Measures in Classification of Human Communications, spring link, volume.4509, pp.4-5.