I J C T A, 8(5), 2015, pp. 1955-1962

Semantic Story Retrieval Based on Gender Centric and Language Style Identification

Emilet Annie A.*, P. Pabitha and Sivaprabha T.*****

Abstract: Story search helps the readers to search for the stories of their interest based on the query given by them. A general search compares the terms in the query and the terms occurring in the story, if they match then the story is retrieved which is prevalent with the existing retrieval system. The proposed system considers a Language centric and a gender centric search. The language centric search allows the readers to search the stories based on their proficiency in the language. The gender centric search allows the readers to choose either a Hero (male) or Heroin (female) centric stories. The system is purely designed to satisfy the users taste and preferences. A proper classification of story followed by indexing is done for retrieval. Experiments were conducted to perform gender and language style classification. Both the feature set and the classification algorithm influence the classification accuracy and error rate.

Index Terms: Story search, gender centric classification, language centric classification.

1. INTRODUCTION

Reading books is very important because it seem to improve our readability, imagination, stress reduction, knowledge and vocabulary expansion, moral values etc. To develop interest of reading books among the people, many books are available online. There is a necessity for an efficient story retrieval systems. Story search helps the readers to retrieve the story of their choice based on the query. The query is a formal statement of the user information needs is like a search string in the search engine [9][10].The query does not uniquely identify a single object in the collection but instead several objects that matches the query with different degrees of relevancy. Hence the objects must be classified based on their similarity. Stories can be classified by considering various aspects such as document type, genre, readability assessment, language identification, spam filtering, sentiment analysis etc [11][12][13]. Classification of stories can be of two types namely the content based classification and the request based classification. In the content based classification, a particular subject in the story determines the class to which the story is to be assigned [14][15]. A rule is followed that at least 20% of the content of the story should be about that class. In the automatic classification it could be the number of times the given words appear in the story. In the request oriented classification, a request from users influences how the story is to be classified. The classifier should be able to find the descriptors to which the entity belongs and to find all the queries for which the entity in hand is relevant [16]. The request oriented classification is used in the semantic story search. The system gets the users request and retrieves all the stories that are relevant to the request. The classification helps in grouping the stories based on a particular criterion. When a request for that criterion is made, all the relevant stories for the requested criterion is retrieved. To provide the readers with a good reading environment, it is necessary that we need to get the user taste and preference and then give them the stories that best suites their needs.

* Madras Institute of Technology, Anna University, Chrompet, Chennai-600044, *Email: emiletannie92@gmail.com*

** Department of Computer Technology, Faculty of Information and Communication Engineering, Madras Institute of Technology, Anna University, Chrompet, Chennai-600044, *Email: pabithap@gmail.com*

*** Madras Institute of Technology, Anna University, Chrompet, Chennai-600044, *Email: sivaprabha26@gmail.com*

The classification of the stories based on the gender identification helps to classify the story as male centric or female centric.(E.g. when considering the stories for kids, Grimm's fairy tales like Cinderella, Rapunzel are classified as female centric because more importance is given for the female character. Tom Sawyer, huckleberry finn can be classified as a male centric story because more importance is given for the male character). Behavioural analysis states that the female centric stories seems to attract girls and women. Male centric stories seem to attract boys and men. The gender centric classification is designed to suite individual needs and preferences. Language style identification is used to classify the stories based on the readability of the user. This is based on the proficiency of the user with the English language. A three point classification of the readers is done namely beginner, intermediary and proficient. The classification helps the readers to choose the stories based on their fluency with the language and this will provide the readers a better interpretation of the story.

The paper is organized as follows. The second section discusses about the related work on document search and classification. The third section discusses about the design of the system. Then the fourth section discusses about the experimental evaluation of the classification system and finally concludes with a short discussion on future work.

2. RELATED WORK

The conventional method for the document classification includes the following, Improving the classification, reducing the dimensionality of the features, proper indexing for the faster retrieval, building a good classifier models, improving the discrimination between the classes.

XU Jiao and LI Lian (2015**)** stated that the document classification is the task of sorting a set of documents into categories. Improving the accuracy of the text categorization and reducing the dimension of the feature space is very important. A two stage feature selection method called the category correlation degree (CCD) is used in feature selection and latent semantic indexing (LSI) is used to discover the correlative relation between features and to reduce the feature space dimension. In CCD, each feature can be ranked depending on importance of the classification. LSI is used to construct a new semantic space. The main advantage is that CCD reduces the dimension of features and LSI reduces the computational complexity. When the number of LSI dimensions for CCD goes beyond 100, the accuracy and the macro-F1 decreases.

Xiao and ZincirHeywood (2004) stated that a data processing methods and dimensionality reduction can be used together. A evaluation on the three dimensionality reduction techniques for a high dimensional document collection is used for efficient feature reduction. Three dimensionality reduction methods namely the LSI, random mapping and the combination of the two algorithms is explained. On comparison with the F1 measure the third method of combination of the random mapping and LSI seem to perform better than just the LSI algorithm[7].

Gonen (2013) reported that The dimensionality reduction in general used as a pre-processing step can be coupled with the supervised learning steps to improve the predictive performance. Bayesian supervised dimensionality reduction method is used. Three benchmark datasets were considered and seven dimensionality reduction algorithms were used. The classification and the retrieval performance were better than the traditional techniques.

Lan et al.(2009) stated that when using the vector space model(VSM), a document can be transformed into vector in the term space which can be classified. To improve the text classification, we can assign different weights to the terms using the term weighting method. A supervised and unsupervised term weighting methods is used along with the SVM and KNN to do the classification. In a supervised term weighting method called the tfrf is used and it seems to perform better when compared to the tf-idf weighting schemes whether it is by using the linear or the non-linear svm classification algorithm.

Kim et al.(2006) stated that the traditional Naïve Bayesian algorithm has a poor performance for the text classification. The main reason for this is due to the poor parameter estimation. Two heuristic approaches are used namely per-document text normalization and the feature weighting methods are used to overcome the problem of poor parameter estimation. So this modified naïve Bayesian performs better than the SVM classifier.

Dennis and Shreyes (2009) proved that natural language processing can be used to classify the documents efficiently. A simple hypothesis is used where the documents in different categories can distinguish themselves by using the features obtained by the natural language processing. The features are the word structure, word frequency and the natural language structure. The classified document is randomly spit into training and the testing group. Experiments were conducted using the feature sets for the Bayes Classification, Maximum entropy classification and for examining the sentence structure and their differences using probabilistic grammar parsers. While trying to use the content based classification features did not give a better result because the training set was not taken properly which causes the class imbalance problem.

Yanguang et al.(2015) reported a comparison between the four text classifiers. These classifiers were tested on movie reviews whether they were able to classify the movie reviews as either positive or negative. This involves the analysis of the features in the reviews and sometimes this could result in curse of dimensionality which means the analysis of both the useful and useless features and hence it necessary to carefully select the features for the correct classification.

Lianjing et al.(2015) stated that the text classification is the base for text mining. Naive Bayes is an effective method for text classification. This paper improves the accuracy of Naive Bayes classification using improved information gain is one of the methods of feature extraction. This can be done by reducing the impact of low-frequency word. A corpus of NLTK is used. The accuracy of the classification was improved significantly.

3. PROPOSED SYSTEM

Document retrieval is very important because maintaining and manually retrieving large volumes of document is very tedious. The proposed work involves the efficient classification of the story based on the gender and language style identification for search and retrieval

3.1. Document Pre-processing

Classification of story begins with preprocessing which is taken as the training document. Preprocessing involves tokenizing, stemming, stopwords removal, case folding or capitalization and parts of speech tagging which can be done efficiently using the natural language processing because it is an efficient technique for the computers to understand our human language. Tokenization includes the word and sentence tokenizers. Stemming is used to remove the suffixes. Stopwords are the unwanted words which does not have any meaning and not effective for classification (e.g. for, the, an etc). Case folding converts the entire story to a standard lower or upper case for easy processing. Parts of speech tagging identifies the parts of speech (nouns, verbs, adjectives. etc) in the story that can be very useful features for classification.

3.2. Feature selection & Extraction

Feature selection and extraction helps in finding the most informative features that provide the distinction between the stories. The defined features are then extracted from the story, based on which the classification is done. The feature set selected determines the accuracy of the classification and hence the selection of the proper features is very important. Feature selection means selecting the existing features without any transformation but whereas the feature extraction if used to transform the features to a lower dimensional space.

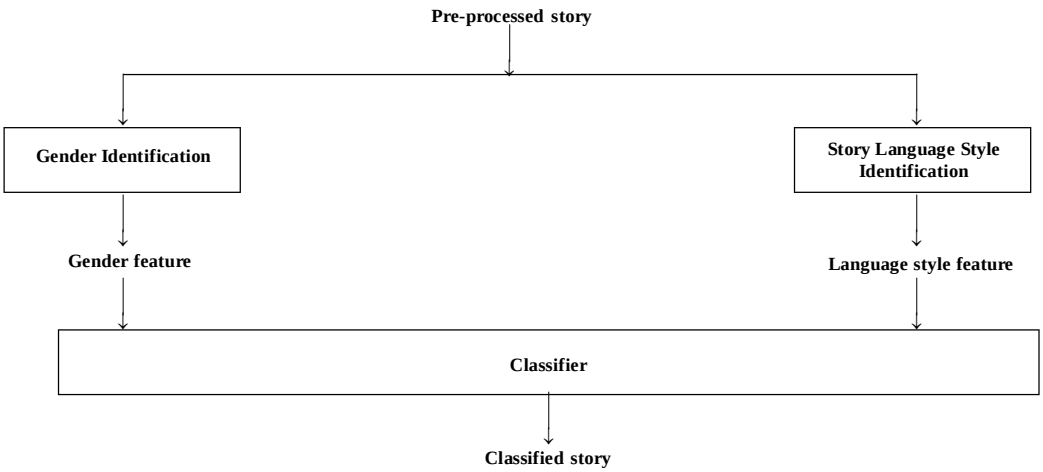

Figure1: Feature selection & extraction

A figure 1 show that after the document is pre-processed, features need to extracted and selected. So we get the gender identification feature and the language style features. The Gender features taken for classification are given below.

Feature set for Gender Identification of names = {(last1l), (first1l, last1l, count), (first1l, last1l, first2l), (last1l, last2l), (first1l, first2l), (first1l), (last3l), (last2l), (last1l, last2l, last3l), (first2l), (first3l)}

Where, l = letter in the name. Example: On considering the name 'Jennifer',

The Gender feature set = {('r'), ('J', 'r', count(j) = 1, count(e) = 2, count(n) = 2, count(i) = 1, count(f) = 1, count(r) = 1), ('J', 'r', 'er'), ('r', 'er'), ('J', 'Je'), ('J'), ('fer'), ('er'), ('r', 'er', 'fer'), ('Je'), ('Jen')}

The feature set described above was used for the test data on the gender names and it was able to classify the names as male or female.

Table 1
Informative feature for gender identification

Gender feature set No	*Male: Female ratio*	*Description*
1,2	{last letter = ′a′ female : male = 30.8 : 1.0, ′k′ male : female = 27.0 : 1.0, ′f′ male : female = 24.4 : 1.0, ′v′ male : female = 17.6 : 1.0, ′p′ male : female = 9.9 : 1.0}	Considering last letter
3,4	{lw = ′na′ female : male = 89.8 : 1.0, lw = ′la′ female : male = 65.7 : 1.0, lw = ′ra′ female : male = 54.5 : 1.0, lw = ′ia′ female : male = 32.5 : 1.0, lw = ′ld′ male : female = 31.7 : 1.0}	lw considers last two letters
5	{firstwo = ′hu′ male : female =17.3 : 1.0, firstwo = ′wh′ male : female = 8.4 : 1.0, firstwo = ′fo′ male : female =7.7 : 1.0, firstwo = ′tu′ male : female = 7.2 : 1.0, firstwo = ′wa′ male : female = 7.0 : 1.0}	Firstwo considers first two letters
6	{prefix1 = ′w′ male : female = 4.1 : 1.0, prefix1 = ′q′ male : female = 2.4 : 1.0, prefix1 = ′x′ male : female = 2.4 : 1.0, prefix1 = ′u′ male : female = 2.3 : 1.0, prefix1 = ′k′ female : male = 2.3 : 1.0}	Prefix1 considers only the first letter
7	{suffix1 = ′nne′ female : male = 29.4 : 1.0, suffix1 = ′ana′ female : male =23.8 : 1.0, suffix1 = ′tta′ female : male = 21.7 : 1.0, suffix1 = ′lyn′ female : male = 18.7 : 1.0}	Last three letters
8,9	{suffix1 = ′na′ female : male = 89.8 : 1.0, suffix1 = ′la′ female : male = 65.7 : 1.0, suffix1 = ′ra′ female : male = 54.5 : 1.0, suffix1 = ′ia′ female : male = 32.5 : 1.0, suffix1 = ′ld′ male : female = 31.7 : 1.0}	Last two letters
10	{ft = ′hu′ male : female = 17.3 : 1.0, ft = ′wh′ male : female = 8.4 : 1.0, ft = ′fo′ male : female = 7.7 : 1.0, ft = ′tu′ male : female = 7.2 : 1.0, ft = ′wa′ male : female = 7.0 : 1.0}	First two letters
11	{ft = ′gar′ male : female = 13.4 : 1.0, ft = ′tha′ male : female = 13.0 : 1.0, ft = ′cat′ female : male = 11.3 : 1.0, ft = ′ros′ female : male = 9.6 : 1.0, ft = ′dor′ female : male = 8.8 : 1.0}	First three letters

Feature set for readability identification= These feature set helps in identifying the readability of the words that can be classified into three point scale as beginner, knowledgeable and highly proficient.

Readability = {(length), (a, e, i, o, u), (count of vowels, length), (prefix letter, length), (first 1l, first 2l), (last 1l), (count of vowels, length, first 1l)}. Example: on considering the word 'abrogate'

The Readability feature set = {(length = 8), (a = 2, e = 1, o = 1), (count = 4, length = 8), ('a', 8), ('a', 'ab'), ('e'), (4, 8, 'a')}

The proposed system should be able to address the problem of dimensionality reduction. An term frequency technique is used which is used to assign weights to the terms that is considered important for classification.

Table 2
Informative feature for Readability identification

ReadabilityFeature No	*Informative feature (Begin=Beginner, Know=Knowledgeable, high= highly proficient}*	*Description*
1,4	{length = 3 Begin : high = 45.5 : 1.0, length = 10 know: Begin = 34.2 : 1.0, length = 9 high : Begin =19.7 : 1.0, length = 4 Begin : high = 17.4 : 1.0, length = 7 know : Begin = 11.8 : 1.0}	Length of the word, prefix letter
2	{count(e) = 3 know: Begin = 17.1 : 1.0, count(a) = 2 know : Begin = 5.6 : 1.0, count(o) = 2 high: know = 3.8 : 1.0, count(e) = 2 know : Begin = 3.7 : 1.0, count(e) = 0 Begin : know = 2.8 : 1.0}	Count of the vowels
3	{count(u) = (0, 3) Begin: high = 40.1 : 1.0, count(i) = (0, 3) Begin : high = 39.2 : 1.0 count(a) = (0, 3) Begin: high = 35.1 : 1.0, count(o) = (0, 3) Begin : high = 33.4 : 1.0 count(e) = (0, 4) Begin : high = 27.2 : 1.0}	Count of vowels & length
5	{prefix1 = ′e′ know: Begin = 6.2 : 1.0, firstwo = ′re′ know: Begin = 5.3 : 1.0, firstwo = ′ap′ know: Begin = 4.8 : 1.0, firstwo = ′un′ know : Begin = 4.8 : 1.0, prefix1 = ′v′ high : know = 4.8 : 1.0}	First letter, first 2 letter
6	{suffix1 = ′s′ high : know = 18.1 : 1.0, suffix1 = ′d′ high: know = 8.5 : 1.0, suffix1 = ′y′ high : know = 6.1 : 1.0, suffix1 = ′p′ Begin: high = 5.2 : 1.0, suffix1 = ′e′ high: know = 3.6 : 1.0}	Last letter
7	{count(u) = (0, 7, ′d′) know: high = 3.0 : 1.0, count(i) = (1, 7, ′d′) know : high = 3.0 : 1.0 count(o) = (0, 8, ′s′) know: high = 3.0 : 1.0, count(o) = (0, 7, ′d′) know: high = 3.0 : 1.0count(o) = (1, 10, ′p′) high : know = 2.9: 1.0	count of vowels, length, first letter

3.3. Document Classification

Document/story Classification classifies the story based on gender and language style identification. The gender and the language style features are given as input to the classifiers for classification. A classifier is used to perform the classification based on the classes defined during training. The training data is given as input to the classification algorithm with the predefined class labels. So by defining the feature and the classification algorithm, a trained classifier model is created which is able to differentiate between the classes based on the features. Hence when a test document without any class label is given as input to the classifier model. Now the classifier classifies the test data to a particular class. Hence a classified document is got as an output. The gender based identification of the names involves the classification of the names into two classes which includes male and female. In language style identification, the classifier classifies the story into three classes as beginner, intermediary and proficient.

4. EXPERIMENTAL EVALUATION

4.1. Experimental setup

NLTK (Natural Language Processing Toolkit) has an inbuilt corpus that has the names dataset. This dataset had the male and the female dataset. Total of 7944 names were considered of which there were 2943 male

names and 5001 female names. About eleven features sets defined in table1 is considered and was applied to the machine learning algorithm like the naïve Bayes, decision tree and the maximum entropy algorithms. The training dataset contains 7444 names and test dataset contains 500 names. NLTK is an NLP package that can be used in python.

In this section, The classification accuracy and the error rate for the gender identification of names and the language style identification is discussed based on which a comparative study on three algorithms namely the naïve Bayesian, decision tree and maximum entropy is done.

4.1.1. Impact of the Gender feature sets on Classification Accuracy

In this section, we will discuss on the impact of the feature set on the accuracy of the classification in figure 2a &2b. The classification accuracy seems to vary with the British names and the Indian names. Classification accuracy was high with maximum entropy algorithm when considering the British names but whereas with the Indian names the naïve Bayesian was better when compared with the other algorithms.

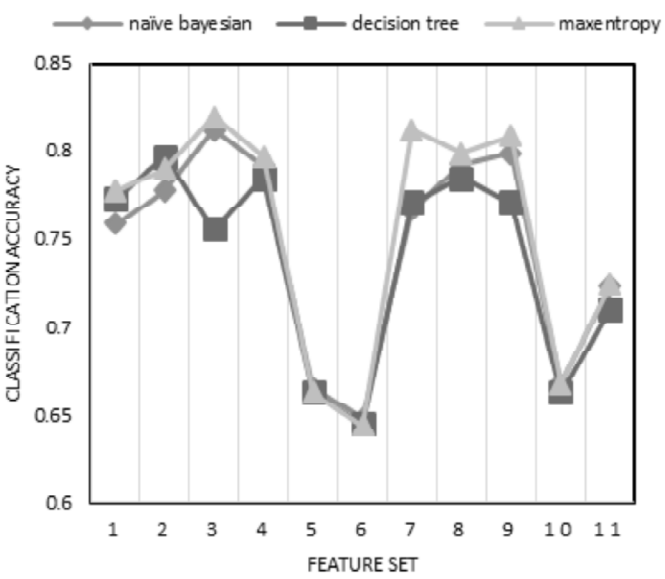

Figure 2a: Classification accuracy variation with the feature set for British names

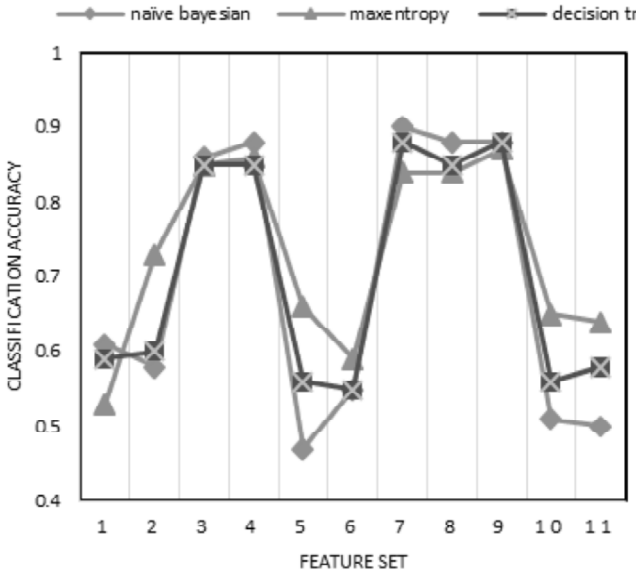

Figure 2b: Classification accuracy variation with the feature set for Indian names

The classification accuracy of foreign names using the maximum entropy and Indian names using the Naïve Bayesian. Only the best five feature sets are shown in Table 2. Feature set 7 seem to be the best for both the foreign names and Indian names. The feature sets can be referred from table 1.

Table 2
Classification accuracy of Gender identification

Gender Feature set	*Accuracy (Foreign names)*	*Accuracy (Indian names)*
3	82%	85%
4	79.7%	88%
7	81.2%	90%
8	79.9%	88%
9	80.9%	88%

4.1.2. Impact of the Language style feature sets on Classification Accuracy

In this section, we will discuss on the impact of the feature set on the accuracy of the classification in figure 3. The classification accuracy seems to vary with the language style feature sets. Classification accuracy

Table 3
Classification accuracy of Language style identification

Language style feature set	*Classification Accuracy*
3	99.6%
4	87%
7	85.6%

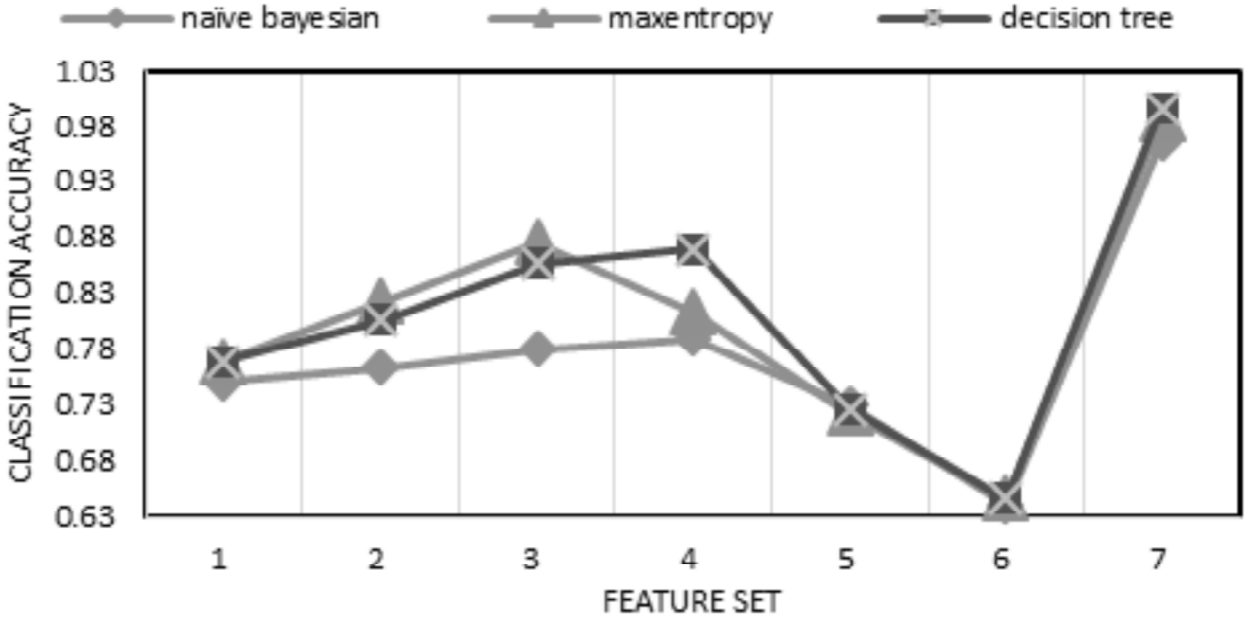

Figure 3: Classification accuracy variation with the feature set for Language style identification

seems to be high for the decision tree for features 4 & 7 and for the features 2 & 3 maximum entropy seems to be better. Naïve Bayesian does not perform well for the feature set 1,2,3,4. Totally 731 words were taken out of which 387 words are labelled as beginner, 147 words are labelled as intermediary, 244 wordsare labelled as proficient. About 431 words are taken for training the classifier model and the remaining 300 words were taken for testing.

Table 3 shows the classification accuracy of language style identification using decision tree algorithm. The best three feature set is shown along with the accuracy.

5. CONCLUSION

Gender based classification of the names was able to classify the names according to the gender as Male and Female using the naïve Bayes, maximum entropy and decision tree algorithms. Experimental results showed that maximum entropy seem to perform better than the naïve Bayesian and the decision tree algorithm when considering the classification accuracy and the error rates for the Gender based classification of the foreign names. But however for the Indian names the naïve Bayesian performs better than the maxentropy and the decision tree in prediction. Language style identification was able to classify the English words according to the readability such as whether it is for the beginners, knowledgeable or for the highly proficient readers where the decision tree seem to outperform the maximum entropy algorithm. So each machine learning algorithm perform better than the others depending on the application for which they are used. The future work involves considering various different features so as to further improve the accuracy of the classification and also to experiment with other performance metrices like the precision, recall and the F1 measure. Based on the results obtained can be used in classifying the document based on the gender content and the readability and also classification of the documents based on the semantic meaning can be done.

Reference

[1] Dennis Ramdass, Shreyes Seshasai, *"Document Classification for Newspaper Articles"*, 2009, Final Project Springer 2009.

[2] Gonen M., *"Bayesian Supervised Dimensionality Reduction"*, IEEE Transactions on Cybernetics, Vol. 43, No. 6, pp. 2179 – 2189, 2013

[3] Kim B., Han S., Rim C., and Myaeng H., *"Some effective techniques for naive bayes text classification,"* IEEE Transactions on Knowledge and Data Engineering, pp. 1457-1466, 2006.

[4] Lan, Man, et al, *"Supervised and traditional term weighting methods for automatic text categorization,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 721-735, 2009.

[5] Lianjing Jin, Wei Gong, Wenlong Fu and Hongbin Wu, *"A Text Classifier of English Movie Reviews Based on Information Gain"*, 2nd International Conference on Computational Science and Intelligence (ACIT-CSI), pp. 454 – 457, 2015.

[6] XU Jiao and LI Lian,*"A Two-Stage Feature Selection Method for Text Categorization by Using Category Correlation Degree and Latent Semantic Indexing"*, Journal of Shanghai Jiaotong University (Science), Vol. 20, No. 1, pp. 44-50, 2015.

[7] Xiao Luo and Zincir-HeywoodA.N., *"Evaluation of three dimensionality reduction techniques for document classification"*, Canadian Conference on Electrical and Computer Engineering, Vol. 1, pp. 181 – 184, 2004.

[8] Yaguang Wang, Wenlong Fu, Aina Su and Yuqing Ding, *"Comparison of Four Text Classifiers on Movie Reviews"*, 2nd International Conference on Computational Science and Intelligence (ACIT-CSI), pp. 495 – 498, 2015.

[9] SandersonM and CroftW.B,*"TheHistoryof Information RetrievalResearch"*, Proceedings of the IEEE, Vol: 100, pp: 1444 – 1451, 2012.

[10] Kushchu I, *"Web-based evolutionary and adaptive information retrieval"*, IEEE Transactions on Evolutionary Computation, Vol. 9, No. 2, pp. 117 – 125, 2005.

[11] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee, *"A similarity measure for text classification and clustering"* Knowledge and Data Engineering, IEEE Transactions, pp. 1575-1590, 2014.

[12] Pouilloux F, *"Extracting Named Entities at Web Scale for Competitive Intelligence"*, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1, pp. 501–501, 2011.

[13] Miltsakaki E, & Troutt A, *"Real-time web text classification and analysis of reading difficulty"*. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp. 89-97, 2008.

[14] Ebook: *"Machine Learning in Document Analysis and Recognition"* byMarinai, Simone, Fujisawa, Hiromichi .

[15] Classification: http://www.avepoint.com/community/avepoint-blog/challenges-data-classification/

[16] Taha K,(2013),*"Determining the Semantic Similarities Among Gene Ontology Terms"*, IEEE Journal of Biomedical and Health Informatics, Vol. 17, No. 3, pp. 512–525