RESEARCH ARTICLE OPEN ACCESS

Semantic Similarity Based Classification Of Narratives

S. Gopi [1], A. Berno Raj [2], M. Abinav [3], P.Gokul Sarathy [4], D.P. Bharath [5]

Asst.Professor [1] , UG Student [2], [3], [4] & [5]

Panimalar Engineering College, Chennai

Tamil Nadu – India

ABSTRACT

Semantic similarity between the Narratives is found by constructing a semantic network that defines the semantic relatedness which can be introduced to classify the document. The story retrieval is mainly due to the identification of the protagonist which can be classified as male based or female based narratives. The comprehensibility can be found based on the algorithm with the improvised accuracy

***Keywords*:-** Narratives classification, Protagonist identification, improvised comprehensibility

I. INTRODUCTION

Extensive reading is done for recreation, to reduce stress, to think out of the box while intensive reading is done for knowledge acquisition, vocabulary development and to understand the moral values. At this point there comes a necessity for automated narrative retrieval system. Search engine uses keyword to fetch the narratives. The keyword doesn't identify single article but multiple articles with less degree of relevancy. Hence there is a need for semantics to increase the relevancy. Classification of emails, natural language processing uses semantics [xi] [xii].Request based classification points out the narratives based on the users query. The crawler retrieves the narratives while the result doesn't slip from relevancy.

The classification of narratives is based on the Protagonist identification and clustering the names based on gender. Tom Clancy's novels are based on military subjects and warfare which are subjected to attract male readers. On the other hand the princess diaries novel is liked by female readers. Comprehensibility of the content is estimated and retrieved based on the users proficiency.

II. RELATED WORK

XU jiao and LI Lian(2015):

Dimensionality reduction of the feature is the key aspect which is done by category correlation and the latent semantic index. Here CCD reduces the features dimension while the LSI reduces the formulation complexity. When the CCD>100 then the accuracy decreases.

Gonen(2013):

In supervised learning, the dimensionality reduction is the pre-processing mechanism for performance enhancement.

Dimensionality reduction of feature is set out by using Bayes' algorithm. Here the performance is superior than the conventional methods.

Lan et al(2009):

Vector space model can be used to transform a document into vector by assigning weights to the text for improvised classification. Here SVM and KNN as a part of supervised and unsupervised learning scheme does the term frequency and inverse document frequency weighing pattern.

Kim *et al*(2006):

Text normalization and feature weighing pattern are introduced to resolve the poor parameter estimation in the conventional Naïve Bayesian paradigm. So the improvised Naïve Bayesian has some added advantages over the SVM.

III . EXISTING SYSTEM

Text categorization finds its application in the automated indexing, dissemination of information to consumers, filing patents into patent directories, spam filtering, identification of document genre, essay grading. So to improve the accuracy of the text categorization and for reducing the dimension of the feature space is very important.

IV. PROPOSED SYSTEM

Manual classification and retrieval of large volumes of document is monotonous and tedious. Hence the Proposed system actuates the automated classification of narratives.

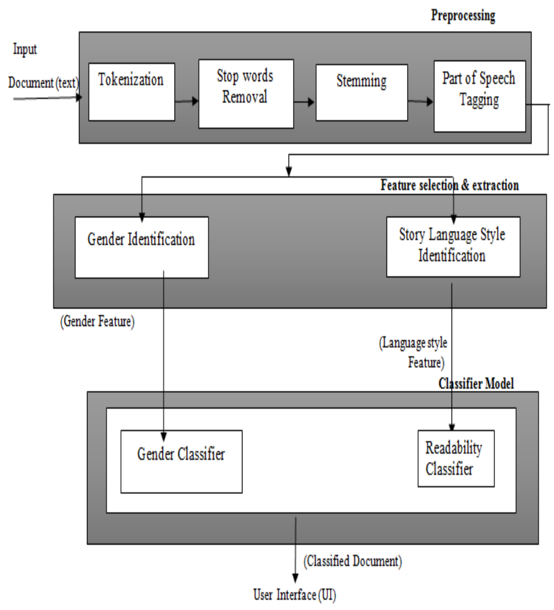

Fig.3.1 **Architecture**

Document pre-processing

Classification of narratives starts with the pre-processing. It involves tokenization, removal of stop words, stemming and POS tagging. Word and sentence tokenizer performs the process of breaking the stream of text into tokens. Followed by Stemming for the removal of suffix. Stop words doesn't possess any significant meaning in the documents and so they are eliminated. POS tagging finds the parts of speech(such as noun, verb) in the document to maintain the semantics of the narration. {**RB** = Adverb, **VBD** = Verb, past tense , **NNP** = Noun, proper, **JJ** = adjective, **CD** = Numerical, cardinal, **NNS** = Noun, common, plural , **VBN** = Verb, past participle , **IN** = Preposition or conjunction, **NN** = Noun, common, singular or mass , **VBG** = Verb, present participle or gerund, **CC** = Conjunction, coordinating }

Attribute selection and extraction

In this phase, the distinction between the narrations is found. The attribute selected determines the accuracy of the classification. Hence the attribute selection is predominant for the classification. The attribute extraction is used to transform the attribute into a low dimension space.

Narrative Classification

Narrative classifier performs the protagonist identification and then classifies the narrative based on the gender. It also identifies the comprehensibility of the language. During the training, the classifier understands the classes for classification and then stores the corresponding values to the male and female classes. Similarly the comprehensibility is done where the classifier classifies the narratives into 3 classes namely beginner, intermediate and proficient.

V. EXPERIMENTAL EVALUATION

A. Experimental setup

The dataset resides in corpus of the Natural language processing toolkit which holds the male and female names. The dataset had 7933 names in which 2923 were male names while 5010 were female names. Machine learning algorithms like Naïve Bayes, Maximum entropy were used.

Performance and Classification accuracy on Gender feature. Maximum entropy algorithm is found to have higher accuracy for British names while the Naïve Bayes algorithm has higher accuracy with the Indian

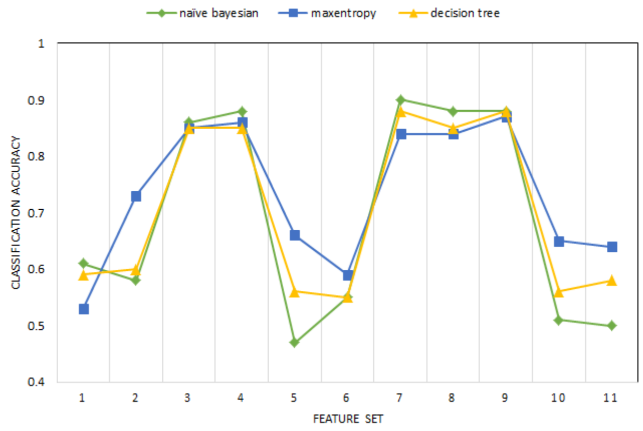

Fig.4.1 Classification accuracy for feature set of indian names

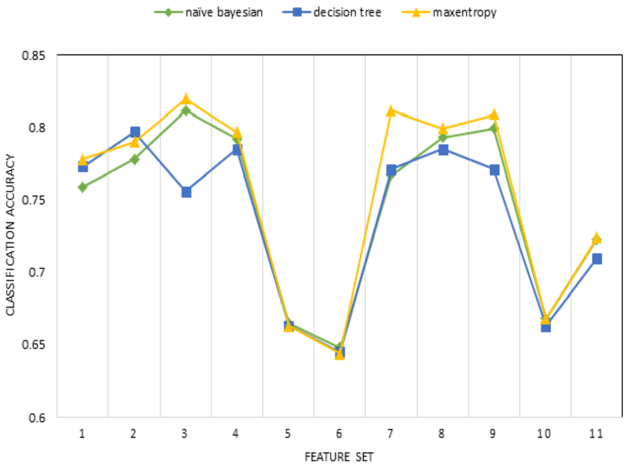

Fig.4.2 Classification accuracy for feature set of foreign names

Table 1

Classification Accuracy of Gender Identification

Gender FeatureSet	Accuracy (Foreign Names)	Accuracy (Indian Names)
3	82%	85%
4	79.7%	88%
7	81.2%	90%
8	79.9%	88%

Performance and Classification accuracy on Comprehensibility

Decision tree is found to have higher accuracy for the feature 4 and 7 while Maximum entropy has upper hand for feature 2 and 3.Naive Bayesian has lesser accuracy for 1,2,3,4.The dataset had 788 words, in which 384 words were categorized as beginner while 155 words were categorized as intermediate and 249 words as proficient.

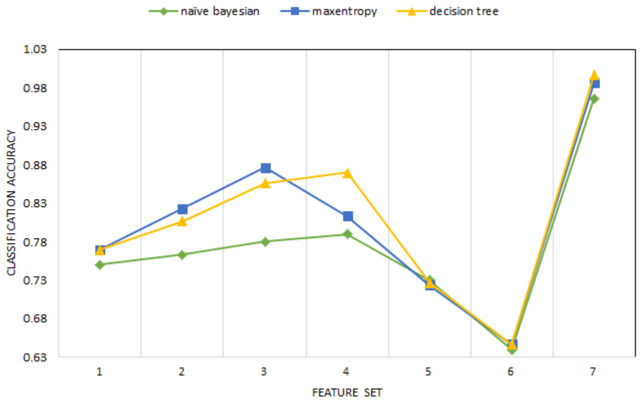

Fig.4.3 Classification accuracy for Comprehensibility

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

Table 2

Classification accuracy for comprehensibility

VI. CONCLUSION

Maximum Entropy has higher performance than Naïve and decision tree algorithm when foreign name was taken under consideration while Naïve has higher accuracy over max entropy and decision tree. Comprehensibility of the language classifies the English words into beginner, intermediate and proficient. The future work involves combining the system with the personal assistant for enhanced personalization.

REFERENCES

[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.