Classification of Children Stories in Hindi Using Keywords and POS Density

Harikrishna D M, K. Sreenivasa Rao
School of Information Technology
Indian Institute of Technology
Kharagpur, India
Email: harikrishna.dm@sit.iitkgp.ernet.in, ksrao@iitkgp.ac.in

***Abstract*—The main objective of this work is to classify Hindi stories into three genres: Fable, Folk-tale and Legend. In this paper, we are proposing a framework for story classification using keyword and Part-of-speech (POS) based features. Keyword based features like Term Frequency (TF) and Term Frequency Inverse Document Frequency (TFIDF) are used. Effect of POS tags like Noun, Pronoun, Adjective etc., are analyzed for different story genres. Classification performance is analyzed using different combinations of features with three classifiers; Naive Bayes (NB), k-Nearest Neighbour (KNN) and Support Vector Machine (SVM). From the experimental studies, it is observed that combining linguistic and keyword based features do not improve significantly the classifier performance. Among the classifiers, SVM models outperformed the other models.**

***K*eywords —** Hindi Story Classification; Text-to-Speech; Document Classification; Part-of-Speech; Vector Space Model; Naive Bayes; KNN; SVM

I. Introduction

A story can be defined as series of events and character actions. Story narration style varies across different genres because of the specific emotional content and characters present in it. Emotions such as anger, sad, happy, fear, surprise and disgust are observed in stories. To attract the attention of listeners, a story teller will interpret these emotions and narrate a story by modifying prosody i.e., pitch, duration and intensity. For a Text-to-Speech (TTS) system to be more natural and realistic, it is important that the text analysis module in TTS system should predict the story genre and emotions present in the text. In this work, we are addressing the text classification problem and the results will be used in text processing module of Hindi story TTS system.

In this work, we have considered three story genres namely, fable, folk-tale and legend. Fable is a short tale involving animals as essential characters. Folk-tale is a popular story which is passed on in spoken form from one generation to the next. Legend is a semi-true story carrying important meaning or symbolism for the culture in which it originates. It is based on historic factors of a particular geographic region.

In [1], a multi-step system for children story analysis was proposed. Story analysis tasks like identification of characters, personality attributes of character like age and gender were carried out. From the literature, it is observed that there is no existing works on story classification in Hindi and features related to POS have not been explored for story classification in Hindi. In this paper, we propose new feature based on Part-of-speech (POS). The motivation for considering the POS information for story classification is supported by the observations like more named entities in stories and importance of POS tags like nouns, adjectives, quantifiers for distinguishing between story genres. The rest of this paper is organized as follows. Section II gives an overview of the related work. Overview of the proposed framework is explained in Section III. Section IV presents the detailed experimental setup. Results of the experiments are discussed in Section V. Conclusions are drawn in Section VI.

II. Related Work

Text classification was carried out for different domains using different approaches across languages. Most widely used approaches for text classification are WordNet and Machine Learning approach. In [2], Sanskrit documents were classified in linear-time using lexical chain which links significant words. In [3], a lexicon-pooled Naive Bayes approach for classifying Nepalese news stories was proposed. Domain specific lexicon were created for each class and incorporated into Naive Bayes classifier, which substantially improved classifier performance. They also contributed resources in the form of Nepali news stories corpus and domain-specific lexicon for Nepali news stories. In [4], ontology and hybrid based approach for classification of Punjabi text documents was proposed. They developed sports specific ontology for Punjabi and conducted in-depth analysis of Punjabi corpus for the preparation of gazetteer lists such as middle names, last names, abbreviations etc., for Named Entity Recognition task. In [5], Marathi articles were classified using different classifiers and compared their accuracy and classification time parameters. Rule based stemmer and Marathi word dictionary were built to reduce the dimensionality of feature vectors. In [6], Kannada web pages were classified using various pre-processing agents. Pre-processing steps like language identification, sentence boundary detection, stemming and stopword removal are applied on the webpage content before classification. In [7], manually collected Kannada sentences from Kannada Wikipedia were classified. Stop words and restrictions based on word occurrence were used for dimensionality reduction. In [8], Tamil documents were classified using Artificial Neural Network (ANN) and Vector Space Model (VSM). Their experiments concluded that ANN is better for more representative collection and captured the non-linear relationships between the input document vectors and the document categories than that of VSM. In [9], Advanced Back Propagation Algorithm (ABPA) was used for the

classification of Tamil documents. In [10], Telugu news articles were classified into four categories: Politics, Sports, Business and Cinema using NB classifier. In [11], language independent, corpus-based machine learning techniques were used for text categorization in ten major Indian Languages. Challenges in text categorization of Indian languages like a large amount of word-forms, the morphological richness of the languages and feature-space were explored in their work. But, there is no existing work on story classification for Indian languages. This led to the motivation for the present study.

III. Hindi Story Classification

Fig. 1 shows the process for Hindi story classification. 300 short stories are collected from story books and blogs. Story corpus is cleaned and stopwords are removed. Using Hindi Shallow Parser, Lemmatization and POS tagging are carried out for entire stories. Feature vectors are computed using a combination of different weighing schemes for keywords and density of POS tags. Classifiers are used to predict the output class labels. Story TTS system uses predicted class label and story text to modify the prosody to convey the story-telling style.

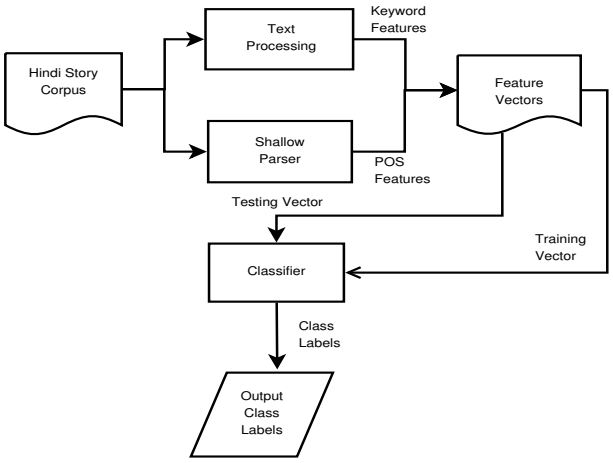

Fig. 1. Flow diagram of Hindi story classification

A. Database Preparation

Hindi story corpus consists of 300 short stories collected from Blogs, Panchatantra and Akbar-Birbal books. No standard dataset is available for Hindi story classification. Details of the Hindi story corpus are presented in Table I. Categories of stories and their definitions are explained in Section I.

TABLE I. Statistical Information of Hindi Story Corpus

Category	Total Stories	Words	Unique Words
Fable	100	50344	3313
Folk-tale	100	46900	3710
Legendary	100	35991	2963

B. Text Processing and POS Tagging

Corpus is cleaned as a part of pre-processing. Multiple white spaces are stripped and punctuation marks, special symbols and numbers are removed. Furthermore, Hindi Shallow Parser[1] developed by IIIT Hyderabad is used for lemmatization or stemming, to convert each word into its root word (base form). Lemmatization is an important task as it decreases the feature vector dimension which leads to a better representation of the document. No standard stopword list for Hindi is available. A list of 164 stopwords is prepared and used for this work.

C. Feature Extraction

Hindi story classification is treated as a supervised machine-learning problem, where stories are projected into Vector Space Model (VSM) which uses words as features. Motivated by the observations in section I, linguistic-based features like density of POS (PD) and different weighting scheme like Term frequency (TF), Term frequency inverse document frequency (TFIDF) are explored. Furthermore, each story in the collection can be viewed as a vector with one component corresponding to each term in story corpus, together with a weight for each component given TF or TFIDF. Different combinations of features vectors are considered for evaluation. R statistical programming language is used for feature extraction [12].

- **Term Frequency (TF)**: Frequency of terms in a document is calculated. Importance of a word within a story genre is given by TF measure.
- **Term Frequency Inverse Document Frequency (TFIDF)**: For a term, weight is assigned as a product of TF and IDF. IDF is calculated as

$$idf(t_i) = log\frac{N}{n_i}$$

where N is the total number of stories and n_i is the number of stories in corpus that contains word t_i. Importance of a word across story genre is given by TFIDF measure.

- **POS Density (PD)**: The relevance of the POS tags with respect to Indian languages are explained in Section IV-A. POS tags used here are: Noun (NN), Proper Noun (NNP), Spatial and Temporal Nouns (NST), Pronoun (PRP), Finite Verb (VM), Auxiliary Verb (VAUX), Post Position (PSP), Particles (RP), Adjective (JJ) and Quantifiers (QF). For each document, PD is used as a feature vector. It is calculated as

$$PD = \sum_{p \in P} \frac{count(p)}{Total\ words\ in\ document}$$

where P = NN, VM, PRP, VAUX, NNP, NST, PSP, RP, JJ and QF.

IV. Experiments

A. Analysis of POS Tags According to Story Genres

Guidelines for POS tagging in Indian languages are described in shallow parser manual[2]. Details of the POS tags selected in this work are presented in Table II and the importance of some POS tags are explained below.

[1] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
[2] http://ltrc.iiit.ac.in/tr031/posguidelines.pdf

- **Noun (NN) and Proper Noun (NNP):** A common noun refers to person, place or thing. Proper nouns helps in identifying specific names of person and place. Unlike English, Indian languages do not have orthographic conventions for proper nouns. Common nouns and proper nouns as a good discriminator for story classification.
- **Spatial and Temporal Noun (NST):** A noun denoting spatial and temporal expressions. It is used to differentiate from common nouns. This tag is equivalent to an adverb in some context.
- **Adjective (JJ):** Importance of adjective in stories are noticed. The comparative and superlative form of adjectives are also considered in JJ tag.
- **Quantifiers (QF):** Quantifiers are used to intensify adjectives. The combination of quantifiers and adjectives can be used to distinguish between genres of stories.

TABLE II. COUNT OF POS TAGS FOR STORY GENRES

POS Tags	Category		
	Fable	Folk-tale	Legend
NN	10975	9985	7277
VM	9298	8439	6098
PSP	6788	6249	4898
PRP	5286	4910	3761
VAUX	4278	3735	2817
JJ	1691	1698	1420
NNP	1534	1497	1554
RP	1456	1353	1011
NST	1035	764	584
QF	635	530	503

Distribution of POS tags across story genres are given in Figure 2. It is unclear that which POS tags may be important for recognition of story genres. For better investigation of the effect of linguistic information on story classification, we carried out experiments with variable sets of POS tags. The POS groups are listed in Table III.

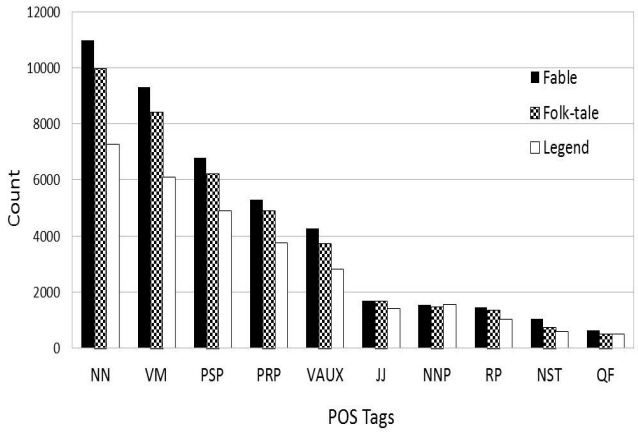

Fig. 2. POS Distribution across story genres

B. Evaluation

In this work, we used WEKA [13] as a framework combined with LibSVM [14] for classification. As discussed in Section III-C, five different combinations of features are used in this work: PD, TF, TFIDF, TF + PD and TFIDF + PD. POS tags mentioned in Set 2 (Table III) are used as features for calculating POS Density (PD). Features are normalized to fit in the range of [0,1]. The performance of different POS tag sets and feature selection methods are evaluated using three classifiers: Naive Bayes (NB), k-Nearest Neighbour (KNN) and Support Vector Machine (SVM).

TABLE III. DIFFERENT SETS OF POS TAGS

Set	POS Tags
Set 1	$\{NN, NNP, NST, PRP, JJ, QF, VM, VAUX, PSP, RP\}$
Set 2	$\{NN, NNP, NST, PRP, JJ, QF\}$
Set 3	$\{NN, NNP, NST, JJ, QF\}$
Set 4	$\{NNP, NST, JJ, QF\}$
Set 5	$\{NN, NNP, NST, PRP\}$

Classifier performance is evaluated using 10-fold cross validation. For KNN, Nine nearest neighbours are used i.e. k=9. For SVM, Linear kernel is used with other default settings in WEKA. Results are evaluated using Precision (P), Recall (R), F-measure (F) and Accuracy.

$$P = \frac{No.\ of\ documents\ correctly\ classified\ as\ class\ x}{No.\ of\ documents\ classified\ as\ class\ x}$$

$$R = \frac{Proportion\ classified\ as\ class\ x}{Actual\ classified\ as\ class\ x}$$

$$F = \frac{2 \times P \times R}{(P + R)}$$

$$Accuracy = \frac{No.\ of\ documents\ correctly\ classified}{Total\ No.\ of\ documents}$$

V. RESULTS AND DISCUSSIONS

Classifier accuracies for different POS tag sets are given in Table IV. Highest classifier accuracy for NB and SVM is observed in Set 2. Even though tags like Verb-finite (VM), Auxiliary Verb (VAUX), Post position (PSP), Particles (RP) are dominant in stories, they do not contribute significantly for discriminating the story genres. For Hindi, it is observed that most of post positions are conjoined with pronoun and particles will be included in stopwords list.

Tables V, VI, VII represents performance measures of different classifiers for story genres Fable, Folk-tale and Legend respectively. In case of fable, highest F-measure is achieved by TF + PD feature using NB classifier. In case of folk-tale and legend, highest F-measure is achieved using TF + PD features using SVM classifier. For SVM classifier, there is no change in Precision, Recall and F-measure for TFIDF and TFIDF + PD feature. For all story genres, there is little improvement in Precision, Recall and F-measures after adding PD features to TF or TFIDF because of the high dimensionality of TF and TFIDF features.

TABLE IV. CLASSIFIER ACCURACIES FOR DIFFERENT POS TAG SETS

Tag Sets	Accuracy (%)		
	NB	KNN	SVM
Set 1	47	**49.3**	44.67
Set 2	**49.3**	45.3	**46**
Set 3	46.3	44	42.3
Set 4	44.3	46	43.67
Set 5	45.3	43.33	43.67

TABLE V. PERFORMANCE MEASURES FOR FABLE

Features	NB			KNN			SVM		
	P	R	F	P	R	F	P	R	F
PD	0.47	0.39	0.42	0.46	0.66	0.54	0.47	0.62	0.53
TF	0.74	**0.83**	0.78	0.52	0.7	0.59	0.91	0.42	0.57
TFIDF	0.58	0.71	0.64	0.53	0.57	0.55	0.85	0.41	0.55
TF + PD	0.75	**0.83**	**0.79**	0.53	0.71	0.61	**0.92**	0.43	0.59
TFIDF + PD	0.59	0.71	0.65	0.54	0.6	0.57	0.86	0.42	0.56

TABLE VI. PERFORMANCE MEASURES FOR FOLK-TALE

Features	NB			KNN			SVM		
	P	R	F	P	R	F	P	R	F
PD	0.36	0.48	0.41	0.37	0.38	0.37	0.34	0.46	0.39
TF	0.62	0.53	0.57	0.52	0.74	0.61	0.58	0.85	0.69
TFIDF	0.43	0.47	0.45	0.51	0.58	0.54	0.45	0.61	0.52
TF + PD	**0.64**	0.54	0.58	0.55	0.74	0.63	0.59	**0.86**	**0.7**
TFIDF + PD	0.42	0.47	0.44	0.5	0.68	0.58	0.45	0.61	0.52

TABLE VII. PERFORMANCE MEASURES FOR LEGEND

Features	NB			KNN			SVM		
	P	R	F	P	R	F	P	R	F
PD	0.54	0.46	0.49	0.6	0.34	0.43	0.67	0.23	0.34
TF	0.76	0.79	0.77	0.51	0.86	0.64	0.86	0.94	0.9
TFIDF	0.71	0.5	0.58	0.41	0.8	0.54	0.58	0.69	0.63
TF + PD	0.77	0.81	0.79	0.57	0.75	0.65	**0.87**	**0.95**	**0.91**
TFIDF + PD	0.71	0.49	0.58	0.44	0.77	0.56	0.58	0.69	0.63

VI. CONCLUSION

In this paper, Hindi stories are classified into three genres namely, fable, folk-tale and legend. The combination of keyword and linguistic features for classifying Hindi stories are proposed. A Hindi shallow parser is used for lemmatization and POS tagging. Feature vector dimension is reduced after lemmatization and stopword removal. The relevance of the POS tags for story classification with respect to Indian languages are explained. Experiments with different sets of POS tags are carried out to investigate the effect of linguistic features on story genre recognition. The performance of different POS tag sets and feature selection methods are evaluated using three classifiers: NB, KNN and SVM. 10-fold cross validation is used to evaluate classifier performance. It is observed that combining linguistic and keyword based feature do not improve significantly the classifier performance. The performance of SVM is better than NB and KNN in prediction accuracy.

Future work includes more detailed evaluations with more stories and integrating the results of story classification into text processing module of the story TTS. Apart from Hindi, the current study can be extended to other Indian languages.

ACKNOWLEDGMENT

The authors would like to thank Department of Information Technology, Govt. of India for supporting this research work, *Development of Text-to-Speech synthesis for Indian Languages Phase II*, Ref. no. 11(7)/2011HCC(TDIL).

REFERENCES

[1] E. Iosif and T. Mishra, "From Speaker Identification to Affective Analysis: A Multi-Step System for Analyzing Children Stories," *EACL*, pp. 40–49, 2014.

[2] S. Mohanty, P. Santi, R. Mishra, R. Mohapatra, and S. Swain, "Semantic Based Text Classification Using WordNets: Indian Language Perspective," in *Proceedings of the 3th International Global WordNejuh t Conf., South Jeju Island, Korea*, 2006, pp. 321–324.

[3] S. Thakur and V. Singh, "A lexicon pool augmented Naive Bayes Classifier for Nepali Text," in *Seventh International Conference on Contemporary Computing (IC3)*. IEEE, 2014, pp. 542–546.

[4] V. G. Nidhi, "Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach," in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, COLING*, 2012.

[5] M. P. Game and Pravin, "Comparison of Marathi Text Classifiers," *Association of Computer Electronics and Electrical Engineers*, vol. 4, 2014.

[6] N. Deepamala and P. R. Kumar, "Text Classification of Kannada Webpages Using Various Pre-processing Agents," in *Recent Advances in Intelligent Informatics*. Springer, 2014.

[7] R. Jayashree and M. K. Srikanta, "An analysis of sentence level text classification for the Kannada language," in *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2011.

[8] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, "Automatic classification of Tamil documents using vector space model and artificial neural network," *Expert Systems with Applications*, vol. 36, pp. 10 914–10 918, 2009.

[9] S. Kanimozhi, "Web based classification of Tamil documents using ABPA," *International Journal of Scientific & Engineering Research*, vol. 3, 2012.

[10] K. N. Murthy, "Automatic Categorization of Telugu News Articles," *Department of Computer and Information Sciences*, 2003.

[11] K. Raghuveer and K. N. Murthy, "Text Categorization in Indian Languages using Machine Learning Approaches," in *Proceedings of the 3rd Indian International Conference on Artificial Intelligence*, 2007.

[12] R. D. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.