Children Story Classification based on Structure of the Story

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 and Telugu stories based on their structure into three genres: Fable, Folk-tale and Legend. In this work, each story is divided into three parts: (i) introduction, (ii) main and (iii) climax. The objective of this work is to explore how story genre information is embedded in different parts of the story. 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. Classification performance is analyzed for different story parts using various combinations of features with three classifiers: (i) Naive Bayes (NB), (ii) k-Nearest Neighbour (KNN) and (iii) Support Vector Machine (SVM). From the experimental studies, it has been observed that classification performance has not significantly improved by combining linguistic (POS) and keyword based features. Among classifiers, SVM outperformed the other classifiers. The main part of the story has the highest classification accuracy compared to introduction and climax parts of the story.**

***K*eywords —** Story Classification; Text-to-Speech; Part-of-Speech; Vector Space Model; Naive Bayes; KNN; SVM; Structure of Story; Introduction; Main; Climax

I. Introduction

A children story can be divided into three parts: (i) introduction, (ii) main and (iii) climax. Introduction part comprises of introducing characters that are involved in the rest of story. It also describes time and place of the event. Main part constitutes the core component of the story. It has series of events and actions that relate to a central theme of the story. The central idea of the story and moral are described in the climax part. In this work, we are classifying children stories into three story genres namely fable, folk-tale and legend based on their structure. Fable is a short tale involving animals as essential characters. Folk-tale is a traditional story that is passed on in spoken form from one generation to the next. Legend is a semi-true story carrying significant meaning or symbolism for the culture in which it originates. It is based on historical factors of a particular geographic region.

The basic goal of this work is to develop a story speech synthesis system. Given a story text, the system should synthesize speech as narrated like a story teller. A story-teller narrates a story by varying prosody like pitch, duration and intensity. It is also observed that narration style depends on the story genre and hence there is a need to identify story genre from the given story text for story synthesis. Given a neutral Text-to-speech (TTS) system and prosody modification rules associated with each story genre, the TTS system should synthesize story speech. Recently, syllable-based unit selection neutral TTS systems were developed in 13 Indian languages [1]. We need to integrate prosody modification and text processing modules for the existing neutral TTS systems to synthesize the desired story speech. With this motivation, we tried to explore story genre classification and how story genre information is embedded in different parts of the story for Hindi and Telugu.

In this work, we have manually divided the stories based on their structure. We proposed a new feature based on Part-of-speech (POS) for story classification. 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 discriminating between story genres. From the literature, it has been observed that there is no existing works on story classification for Indian languages and features related to POS have not been explored for story classification. Hence, in this study we are using POS features for story classification in addition to keyword based features.

The rest of this paper is organized as follows. Section II gives an overview of the related work. Overview of the proposed framework for story classification 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

Within the context of a storytelling TTS application, a perceptual study to identify emotions in children stories was carried out [2]. In [3], children stories have been analyzed for identification of characters and personality attributes of character like age and gender. Story classification is a typical document classification problem, and it has been carried out for different domains using different approaches across languages. Most widely used approaches for text classification for Indian languages are WordNet and Machine Learning approach. In [4], ontology and hybrid based approach for classification of Punjabi text documents was proposed. They developed sports specific ontology for Punjabi and prepared 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 built rule based stemmer and Marathi word dictionary 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], stop words and restrictions based on word occurrence were used for dimensionality reduction and classified manually collected Kannada sentences from Kannada Wikipedia. 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], Telugu news articles were classified into four categories: Politics, Sports, Business and Cinema using NB classifier. In [10], language independent, corpus-based machine learning techniques were used for text categorization in ten major Indian Languages. But, there is no existing work on story classification for Indian languages which led to the motivation of the present study.

III. Story Classification Framework

Fig. 1 shows the overall framework for story classification. Short stories are collected from blogs and story books. Story corpus is cleaned and stopwords are removed. Lemmatization and POS tagging are carried out to the entire stories using a shallow parser. Feature vectors are computed using the combination of POS tag density and different weighting schemes for keywords. Output class labels are predicted using classifiers.

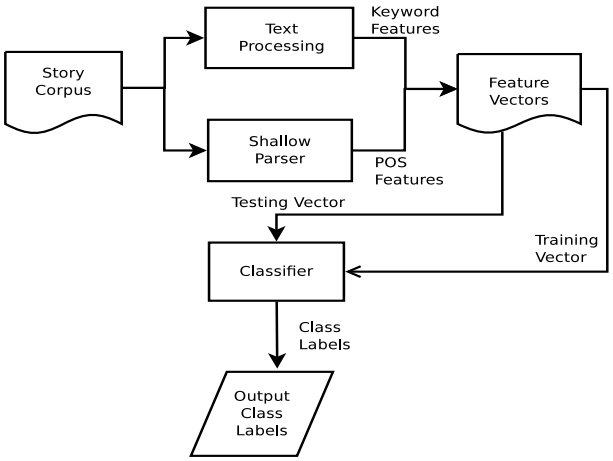

Fig. 1. Flow diagram of Story Classification Framework

A. Database Preparation

Hindi and Telugu story corpus consisting of 300 and 150 short stories respectively are collected from Blogs[1], Panchatantra and Akbar-Birbal books. No standard dataset is available for story classification for Indian languages. Details of the Hindi and Telugu story corpus are presented in Table I. Categories of stories and their definitions are explained in Section I.

B. Text Processing and POS Tagging

Text pre-processing is done to clean the story corpus which includes stripping multiple white spaces, removing special symbols and numbers. Furthermore, Hindi and Telugu shallow parser[2] developed by IIIT Hyderabad is used for lemmatization i.e. to convert each word into its root word. No standard stopword lists for Hindi and Telugu are available. A list of 164 and 138 stopwords respectively for Hindi and Telugu are prepared and used for this work. Better representation of stories is achieved after lemmatization and stopword removal because it removes unnecessary words which help in reducing feature vector dimension.

TABLE I. Statistical Information of Hindi and Telugu Story Corpus

Story genre	Hindi		Telugu	
	# Stories	# Words	# Stories	# Words
Fable	100	50344	50	6668
Folk-tale	100	46900	50	6144
Legend	100	35991	50	8540

C. Feature Extraction

Stories are projected into Vector Space Model (VSM) which uses words as features. 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. 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. Different combinations of features vectors are considered for evaluation. R statistical programming language is used for feature extraction [11].

- **Term Frequency (TF)**: Frequency of terms in a story are calculated. TF measure explains the importance of a word within a story genre.
- **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 the corpus that contains word t_i. Importance of a word across story genre is given by TFIDF measure.

- **POS Density (PD)**: Relevance of the POS tags with respect to Indian languages are explained in shallow parser manual[3]. 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

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

[1] http://telugubalalu.blogspot.in/

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

[3] http://ltrc.iiit.ac.in/tr031/posguidelines.pdf

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

IV. Experiments

A. Analysis of Story Structure

While analyzing the stories, it was observed that the story genre information is embedded in different parts of the story. We tried to explore the embedded story genre information by dividing the story into parts based on their structure. One advantage of classifying the stories based on parts is the reduction of feature vector dimension as the features are directly proportional to the number of terms in the story. With this motivation, entire stories were manually divided into three parts and labelled as: (i) introduction, (ii) main and (iii) climax. To avoid any annotation bias we took inter-annotator agreement. Three annotators were instructed regarding the annotation task. Each annotator reads the individual parts of the story and agrees whether the parts belongs to corresponding labels or not. A measure of agreement among multi-annotators is discussed in [12]. Fleiss Kappa (κ) is a statistical measure of inter-annotator agreement and κ = 0.774 is observed from the annotation that is considered to be substantial agreement. Story classification have been explored by building models for the individual parts of stories divided based on their structure.

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 employed in this work: PD, TF, TFIDF, TF + PD and TFIDF + PD. Baseline corresponds to the TF features for the individual part of stories without lemmatization and stopword removal. POS tags mentioned in III-C are used as features for calculating POS Density (PD). The performance of feature selection methods are evaluated using three classifiers: Naive Bayes (NB), k-Nearest Neighbour (KNN) and Support Vector Machine (SVM). In [15], a comparative study of different types of approaches for text classification are presented and concluded that NB, KNN and SVM are the most appropriate learning algorithms for text classification in the existing literature. Moreover, these three algorithms have shown better performance than other algorithms [16], [17]. These results motivated us to select the three promising machine learning algorithms in this study. Classifier performance is evaluated using 10-fold cross validation. For KNN, nine nearest neighbours are used i.e. k=9. For SVM, a linear kernel is used with other default settings in WEKA. Results are evaluated using Precision (P), Recall (R), F-measure (F) and Accuracy.

V. Results and Discussions

Tables II, III and IV represents story classification performance measures for introduction, main and climax part of the story respectively. In case of models built using only the introduction part of the story, highest F-measure of 0.88 and 0.85 is achieved by TF + PD feature using NB and SVM classifier for Hindi and Telugu respectively. For the models built using only the main part and climax part of the story, highest F-measure of 0.86 and 0.83 is achieved by TF + PD feature using SVM classifier for Hindi and Telugu respectively. It is noted that 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.

Figures 2, 3 and 4 represents the classification accuracies for introduction, main and climax part of the story for different features for Hindi and Telugu. From the figures, it can be noted that for most of the feature groups,(i) KNN has the least accuracy, (ii) SVM has the highest accuracy and (iii) NB has relatively lesser accuracy compared to SVM but better than KNN. For different story types, the accuracy achieved using TF + PD features are higher than the rest of the features. The main part of the story has the highest classification accuracy compared to introduction and climax part of the story.

VI. Conclusion

In this paper, based on the structure of the children stories, Hindi and Telugu stories are classified into three genres namely, fable, folk-tale and legend. Entire stories are manually divided into introduction, main and climax parts. Story structure is analyzed to explore the embedded story genre information in different parts of the story. The combination of keyword and linguistic features for classifying stories are proposed. The shallow parser is used for lemmatization and POS tagging. Models are built by considering only the introduction, main and climax part of the story. The performance of story classification for different story parts using various feature selection methods is evaluated with 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 the classifier performance significantly. The performance of SVM is better than NB and KNN in classification accuracy. Models built using only main part of the story outperformed the models built using introduction and climax part in terms of classification 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.

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] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra *et al.*, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–8.

[2] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling." in *INTERSPEECH*, 2005, pp. 533–536.

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

[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.

TABLE II. STORY CLASSIFICAITON PERFORMANCE MEASURES FOR INTRODUCTION PART OF STORY

Story Genre	Features	Hindi									Telugu								
		NB			KNN			SVM			NB			KNN			SVM		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Fable	Baseline	0.6	0.66	0.63	0.45	0.56	0.5	0.65	0.7	0.76	0.68	0.62	0.65	0.34	0.92	0.5	0.76	0.64	0.7
	PD	0.46	0.54	0.5	0.38	0.65	0.48	0.44	0.66	0.53	0.42	0.5	0.46	0.42	0.56	0.48	0.39	0.3	0.34
	TF	0.72	0.72	0.72	0.69	0.43	0.53	0.82	0.77	0.79	0.75	0.7	0.72	0.42	0.72	0.53	0.69	0.72	0.71
	TFIDF	0.67	0.71	0.69	0.65	0.34	0.45	0.7	0.82	0.75	0.69	0.7	0.69	0.37	0.8	0.51	0.72	0.66	0.69
	TF + PD	0.75	0.77	**0.76**	0.68	0.63	**0.65**	0.83	0.78	**0.8**	0.76	0.74	**0.75**	0.57	0.78	**0.66**	0.7	0.78	**0.74**
	TFIDF + PD	0.69	0.73	0.71	0.6	0.66	0.63	0.69	0.83	0.76	0.7	0.7	0.7	0.37	0.9	0.52	0.73	0.7	0.71
Folk-tale	Baseline	0.57	0.46	0.51	0.4	0.31	0.35	0.59	0.43	0.5	0.52	0.48	0.5	0.5	0.46	0.48	0.48	0.22	0.3
	PD	0.43	0.42	0.43	0.34	0.25	0.29	0.47	0.36	0.41	0.29	0.34	0.32	0.28	0.24	0.26	0.39	0.56	0.46
	TF	0.7	0.66	0.68	0.66	0.29	0.4	0.71	0.75	0.73	0.54	0.64	0.59	0.51	0.55	0.53	0.57	0.7	0.63
	TFIDF	0.6	0.47	0.53	0.49	0.26	0.34	0.61	0.44	0.51	0.46	0.44	0.45	0.49	0.53	0.51	0.46	0.24	0.32
	TF + PD	0.72	0.68	**0.7**	0.61	0.42	**0.5**	0.72	0.76	**0.74**	0.59	0.7	**0.64**	0.53	0.57	**0.55**	0.6	0.78	**0.68**
	TFIDF + PD	0.62	0.49	0.55	0.45	0.47	0.46	0.64	0.45	0.53	0.5	0.48	0.49	0.51	0.57	0.54	0.56	0.3	0.39
Legend	Baseline	0.76	0.87	0.81	0.62	0.51	0.56	0.83	0.78	0.8	0.69	0.72	0.71	0.55	0.46	0.5	0.53	0.92	0.67
	PD	0.57	0.48	0.52	0.63	0.35	0.45	0.6	0.44	0.51	0.58	0.38	0.46	0.6	0.48	0.53	0.67	0.52	0.58
	TF	0.84	0.9	0.87	0.56	0.81	0.66	0.87	0.87	**0.87**	0.81	0.68	0.74	0.48	0.6	0.53	0.97	0.7	0.81
	TFIDF	0.75	0.83	0.79	0.46	0.85	0.6	0.76	0.83	0.79	0.73	0.74	0.73	0.61	0.44	0.51	0.59	0.92	0.72
	TF + PD	0.85	0.91	**0.88**	0.59	0.81	**0.68**	0.87	0.87	**0.87**	0.8	0.7	0.75	0.47	0.72	**0.57**	0.98	0.75	**0.85**
	TFIDF + PD	0.76	0.87	0.81	0.52	0.74	0.61	0.77	0.85	0.81	0.75	0.78	**0.77**	0.91	0.38	0.54	0.62	0.95	0.75

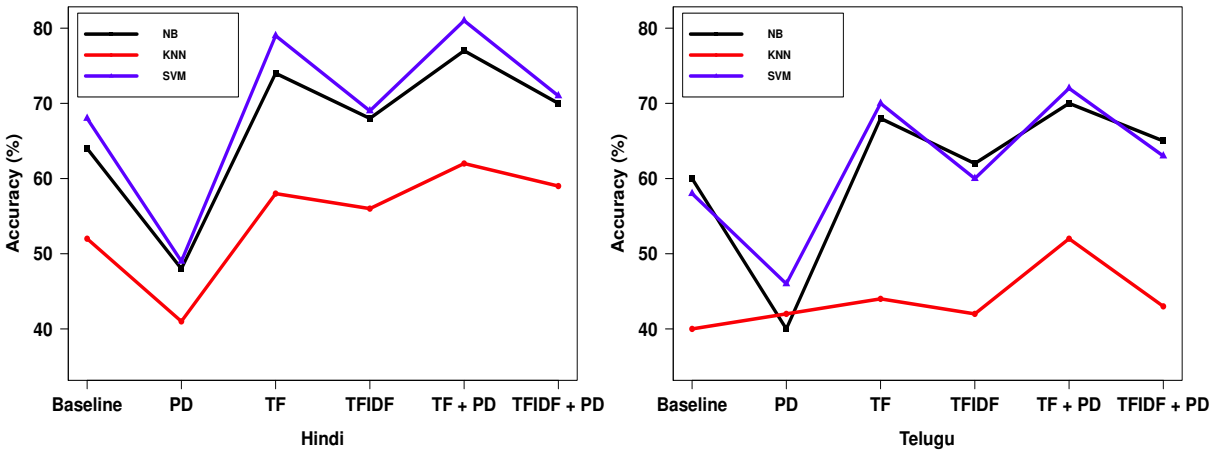

Fig. 2. Classification accuracy for models built using introduction part of story for Hindi and Telugu

[5] M. P. P. Game, "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 *Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of*, 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. 10914–10918, 2009.

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

[10] K. Raghuveer and K. N. Murthy, "Text Categorization in Indian Languages using Machine Learning Approaches." in *IICAI*, 2007.

[11] R Development Core 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

[12] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[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*,

TABLE III. STORY CLASSIFICAITON PERFORMANCE MEASURE FOR MAIN PART OF STORY

Story Genre	Features	Hindi									Telugu								
		NB			KNN			SVM			NB			KNN			SVM		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Fable	Baseline	0.66	0.6	0.63	0.58	0.32	0.41	0.63	0.44	0.52	0.71	0.7	0.71	0.59	0.2	0.3	0.79	0.66	0.72
	PD	0.39	0.45	0.42	0.48	0.33	0.39	0.38	0.65	0.48	0.54	0.88	0.67	0.48	0.78	**0.60**	0.47	0.68	0.56
	TF	0.7	0.86	0.77	0.63	0.36	0.46	0.83	0.8	0.82	0.71	0.78	0.74	0.45	0.28	0.35	0.8	0.76	0.78
	TFIDF	0.68	0.64	0.66	0.6	0.36	0.45	0.66	0.5	0.57	0.59	0.72	0.65	0.62	0.16	0.25	0.77	0.6	0.67
	TF + PD	0.79	0.86	**0.79**	0.69	0.55	**0.61**	0.84	0.86	**0.85**	0.77	0.82	**0.8**	0.38	0.38	0.38	0.87	0.8	**0.83**
	TFIDF + PD	0.7	0.65	0.67	0.64	0.44	0.52	0.69	0.72	0.71	0.62	0.75	0.68	0.81	0.26	0.39	0.77	0.72	0.74
Folk-tale	Baseline	0.39	0.38	0.38	0.4	0.28	0.33	0.38	0.65	0.48	0.48	0.46	0.47	0.34	0.71	0.46	0.42	0.76	0.54
	PD	0.4	0.38	0.39	0.42	0.4	0.41	0.31	0.23	0.26	0.39	0.20	0.26	0.24	0.16	0.19	0.36	0.4	0.38
	TF	0.63	0.55	0.59	0.52	0.35	0.42	0.73	0.71	0.72	0.63	0.58	0.6	0.35	0.7	0.47	0.61	0.82	0.7
	TFIDF	0.39	0.38	0.38	0.5	0.25	0.33	0.39	0.76	0.52	0.52	0.48	0.5	0.34	0.94	**0.5**	0.46	0.78	0.58
	TF + PD	0.7	0.54	**0.61**	0.6	0.36	**0.45**	0.72	0.76	**0.74**	0.69	0.62	**0.65**	0.34	0.82	0.48	0.64	0.86	**0.74**
	TFIDF + PD	0.41	0.43	0.42	0.52	0.26	0.35	0.53	0.64	0.58	0.53	0.51	0.52	0.36	0.96	0.52	0.48	0.8	0.6
Legend	Baseline	0.6	0.56	0.58	0.38	0.73	0.5	0.8	0.26	0.39	0.6	0.64	0.62	0.48	0.52	0.5	0.44	0.96	0.61
	PD	0.51	0.46	0.48	0.51	0.4	0.45	0.58	0.32	0.41	0.62	0.52	0.57	0.56	0.4	0.47	0.73	0.32	0.44
	TF	0.78	0.73	0.75	0.4	0.83	0.54	0.81	0.87	0.84	0.76	0.75	0.75	0.52	0.45	0.48	0.9	0.69	0.78
	TFIDF	0.62	0.67	0.64	0.38	0.88	0.53	0.9	0.28	0.43	0.61	0.7	0.65	0.5	0.48	0.49	0.46	0.92	0.61
	TF + PD	0.76	0.78	**0.77**	0.42	0.89	**0.57**	0.83	0.89	**0.86**	0.77	0.8	**0.78**	0.54	0.5	**0.52**	0.95	0.7	**0.81**
	TFIDF + PD	0.63	0.65	0.64	0.4	0.93	0.56	0.85	0.63	0.72	0.6	0.7	0.65	0.52	0.48	0.5	0.5	0.9	0.64

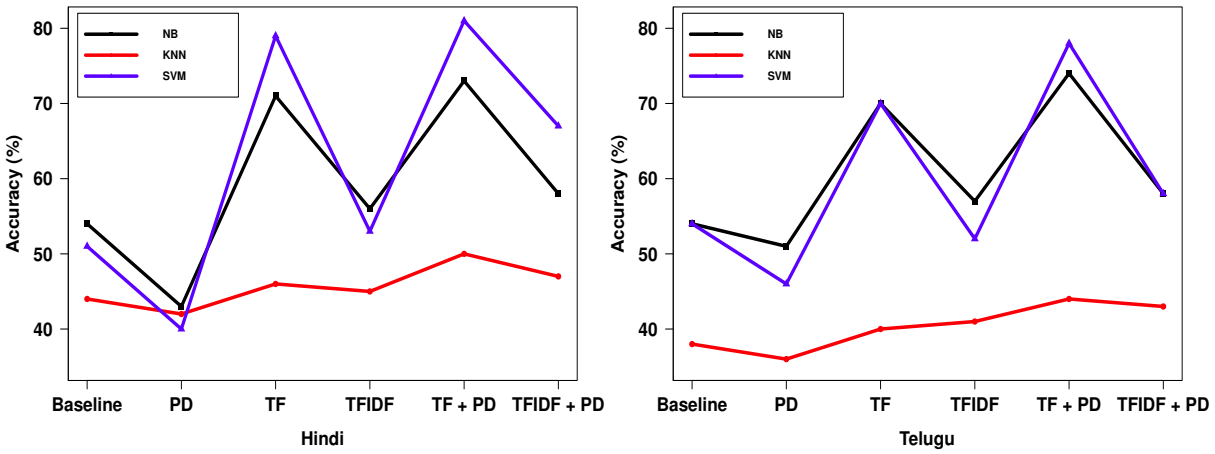

Fig. 3. Classification accuracy for models built using main part of story for Hindi and Telugu

vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] P. Y. Pawar and S. Gawande, "A comparative study on different types of approaches to text categorization," *International Journal of Machine Learning and Computing*, vol. 2, no. 4, pp. 423–426, 2012.

[16] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[17] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 42–49.

TABLE IV. STORY CLASSIFICAITON PERFORMANCE MEASURE FOR CLIMAX PART OF STORY

Story Genre	Features	Hindi									Telugu								
		NB			KNN			SVM			NB			KNN			SVM		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Fable	Baseline	0.62	0.64	0.63	0.77	0.27	0.4	0.73	0.57	0.64	0.59	0.74	0.66	0.44	0.24	0.31	0.8	0.32	0.46
	PD	0.44	0.52	0.48	0.42	0.54	0.47	0.47	0.6	0.53	0.47	0.76	0.58	0.38	0.58	0.46	0.38	0.24	0.29
	TF	0.74	0.7	0.72	0.56	0.31	0.4	0.85	0.74	0.79	0.72	0.74	0.73	0.42	0.7	0.52	0.76	0.8	0.78
	TFIDF	0.65	0.68	0.66	0.86	0.31	0.46	0.71	0.74	0.73	0.57	0.66	0.61	0.35	0.88	0.5	0.61	0.52	0.56
	TF + PD	0.75	0.71	**0.73**	0.66	0.37	0.47	0.86	0.75	**0.8**	0.76	0.76	**0.76**	0.39	0.86	**0.54**	0.78	0.82	**0.8**
	TFIDF + PD	0.66	0.7	0.68	0.84	0.34	**0.48**	0.72	0.76	0.74	0.59	0.66	0.62	0.37	0.92	0.53	0.66	0.54	0.59
Folk-tale	Baseline	0.44	0.44	0.44	0.45	0.38	0.41	0.41	0.39	0.4	0.5	0.36	0.42	0.31	0.76	0.44	0.6	0.43	0.5
	PD	0.35	0.28	0.31	0.38	0.36	0.37	0.33	0.26	0.29	0.26	0.14	0.18	0.33	0.26	0.29	0.37	0.64	0.47
	TF	0.51	0.54	0.53	0.49	0.43	0.46	0.66	0.76	0.71	0.62	0.64	0.63	0.4	0.46	0.43	0.62	0.72	0.67
	TFIDF	0.46	0.43	0.45	0.46	0.44	0.45	0.49	0.46	0.47	0.52	0.46	0.49	0.4	0.51	0.45	0.6	0.66	0.63
	TF + PD	0.53	0.56	**0.55**	0.49	0.47	**0.48**	0.68	0.77	**0.72**	0.65	0.64	**0.65**	0.41	0.52	0.46	0.64	0.75	**0.69**
	TFIDF + PD	0.47	0.45	0.46	0.46	0.46	0.46	0.51	0.47	0.49	0.54	0.5	0.52	0.42	0.53	**0.47**	0.62	0.68	0.65
Legend	Baseline	0.7	0.68	0.69	0.36	0.6	0.45	0.65	0.59	0.62	0.69	0.7	0.69	0.6	0.66	0.63	0.38	0.94	0.54
	PD	0.47	0.47	0.47	0.42	0.33	0.37	0.48	0.45	0.46	0.71	0.6	0.65	0.71	0.48	0.57	0.77	0.48	0.59
	TF	0.74	0.74	0.74	0.37	0.88	0.52	0.86	0.85	0.85	0.77	0.82	0.8	0.65	0.74	0.69	0.92	0.72	0.81
	TFIDF	0.71	0.72	0.71	0.38	0.97	0.55	0.69	0.7	0.69	0.77	0.74	0.76	0.61	0.59	0.6	0.51	0.82	0.63
	TF + PD	0.75	0.76	**0.76**	0.4	0.95	0.56	0.87	0.85	**0.86**	0.81	0.85	**0.83**	0.68	0.77	**0.72**	0.94	0.74	**0.83**
	TFIDF + PD	0.71	0.73	0.72	0.42	0.89	**0.57**	0.68	0.71	0.7	0.79	0.76	0.78	0.62	0.64	0.63	0.55	0.78	0.65

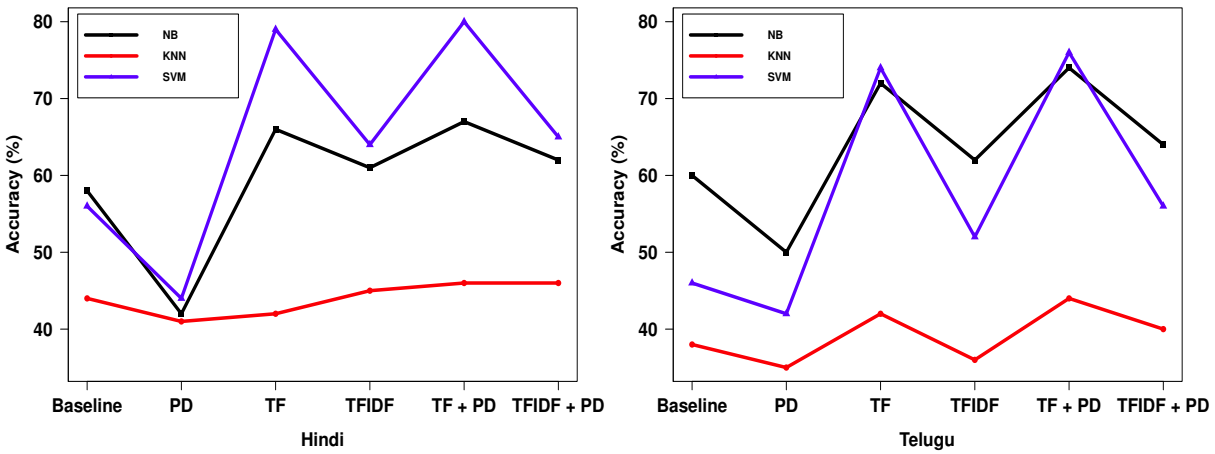

Fig. 4. Classification accuracy for models built using climax part of story for Hindi and Telugu