POS Word Class Based Categorization of Gurmukhi Language Stemmed Stop Words

Kaur Jasleen and R. Saini Jatinderkumar

Abstract Literature in Indian language must be classified for its easy retrieval. In Punjabi literature classifier, five different categories: nature, romantic, religious, patriotic and philosophical, are manually populated with 250 poems. These poems are pre-processed through data cleaning, tokenization, bag of word, stop word identification and stemming phases. Due to unavailability of Punjabi stop words in public domain, manual collection of 256 stop words are done from poetry and articles. After stemming, 184 unique stemmed words are identified. Based on part of speech tagging, 184 stop words are categorized into 98 adverbs, 7 conjunctions, 43 verbs, 24 pronouns and 12 miscellaneous words. These unique 184 stemmed words are being released for other language processing algorithm in Punjabi. This paper concentrates on providing better and deeper understanding of Punjabi stop words in lieu of Punjabi grammar and part of speech based word class categorization.

Keywords Adverb · Conjunction · Verb · Pronoun · Part of speech · Punjabi · Stop word

1 Introduction

With the advent of World Wide Web and Unicode encoding, Indian language content is increasing on the web day by day. In today's internet era, people prefer to use their regional language or mother language to communicate their thoughts. So this data must be classified for its easy retrieval and usage. Text Classification is an act of assigning natural language text into predefined categories [1]. India, being a

K. Jasleen (✉)
Uka Tarsadia University, Bardoli, Gujarat, India
e-mail: sidhurukku@yahoo.com

R. Saini Jatinderkumar
Narmada College of Computer Application, Bhaurch, Gujarat, India
e-mail: saini_expert@yahoo.com

S.C. Satapathy and S. Das (eds.), *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*,
Smart Innovation, Systems and Technologies 51, DOI 10.1007/978-3-319-30927-9_1

multilingual country, consists of wide number of languages and rich literature. Out of these languages, 22 languages are recognized as regional languages [2]. Punjabi, one of them, is widely spoken language in Punjab (India) as well as in Pakistan [3]. Punjabi Language belongs to Indo-Aryan Language Family. Punjab is known for its rich culture and literature. Poem is one form of literary art. Poetry is always imaginative in nature with a message to its reader [4]. Poetry always has a strong association with feelings, thoughts and ideas. An automatic poetry classification is a text classification problem. Input to classifier is poem in Punjabi language and classifier will assign a category on the basis of its content. This paper is focused on the analysis of stop words from grammatical point of view.

2 Indian Language Based Text Classifier

India is a multilingual country. Many languages are being used in India. Indo-Aryan (consists of Hindi, Gujarati, Bengali, Punjabi, Marathi, Urdu, and Sanskrit) and Dravidian (Telugu, Tamil, Kannada) are major language families spoken in India [5]. Brief survey about the text classification works done in Indian languages is given below.

2.1 Text Classification in Indo-Aryan Language Family

Statistical techniques using Naïve Bayes and Support Vector Machine are used to classify subjective sentences from objective sentences for Urdu language. As Urdu language is morphological rich language, this makes the classification task more difficult. The result of this implementation shows that accuracy, performance of Support Vector Machines is much better than Naïve Bayes classification techniques [6]. For Bangla text classification, n-gram based algorithm is used and to analyze the performance of the classifier. Prothom-Alo news corpus is used. The result show that with increase in value of n from 1 to 3, performance of the text classification also increases, but from value 3 to 4 performance decreases [7]. Sanskrit text documents have been classified using Sanskrit Word net. Semantic based classifier is method is built on lexical chain of linking significant words that are about a particular topic with the help of hypernym relation in Word Net [8]. Very few works in literature are found in field of text classification in Punjabi Language. Domain based text classification is done by Nidhi and Vishal [9]. This classification is done on sports category only. Two new algorithms, Ontology based classification and Hybrid approach are proposed for Punjabi text classification. The experimental results conclude that Ontology based classification (85 %) and Hybrid approach (85 %) provides better results. Sarmah et al. [10] presented an approach for classification of Assamese documents using Assamese WordNet. This approach has accuracy of 90.27 % on Assamese documents.

2.2 Text Classification in Dravidian Language Family

Naïve Bayes classifier has been applied to Telugu news articles to classify 800 documents into four major classes. In this, normalized term frequency-inverse document frequency is used to extract the features from the document. Without any stop word removal and morphological analysis, at the threshold of 0.03, the classifier gives 93 % precision [11]. For morphologically rich Dravidian classical language Tamil, text classification is done using vector space model and artificial neural network. The experimental results show that Artificial Neural network model achieves 93.33 % which is better than the performance of Vector Space Model which yields 90.33 % on Tamil document classification [12]. A new technique called Sentence level classification is done for Kannada language; in this sentences are analyzed to classify the Kannada documents. This Technique extended further to sentiment classification, question answering, text summarization and also for customer reviews in Kannada blogs [13].

3 Pre-processing Steps Involved in Punjabi Poetry Classifier

Before classification, data must be pre processed to remove unwanted words and noise [14]. Pre-processing phase of poetry classifier consists of Data Cleaning, Feature Extraction and Feature Selection. As Punjabi is resource scarce language, there is no publicly available corpus, so manual collection of poetry is done. Initially, Data is collected into 5 different categories: ਕੁਦਰਤ [*kudarata*] 'Nature', ਪ੍ਰੀਤ [*prīta*] 'Romantic', ਧਾਰਮਿਕ [*dhāramika*] 'Religious', ਦੇਸ਼ਭਗਤੀ [*dēśabhagatī*] 'Patriotic' and ਦਾਰਸ਼ਨਿਕ [*dāraśanika*] 'Philosophical'. Initially, these categories are populated with 50 poems in each category. Implementation of various subphases (as discussed below) is done in Visual Basic.Net using Microsoft Visual Studio 2010 as front end and Microsoft Access 2007 at back end using Unicode characters [15].

3.1 Cleaning

Preprocessing step is involved to remove the noise from data so that this noisy data don't penetrate into the next higher levels. It includes special symbol deletion. Symbols like: comma (,), dandi (|), double dandi (||), sign of interrogation (?) and sign of exclamation (!) are present in poems. In case of Punjabi language, dandi (|) is used in place of full stop. Double dandi (||) is generally used in ancient Punjabi writings like religious poetry.

3.2 Feature Extraction and Feature Selection

Feature Extraction phase consists of tokenization, unique words and its term frequency calculation, stop word removal and stemming. In tokenization, each poem is tokenized and 'bag of word' model is created. Data structures used for implementation are hash tables, files and arrays. Unique words are identified from poems and its frequency is calculated from tokenized words. After this, stop words are eliminated from unique words. Stop words are most common words occurring in the text which are not significant for classifier. 256 stop words are identified from poetry, news articles and other Punjabi stories. Stemming is way of converting a written text into its root form [16]. Gupta [17] developed different rules for handling stemming for verbs, adverbs and pronouns. These stemming rules are manually applied to 256 identified stop words. After stemming, 184 unique stemmed stop words are identified and presented in Table 1. This table consists of Columns: word in Punjabi (C1), its transliteration in English (C2) and its meaning in English (C3) [18, 19].

Table 1 List of stemmed stop words

S. no.	C1	C2	C3	S. no.	C1	C2	C3
1	ਇਸ	[isa]	This	2	ਜਸਿ	[jisa]	Who, what, which
3	ਵਚਿ	[vica]	In the	4	ਨ	[na]	No
5	ਤਕ	[taka]	Up	6	ਹੁਣ	[huṇa]	Now
7	ਵੀ	[vī]	Too	8	ਜਨਿਾਂ	[jināṁ]	Whom
9	ਉਤੋਂ	[othon]	Upon	10	ਨਾਲ	[nāla]	With
11	ਨਹੀ	[nahīṁ]	No	12	ਚਾਹੇ	[cāhē]	Either
13	ਭੀ	[bhī]	Too	14	ਕਸਿ	[kisa]	What
15	ਵਲੋਂ	[valōṁ]	By	16	ਪਛਿੋਂ	[pichōṁ]	After
17	ਇਹ	[iha]	This	18	ਏਧਰ	[ēdhara]	Around
19	ਏ	[iha]	This	20	ਨੂੰ	[nū]	To
21	ਜਦੋਂ	[jadōṁ]	When, while	22	ਅਜਹਿੇ	[ajihē]	Such
23	ਕਈ	[ka'ī]	Many	24	ਹੀ	[hī]	Only
25	ਤੱਦ	[tada]	Then	26	ਕੇ	[kē]	By
27	ਅੰਦਰ	[andar]	Within	28	ਹਾਂ	[hain]	Yes
29	ਉੱਤੇ	[utē]	Upon	30	ਬਹੁਤ	[bahuta]	Much
31	ਸਾਬੁਤ	[sābuta]	Complete	32	ਕਾਫ਼ੀ	[kāfī]	Enough
33	ਕਦੀ	[kadī]	Sometime	34	ਹੁਣੇ	[huṇē]	Now
35	ਨੇਂ	[nēṁ]	The	36	ਲਈ	[la'ī]	For
37	ਜੀ	[jī]	Respect	38	ਕਿ	[ki]	That
39	ਕਸਿੇ	[kisē]	Someone	40	ਮਗਰ	[magara]	Behind
41	ਪੂਰਾ	[pūrā]	Complete	42	ਦਾ	[dā]	Of
43	ਨੇ	[nē]	The	44	ਤਰ੍ਹਾਂ	[tar'hāṁ]	Like
45	ਹੋਵੇ	[hovē]	If	46	ਫੇਰ	[phēra]	Later

(continued)

Table 1 (continued)

S. no.	C1	C2	C3	S. no.	C1	C2	C3
47	ਜੇਕਰ	[jēkar]	Just in case	48	ਵੇਲੇ	[vēlē]	Times
49	ਦੇ	[dē]	Of	50	ਉੱਥੇ	[othē]	There
51	ਜਹਿੜਾ	[jēhara]	Which	52	ਕਤਿੇ	[kitē]	Somewhere
53	ਬਾਅਦ	[bā’ada]	After	54	ਇੱਥੇ	[ithē]	Here
55	ਸਾਰਾ	[sārā]	all,whole	56	ਜਨਿੂੰ	[jinhanu]	Whom
57	ਚੋ	[cho]	Out	58	ਜਦ	[jad]	When
59	ਕਦੀ	[kadē]	Never	60	ਵਾਂਗ	[vāṅga]	Like
61	ਸਭ	[sab]	All	62	ਦੌਰਾਨ	[doraan]	During
63	ਤਾਂ	[tan]	When	64	ਵਰਗਾ	[varagā]	Like
65	ਕਿ	[ki]	That	66	ਜੋ	[jō]	That
67	ਲਾ	[la]	To attach	68	ਕਰਕੇ	[karkē]	Because
69	ਪੂਰਾ	[pura]	Complete	70	ਬਲਿਕੁਲ	[bilkul]	Absolutely
71	ਨਾਲੇ	[naale]	Also	72	ਐਹੋ	[eho]	Such
73	ਤੋ	[ton]	From	74	ਕੌਣ	[kaun]	Who
75	ਹੋਣਾ	[hona]	Be	76	ਫਰਿ	[pher]	Then
77	ਪਾਸੋ	[paso]	From	78	ਤਦ	[tad]	Then
79	ਜਹਿਾ	[jeha]	Little	80	ਕੋਲੋ	[kolon]	From
81	ਏਸ	[ēs]	This	82	ਕਨਿ੍ਹਾ	[kina]	How much
83	ਜਨਿ੍ਹਾਂ	[jina]	Who	84	ਜਵਿੇ	[jivē]	Such as
85	ਕੁਝ	[kujh]	Some	86	ਹੇਠਾਂ	[hethan]	Below
87	ਦੁਆਰਾ	[dobara]	By	88	ਸਾਰੇ	[sarē]	All
89	ਸਦਾ	[sada]	Forever	90	ਜੱਥਿ	[jithē]	Where
91	ਏਥੇ	[ethē]	Here	92	ਕੋਈ	[koi]	Someone
93	ਬਾਰੇ	[barē]	About	94	ਕੀ	[ki]	What
95	ਕਦ	[kad]	When to	96	ਜੀ	[je]	Please
97	ਕਦੇ	[kadē]	Never	98	ਦੀਆਂ	[dī’āṁ]	Of
99	ਹੋਏ	[hoye]	Happen	100	ਚਲਾ	[chala]	Goes
101	ਰਹੇ	[rahē]	Are	102	ਲੈ	[lai]	Take
103	ਬਣੋ	[bano]	Become	104	ਆਖ	[aakh]	Say
105	ਦੇਣੀ	[dēṇī]	Give	106	ਬਣ	[baṇa]	Made
107	ਪਇਆ	[pi’ā]	Lying	108	ਕਰ	[kara]	Do
109	ਹੋਇਆ	[hō’i’ā]	Happened	110	ਪੈਣ	[pain]	Falling
111	ਗਈ	[ga’ī]	Gone	112	ਕਹਿ	[kēh]	Say
113	ਲਗ	[laga]	Seem	114	ਚੁਕੇ	[chukē]	–
115	ਹੁੰਦਾ	[hudā]	Happen	116	ਕਹਿਾ	[keha]	Said
117	ਜਾਂਦਾ	[jāndā]	Going	118	ਕਰਵਾਈ	[karvayei]	Conducted
119	ਵੇਖ	[vēkha]	See	120	ਬਣਾਏ	[banaye]	Created
121	ਸੁਣ	[suṇa]	Hear	122	ਕੀਤਾ	[kitta]	Carried out
123	ਆਈ	[ā’ī]	Occurred	124	ਜਾਵਣ	[javan]	Going
125	ਸਕਦੇ	[sakdē]	Can	126	ਦੇਖ	[dēkh]	See

(continued)

Table 1 (continued)

S. no.	C1	C2	C3	S. no.	C1	C2	C3
127	ਜਾਵੇ	[javē	Go	128	ਆਦਿ	[ādi]	So on
129	ਜਾਂਦਾ	[janda]	Going	130	ਲਿਆ	[li'ā]	Taken
131	ਕਰਣ	[karana]	Doing	132	ਆ	[ā]	Come
133	ਲਗਾਉਦਾਂ	[lagoda]	Not involving	134	ਰਹਿਾ	[reha]	Going
135	ਆਵੇ	[aavē]	Arrives	136	ਗਿਆ	[geya]	Been
137	ਕਰੀ	[kari]	Do	138	ਉਠ	[otha]	Arise
139	ਲਾਇਆ	[laeya]	Attach	140	ਰਹੀ	[rahi]	Been
141	ਰਹਿ	[reh]	Living	142	ਉਸਨੇ	[usnē]	He
143	ਉਹ	[uha]	He, she	144	ਤੁਸੀ	[tusi]	You
145	ਸਾਂ	[sāṁ]	Was	146	ਮੇਰਾ	[mera]	My
147	ਸਭ	[sabha]	All	148	ਉਸਦੀ	[usdi]	His
149	ਹਨ	[hana]	Are	150	ਤੇਰਾ	[tera]	Your
151	ਤੂੰ	[tu]	You	152	ਉਸ	[us]	His
153	ਸੀ	[si]	Was	154	ਉਏ	[oyē]	Person
155	ਹੋ	[ho]	Are	156	ਆਪ	[aap]	you
157	ਤੈਨੂੰ	[tēnu]	You	158	ਸਨ	[san]	Was
159	ਤੁਸਾਂ	[tusa]	You	160	ਮੈ	[mein]	I
161	ਹੈ	[hain]	Are	162	ਤੁਸੀ	[tusi]	You
163	ਹੈ	[hai]	Is	164	ਅਸੀ	[assi]	We
165	ਆਪਣਾ	[apna]	My	166	ਪਰ	[par]	but
167	ਜੇ	[jē]	If	168	ਤੇ	[tē]	And
169	ਅਤੇ	[aatē]	And	170	ਤਾਂ	[tāṁ]	So
171	ਜਾਂ	[jāṁ]	Or	172	ਭਾਵੇਂ	[bhāvēm]	Although
173	ਕੁਲ	[kal]	Total	174	ਅਗਲੀ	[aagali]	Next
175	ਵਗ਼ੈਰਾ	[vaġairā]	Etc.	176	ਵਰਗ	[varg]	Category
177	ਰੱਖ	[rakh]	Put	178	ਆਮ	[āma]	Common
179	ਲੱਗ	[laag]	Take	180	ਲਾ	[lā]	Apply
181	ਗੱਲ	[gal]	Thing	182	ਹਾਲ	[hāla]	Condition
183	ਪੀ	[pī]	Drink	184	ਇੱਕ	[ek]	One

On lieu of Punjabi Grammar and Part of Speech (POS) based word class categorization, stop words are categorized into 4 different word classes: Adverbs [20], Conjunctions [21], Verbs [20], Pronouns [20] and other miscellaneous words. Any word which is not suitable for first four categories is assigned to miscellaneous one. 98 different adverb forms, 43 different verbs, 24 pronouns, 7 conjunctions are identified from 184 stemmed stop words. And remaining 12 stop words are assigned to miscellaneous category.

Adverb forms in Punjabi language are classified into 2 categories: by function and by form [20]. By function, adverb clauses are categorized into following subclasses: Adverb clause of time: ਜਦ [*jad*] ‘when’, ਅੱਜ [*ajj*] ‘today’. Adverb clause of place: ਉੱਪਰ [*uppar*] ‘upon’, ਉੱਤੇ [*uttē*] ‘over’. Adverb clause of purpose: ਨੂੰ [*nu*] ‘to’, ਲਈ [*laii*] ‘for’. Adverb clause of manner: ਜਵਿੇ [*jive*] ‘as’, ਉਵੇ [*ove*] ‘custom’. Condition clause: ਅਗਰ [*agar*] ‘if’, ਜੇਕਰ *[jekar]* ‘in case’. Result clauses: ਨਾਲੋ [*naalo*] ‘concurrent’, ਤੋ [*to*] ‘from’. Adverb clause of degree: ਬਹੁਤ [*bahut*] ‘much’, ਕਾਫੀ [*kafi*] ‘*enough*’, ਸਾਬੁਤ [*sabut*] ‘complete’. By form, adverb clause is divided into subgroups like derived adverbs, pure adverbs, phrasal adverbs, clausal adverbs, reduplicated adverbs and particles. Few examples are like ਇਥੇ [*ethe*] ‘here’, ਉਥੇ [*othe*] ‘there’, ਕਥਿ [*kithe*] ‘where’, ਜਥਿ [*jitho*] ‘where’, ਕਥਿ [*kitho*] ‘where’. List of Adverbs are shown from serial number 1–98 in Table 1.

Verbs found among stop words are shown in Table 1 from serial number 99–141. For example: ਆਉਣਾ [*aaouna*] ‘to come’, ਜਾਣਾ [*jaana*] ‘to go’.

Pronouns are shown from serial number 142–165. For example ਉਸਦਾ [*usda*] ‘his’, ਅਸੀ [*assi*] ‘we’.

Conjunctions are used to join words, phrases, and clauses [21]. For example, ਅਤੇ [*atē*] ‘and’, ਜਾਂ [*jāṃ*] ‘or’. Serial number 166–172 presents conjunction list.

Miscellaneous words are present from serial number 173–184. For example: ਵਰਗ [*varg*] ‘category’, ਵਗੈਰਾ [*vagera*] ‘etc’.

4 Conclusion

An automatic poetry classifier is used to classify poems according to its content. Before classification starts, these poetries must have to pass through various pre-processing phases. 256 stop words are identified from poetry and news articles written in Punjabi. These 256 stop words are stemmed to its root form using Punjabi stemming rules. Analysis of 184 stemmed stop words from grammatical point of view is discussed in this paper. These stop words are categorized into adverbs, pronouns and conjunctions. In this paper, 184 stemmed stop words are presented for future use in other NLP task in Gurmukhi script. This paper provides enhanced understanding of stop words in light of part of speech tags in Punjabi language.

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**, 1–47 (2002)
2. Languages of India.: http://en.wikipedia.org/wiki/Languages_of_India#Prominent_languages_of_India
3. Punjabi Language.: http://en.wikipedia.org/wiki/Punjabi_language
4. Poem.: http://oxforddictionaries.com/definition/english/poem
5. Kaur, J., Saini, J.R.: A study and analysis of opinion mining research in Indo-Aryan, Dravidian and Tibeto-Burman Language families. Int. J. Data Mining Emerg. **4**(2), 53–60 (2014)

6. Ali, R.A., Maliha, I.: Urdu text classification. In: 7th International Conference on Frontiers of Information Technology, ACM New York, USA, (2009). ISBN 978-1-60558-642-7, doi:10.1145/1838002.1838025
7. Mansur, M., UzZaman, N., Khan, M.: Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus. Center for Research on Bangla Language Processing. BRAC University, Dhaka, Bangladesh (2006)
8. Mohanty, S., Santi, P.K., Mishra, R., Mohapatra, R.N., Swain, S.: Semantic based text classification using wordnets: Indian language perspective. In: 3rd International Wordnet Conference (GWC 06). pp. 321–324 (2006). doi:10.1.1.134.866
9. Nidhi., Gupta, V.: Domain based classification Punjabi text documents. In: International Conference on Computational Linguistics, pp. 297–304 (2012)
10. Sarmah, J., Saharia, N., Sarma, S.K.: A novel approach for document classification using assamese wordnet. In: 6th International Global Wordnet Conference, pp. 324–329 (2012)
11. Murthy, K.N.: Automatic Categorization of Telugu News Articles. Department of Computer and Information Sciences, University of Hyderabad, Hyderabad (2003). doi:202.41.85.68
12. Rajan, K., Ramalingam, V., Ganesan, M., Palanive, S., Palaniappan, B.: Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Syst. Appl. **36**(8), 10914–10918 (2009)
13. Jayashree, R.: An analysis of sentence level text classification for the Kannada language. In: International Conference of Soft Computing and Pattern Recognition, pp. 147–151 (2011)
14. Gupta, V., Lehal, G.S.: Preprocessing phase of Punjabi language text summarization. In: International Conference on Information System for Indian languages, vol. 139, pp. 250–253 (2011)
15. Unicode Table. http://www.tamasoft.co.jp/en/general-info/unicode-decimal.html
16. Stemming. http://en.wikipedia.org/wiki/Stemming
17. Gupta, V.: Automatic stemming of words for Punjabi language. In: Advances in Signal Processing and Intelligent Recognition systems, Advances in Intelligent Systems and Computing, vol. 264, pp. 73–84 (2014)
18. Google Translation. https://translate.google.co.in/#auto/en/%E0%A8%AA%E0%A8%8F
19. Transliteration and Translation. http://www.shabdkosh.com/pa/
20. Bhatia, T.K.: Punjabi: a cognitive-descriptive grammar. Rout ledge Descriptive Grammar Series (1993)
21. Overview of Punjabi Grammar. http://punjabi.aglsoft.com/punjabi/learngrammar/?show=conjunction