Part-of-Speech Tagging System for Indian Social Media Text on Twitter

Anupam Jamatia
Department of Computer Science and Engineering
National Institute of Technology, Agartala, India
anupamjamatia@gmail.com

Amitava Das
Department of Computer Science and Engineering
University of North Texas, Denton, Texas, USA
amitava.das@unt.edu

Abstract

Automatic part-of-speech (POS henceforth) is the primary necessities for any kind of Natural Language Processing (NLP) applications like disambiguate homonyms, text-to-speech processing, information retrieval, natural language parsing, information extraction etc. Here in this paper we are concentrating on POS tagging systems for Hindi and Bengali tweets. Although automatic POS tagging is a well-defined research paradigm even there are significant efforts in literature for these two Indian languages. Making NLP methods for social media text (SMT) has recently received significant attention. Most of the research on SMT till date is concentrated on English therefore making technologies for other languages are as par necessity.

1. Introduction

Rapid growth in social media instigated enormous possibilities for information extraction research but those emergences would have to face several challenges due to the terse nature of the SMT. POS tagging is the prerequisite for any kind of NLP. So far, most of the research on social media texts has concentrated on English, but with best of our knowledge there is no work on Indian social media such as Hindi and Bengali tweets.

India is a nation of languages. It has close to 500 spoken languages (or over 1600, depending on what is counted as a language) and with some 30 languages having more than 1 million speakers. Hindi is the widely spoken language and 4th worldwide in terms of first language speaker whereas Bengali is the second highest one in India, national language in Bangladesh and 6th worldwide in terms of first language speaker.

SMT is characterized by having a high percentage of spelling errors and containing creative spellings (*gr8* for '*great'*), phonetic typing, word play (*goooood* for '*good'*), and abbreviations (*OMG* for '*Oh my God!*'). Non-English speakers do not always use Unicode to write social media text in their own language, frequently insert English elements (through code-mixing and Anglicism), and often mix multiple languages to express their thoughts. Even phonetic typing and creative Romanization are added challenges for Indian social media. Therefore making NLP techniques for Indian SMT is far more challenging than English. Indian SMT has several writing practices:

1. **Monolingual Unicode:** সাংস্কৃতিক সফরে সিঙ্গাপুর ও মালয়েশিয়ায় গেলেন সাংবাদিক সৌরভ
2. **Monolingual Phonetically typed:** sab jhakjhake chokchoke lokjon ..... :)
3. **Unicode-English Mix:** ধুরু! হাজি ক্যাম্পের ওয়াইফাইয়ে পাসওয়ার্ড ঠিক ঠাক দিলেও কানেক্ট নিতেছে না । >:(— feeling angry at Hajj
4. **Unicode-Phonetic-Roman Mix(Bengali):** জানি একদিন দূর থেকে দেখব সবারএই ভুলে যাওয়াও-- জানি একদিন চোখ থেকে পড়বে সুধু অশ্রু-রি ধারা . feelings_more gele sobar nam e las hoye jay
5. **Unicode-Phonetic-Roman Mix(Hindi):** मैं कलेज जा रहा हुँ (कलेज is an English word but, phonetically typed into Devnagri)
6. **Phonetic-Code Mixed:** dadaji budaphe m satiyha gye h ... only namo at any cost

Therefore it is eminent that developing a POS tagging system for all these above kinds demands a new research paradigm altogether whereas we

started with the simple one first: Type 1 monolingual.

We have noticed that monolingual Unicode tweets have relatively lower wordplay or spelling errors, therefore empirical question rises how different/difficult this task is than the general (like NEWS) text POS tagging. To answer this question our rationale is tweets are syntactically very different due to the 140-character length restriction. Moreover URL, hashtags, emoticons and unnecessary symbols made this text genre very different from formal text. Even to establish our rationale we have reported performances of general purpose POS system on our tweet data.

The rest of the paper organized as follows. Section 2 describes related work. Section 3 POS tagset for Indian SMT. As told earlier that tweets are altogether different from the formal text therefore a new tagset for tweets is required. Section 4 is corpus acquisition which elaborates tweet acquisition and annotation process. Annotation-crowd sourcing and bootstrapping methods are described in Section 5 and 6 respectively.. Experiments with various machine learning methods on our corpus are described in Section 7. Performances and learning curves are reported in the Section 8. Performance of general purpose POS tagger on our corpus reported in Section 9. Section 10 describes our pilot POS tagging for Code-Mixed Tweets. The paper concluded with future directions in the section 11.

2. Related Work

POS tagging for Indian language is a well-studied discipline. Here we discuss previous work on general purpose Hindi and Bengali POS tagging first and then will mention about few recent works on POS tagging of English tweets and other language like French SMT.

There are some significant work done on POS tagging on Indian languages like Yoonus and Sinha, (2011) where the authors build a hybrid system to tag for 12 Indian languages i.e. Assamese, Bengali, Bodo, Gujarati, Hindi, Malayalam, Manipuri, Nepali, Oriya, Punjabi, Tamil, and Urdu where it has been noticed that among 12 languages, Punjabi language achieved highest precision (88.97%) and recall (99.77%) and F-score (94.06%). For the experiment, corpora were taken from Linguistic Data Consortium for Indian Languages (LDC-IL) [1] . In Singh et. al.(2006), the authors found 93.45% accuracy of POS tagging on Hindi news corpora using morphological analysis backed by high coverage lexicon and a decision tree based learning algorithm. They made a conclusion that building POS tag for morphologically rich languages could be a better option. In (Ekbal and Bandyopadhyay, 2008), the authors proposed POS tagging system for Bengali news corpus using Support Vector Machine (SVM) which exceed the existing systems based on the Hidden Markov Model (HMM), Maximum Entropy (ME) and Conditional Random Field (CRF) with the final accuracy of 86.84%. Finally, authors concluded that the handling of unknown words using Bengali morphological analyzer might be an important factor to achieve more high accuracy. Mukherjee et.al. (2013) developed a Bengali POS tagging system using Global Linear Model (GLM) where the sentence structure features are defined by syntactical, morphological, ontological properties of Bengali. The system outperforms the existing models based on ME, SVM, CRF and HMM with the final accuracy 93.12%. Dandapat et. al.,(2004) proposed a POS tagger based on a combination of supervised and unsupervised learning with or without morphological analyzer restriction using HMM which achieved a final accuracy of 95%. In the proposed system they have used an untagged corpus and a morphological analyzer and further used those features for POS tagging. Authors concluded that the system's accuracy might increase by applying rule-based post-processing at least for typographical errors. In a later work Dandapat,(2007) used ME based statistical model not only Bengali but also for other Indian language like Hindi and Telegu where the POS tagger reached the overall accuracy on the development data of about 88%, 83% and 68% for Bengali, Hindi and Telugu respectively. In case of poor scenario for morphologically rich language like Bengali, Dandapat et.al. (2007) proposed a combination of HMM and ME based stochastic taggers where the best performance achieved for the supervised learning model along with suffix information and morphological restriction on the possible grammatical categories of a word. Dalal et.al., (2007) proposed maximum entropy Markov

[1]http://www.ldcil.org/standardsTextPOS.aspx

model based statistical POS tagger for a morphologically rich Indian national language Hindi with a rich set of features capturing the lexical and morphological characteristics of the language and achieved the best accuracy and average accuracy of 94.89% and 94.38% respectively using 4-fold cross validation.

There are very few works on POS tagging for tweets, possibly no works for Indian languages tweets. POS tagging for English tweets has first been attempted by Gimpel et.al.,(2011) where they have designed and developed POS inventory for Twitter specific but the accuracy level is obviously lower than the traditional genres. The system[2] is also available online for research purpose. The Gimpel et.al. (2011) tagger was CRF based where the basic features include, checking each word contains digit or hyphens, suffix features up to length 3 and capitalization word pattern. For improvisation authors added more features like regular expressions to detect at-mentions, hashtags, and URLs using frequently-capitalized tokens, traditional tag dictionary based on Penn Treebank (PTB), distributional similarity features for the limited data condition and at lastly phonetic normalization using the Metaphone algorithm[3]. Improved POS tagging for Twitter and Internet Relay Chat[4] (IRC) with unsupervised word clusters tempted by Owoputi et.al., (2013), where twitter tagging has improved 3% than the system developed by Gimpel et.al., (2011) by evaluating the use of large-scale unsupervised word clustering and lexical features. Authors also released[2] a new dataset of English tweets annotated using their own POS annotation guidelines. POS tagger software, annotation guidelines, and large-scale word clusters are available at.

Recently, Nooralahzadeh et. al.,(2014) has proposed a French POS tagging system using discriminative sequence labeling model: CRF, achived 91.9% accuracy on a target corpus collected from various types of French SMT user like Facebook, Twitter, Video games and medical web forums. They have proposed total 28 POS tags, were taken from French Treebank. The same system setup evaluated on a dataset containing 800 English tweets and English social media data such as NPS chat with PTB tags achieved reported accuracies 90.1% and 92.7% respectively.

Vyas et. al. (2014) proposed a POS tagger for Hindi-English code-mixed text collated from Facebook forums, and explored language identification, back transliteration, normalization and POS tagging for code-mixed data. Even multilingual and cross-lingual POS tagging have been explored by several researchers (Yarowsky and Ngai, 2001; Xi and Hwa, 2005; Snyder et.al.,2008; Naseem et.al., 2009).

3. POS Tagset for Indian SMT

Due to the conversational nature of twitter people frequently mix up several non-textual elements in their tweets such as hashtags, emoticons, and URL. Another prime reason of such inclusion is the need of more information propagation; Twitter has 140 characters length restriction. Therefore POS tagging for twitter demands a new tagset designing.

The very first POS tagger for English tweets has been designed by Gimpel et.al., (2011). They introduced several new POS categories. We borrowed several categories from their definition and added them with the Indian languages standard POS tagset as standardized by LDC-IL[1].

Noun (Common & Proper)		**Example**
N_NN	Common Noun	ज़मीन, राजनीति
N_NNV	Verbal Noun	खाने, जाने
N_NST	Spatio-temporal	ऊपर , निचे , आगे
N_NNP	Proper Noun	भारत, अमरीका, ग्लासगो
Pronoun		
PR_PRP	Personal	मैं, तुम
PR_PRL	Relative	जो, जिस, जब
PR_PRF	Reflexive	अपने, स्वयं, खुद
PR_PRC	Reciprocal	अपने आप
PR_PRQ	Wh-Word	किसका,किसकी
Verb		
V_VM	Main	हुई,हटाया
V_VAUX	Auxiliary	रहे, हैं
Adjective		
JJ	Adjective	दुर्घटनाग्रस्त, भीषण
Adverb		
RB_ALC	Adverb of Locations	अबकी, जहां
RB_AMN	Adverb of Manner	आख़िर, जैसे
Demonstratives		
DM_DMD	Absolute	वहां, यहाँ

[2]http://www.ark.cs.cmu.edu/TweetNLP/
[3]http://commons.apache.org/codec/
[4]http://www.irc.org/

DM_DMI	Indefinite	कोई, किस
DM_DMQ	Wh-word	कौन
DM_DMR	Relative	जिस, जो
Quantifier		
QT_QTF	General	थोड़ा, बहुत, कुछ
QT_QTC	Cardinals	एक, दो , तीन
QT_QTO	Ordinals	पहला, दूसरा, तीसरा
Residuel		
RD_ECH	Echowords	खाना-वाना, पानी-बानी
RD_PUNC	Punctuations	, ; .
RD_RDF	Foreign Words	A word written in script other than the script of the original text
RD_SYM	Symbol	$ * ( ) { }
RD_UNK	Unknown	Unknown Words
Conjunction & Postposition		
CC	Conjunction	और , अगर, क्योँकि
PSP	Postposition	ने , को , से , में
Particle & Numerals		
RP_RPD	Default	तो , भी
RP_NEG	Negation	नहीं, बिना
RP_INTF	Intensifier	बहुत, बेहद
RP_INJ	Interjection	अरे, हे, ओ
Twitter-specific		
$	Numerals	1,2,3
@	At-mention	@
~	Re-Tweet/discourse	RT, ~
E	Emoticon	:) :D ☹ ☺
U	url or email	www
#	Hashtag	#

Table 1: POS Tagset for Indian SMT

Finally we concluded with 38 fine-grain tags for Hindi tweets as reported in the Table-1 and 12 coarse-grain tags are reported in the Table- 2. So we have total 38 fine-grain tagset for Hindi Twitter which is shown in Table-1. For our data set we have concise the 38 fine-grain tagset to 12 course-grain tag-set shown in Table- 2.

Noun	N
Pronoun	PR
Adjective	JJ
Verb	V
Adverb	RB
Conjunction	CC
Demonstrative	DM
Particle	RP
Quantifier	QT
Residual	RD
Twitter	TWT
Numerals	$

Table 2: Coarse-grain POS Tagset

4. Corpus Acquisition

We choose NEWS tweets from @BBCHindi and @aajtak. The standard Java based Twitter API[5] has been used for the purpose. For preprocessing CMU tokenizer, a sub-module of CMU twitter POS tagger has been used. Although this tokenizer has been developed for English tweets but still works well for other languages as well. Finally we collected total 3488 tweets from @BBCHindi (995 tweets) and @aajtak (2493 tweets) timeline.

5. Annotation – Crowd Sourcing

Inspired by several success stories of crowd-sourcing we decided to go for crowd-source the annotation task. Most popular and fastest crowd-source service provides by Amazon Mechanical Turk (AMT)[6]. But quality control is the main challenge with this kind of service. Initially we took 50 tweets to POS annotate by crowd. [7][8]In AMT, six workers have participated with an effective hourly incentive $4.500. Total token was 1398 from 50 tweets, incentive per submission was $0.020. So it took total $40.542. Out of these six workers, two workers (i.e. Worker-1 and Worker-2 respectively) were relatively effective to accurate. But the overall annotation experience was not very satisfactory. The main problem is less Hindi speaker turkers for such complex annotation process. Even while analyzing the results we found only two workers was relatively better annotator, resulting 57.142% and 42.857% accuracy compared to a manually annotated golden set. It might be suggested that we should have invested more time with different experiment with AMT such as with increasing the compensation. But looking at the basic result we decided to move on. Even we should mention that the overall annotation process took 2 weeks to complete. So, finally we decided to go for bootstrapping as discussed in the next section.

[5]http://twitter4j.org
[6]https://www.mturk.com/mturk/welcome

6. Bootstrapping

As we failed to get quality data from the AMT we decided to choose the bootstrapping on total 1300 tweets. In the bootstrapping process we annotate 100 tweets in each iteration and trained a CRF based classifier and automatically tag next 100 tweets for next iteration. After automatic tagging using the CRF classifier has been checked manually. This process iterates until learning curves become straight. As it is an ongoing task, for the time being we are able to manage to get 1300 annotated tweets.

For the CRF training we used very basic features such as first 4 chars of any word; if any word is less than 4 chars, then use the whole one; last 4 chars of any word; if any word is less than 4 chars, then use the whole one; previous 3 words and their tags; next 3 words; and Current word (Sarkar, S., and Bandyopadhyay, S., 2008).The bootstrapping setup has described below:

1. Took 100 tweets and annotate manually.
2. Train CRF classifier.
3. Check the 5-fold cross-validation
4. POS Tag new 100 unlabeled tweets using the CRF trained classifier.
5. Automatically tagged 100 tweets then checked and corrected manually. We have developed a GUI based annotation tool.
6. After correction this set added with the previous training set.
7. Retrain the CRF classifier using the new training set. Re-cross validate and made sure performance increased than the previous iteration.
8. Continue step-2 to step-7 until learning curves become flat.

7. Experiments with Various ML Methods

With the annotated 1300 tweets we experimented with several Machine Learning (ML) methods such as SVM, Naive Bayes (NB), and Random Forest (RF) using weka[9] toolkit. To get the results using weka we have developed a system to convert the CRF formatted file to weka compatible ARFF formatted input file. Reported accuracies in the Table 3 are based on 5-fold cross validations using the features of back and forth first, uni, bi, tri, tetra grams and previous word. Among all ML methods Random Forest stand out as the highest performing one. The same feature set, as discussed in the previous section has been used.

Fine-grain tag-set			
ML Method	Correctly Classified Instances	Total Number of Instances	F-Measure
Naive Bayes	6058	18123	33.43
SMO	6879		37.96
Random Forest	14876		82.01
Coarse-grain tag-set			
Naive Bayes	5933	18123	32.74
SMO	6887		38.00
Random Forest	15055		83.01

Table 3: Coarse-grain POS Tagset

8. Learning Graphs

It has been observed (reported in the Figure-1) that after the 13th iteration in the bootstrapping setup the accuracy (78.148%) goes down than the previous i.e. 12th iteration, when it has was 78.418%. A similar accuracy curve on coarse-grain tagset could be noticed in the Figure 2. Accuracies on each bootstrapping iteration are reported in the Table 4.

As reported in the previous section that Random Forest based tagger was the highest performing, therefore that is our final system. We have tested learning graphs for our open POS classes: Noun, Verb, Adverb and Adjective, reported in the Table-4 and Figure-4. From the learning graph it is clear that the system is unable to handle Adjectives and Adverbs very well. The possible reason might be as because we have used basic feature selection only.

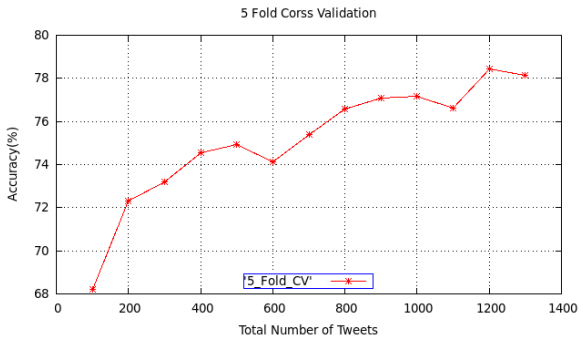

Figure-1: Accuracy Vs. No. of Tweets for 5-Fold Cross Validation (Fine-grain Tagset)

[9]http://www.cs.waikato.ac.nz/ml/weka/

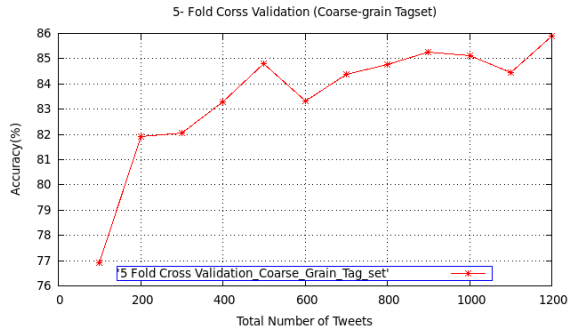

Figure-2: Accuracy Vs. No. of Tweets for 5-Fold Cross Validation (Coarse-grain Tagset)

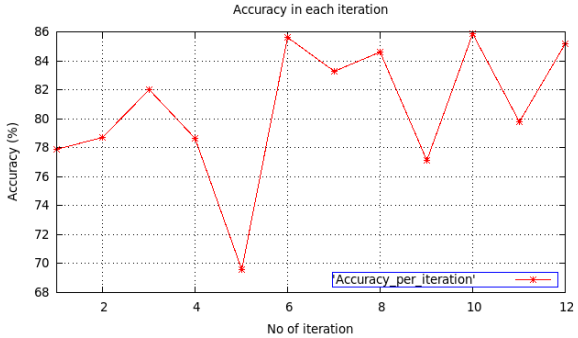

Figure-3: Accuracy Vs. No. of iteration

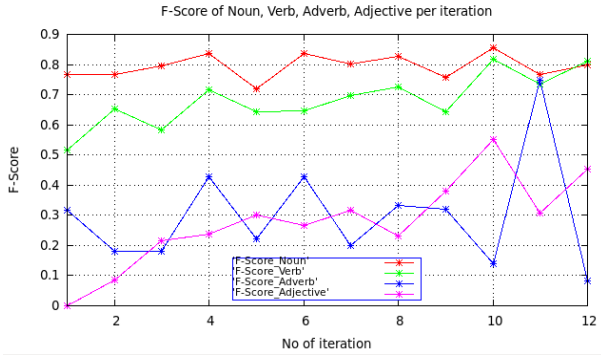

Figure-4: F-Score of Noun, Verb, Adverb, Adjective per iteration

Iteration No.	F-Scores				Accuracies (%)
	Noun	Verb	Adverb	Adjective	
1	0.7662	0.5166	0.3158	0.0	77.87
2	0.7655	0.654	0.1818	0.0857	78.73
3	0.7961	0.5821	0.1818	0.2169	82
4	0.8354	0.717	0.4286	0.2368	78.63
5	0.7204	0.6447	0.2222	0.3022	69.56
6	0.8371	0.6462	0.4286	0.2667	85.59
7	0.8023	0.6957	0.2	0.3168	83.29
8	0.8265	0.7246	0.3333	0.2319	84.58
9	0.7559	0.6426	0.32	0.3788	77.11
10	0.8571	0.8175	0.1404	0.5522	85.89
11	0.7677	0.7343	0.7471	0.3077	79.78
12	0.7978	0.8114	0.0833	0.4545	85.2

Table-4: Accuracies and F-Score of Noun, Verb, Adverb, Adjective per iteration

Some of the examples of erroneous tags are shown in the following example. Red markings are erroneous by the system whereas the first tag is from the golden set.

@BBCHind/@ -/RD_SYM मेरा/PR_PRP/N शो/RD/N तो/CC/QT फ़िल्मों/N से/PSP बड़ा/JJ/N है/V/N :RD_SYM कपिल/N शर्माN http://t.co/XUi23BVktT/U

It has been observed from the confusion matrix of the Random Forest method that there is a high confusion among adjective and common nouns, as 28.65% and 4.86% are wrongly detected as common nouns and adjective respectively whereas 50.41% and 80.69% are correctly detected as adjective and common noun respectively. In case of Auxiliary Verb and Main Verb, 70.60% and 57.47% are rightly detected as auxiliary verb and main verb whereas 15.45% and 11.61% are wrongly detected as auxiliary verb and main verb respectively. There were some prominent error in case of different categories of Proper Noun and Common Noun, 76.42% and 58.39% are correctly detected as proper noun (location) and proper noun (person) but 9.03% and 21.25% are wrongly detected as a common noun in case of for proper noun (person) and proper noun (location) respectively.

9. Hindi Tweet POS Tagging using General Purpose POS Tagger

We have tested performance of a publicly available POS tagger, developed by Society for Natural Language Technology Research [SNLTR][10] on our data. Performance of the SNLTR tagger on our data is only 47.49% but this accuracy with their trained model. An appropriate tag set conversion module has been written to convert the SNLTR tagset to our proposed tagset. When we trained the SNLTR system with 1200 tweets and tested on 100 tweets we have found the accuracy 86.99%. Indeed, this experiment proves that the general purpose Hindi POS Tagger performs poorly on Hindi tweets.

[10] http://nltr.org/snltr-software/

10. POS Tagging for Code-Mixed Tweets

At this point we tested similar setup for code-mixed tweets POS tagging. This data collected from Facebook as well as Twitter. We have annotated manually 400 utterances. In our code-mix corpus, we have around 67.92% Hindi words (Hi) and 32.08% English (En) words. Our language marking system follows the standard defined by Burman et. al., (2014). The same POS tagset has been used here irrespective of word language. As suggested in the Vyas, Y. et.al (2014), we did SNLTR POS tagger on the Hindi part of the data and run CMU POS tagger for the English part respectively. Word sequence plays a great role for syntactic formation and especially for POS tagging. We are not claiming that breaking language-specific sequences and using language specific tagger is a right approach but we have done experiment based on Vyas, Y. et.al (2014) and reporting the accuracy in this paper so that future research could be benefited. With this method we only achieved 2.1% accuracy on the Hindi data and 50.15% accuracy on English part of the data. Obviously the reason is terse nature of the data. SNLTR is a general purpose POS tagger, so naturally it gives lower accuracy than the CMU POS tagger, which is specifically designed for the tweets.

We have also experimented with three different ML methods i.e. SMO, NB and RF on total code-mixed dataset and observed that Random Forest give the highest correctly classified instances: 63.65% with weighted average on F-Measure 0.632, compared to 26.82% with weighted average on F-Measure 0.17 of SMO and 22.51% with weighted average on F-Measure 0.165 of NB which is shown in Table-5.

This is just a pilot POS tagging task on barely 400 tweets for the Code-Mixed social media text. Result implies there is a need to run a separate set of experiments on this data genre. This is the motivation of our future work, ongoing.

11. Conclusions and Future Directions

In this paper we have reported our initial experiments on Hindi tweet POS tagging. Indeed, reported accuracies are far from to be useful. So far we have used linear kernel with default parameters and now investigating with optimized parameters. This is an ongoing task.

As mentioned in the introduction section that there are several writing practices in Indian SMT therefore our next endeavor will be to develop POS tagger for code-mixed SMT.

Different Systems	ML Methods with 5-Fold Cross Validation					
	SMO		Naïve Bayes		Random Forest	
	CCI (%)*	F-Measure	CCI (%)	F-Measure	CCI (%)	F-Measure
EN-Data_Tagged_Proposed_System	44.98	0.322	42.07	0.309	63.94	0.628
EN-Data_Tagged_CMU_System	46.59	0.338	37.40	0.312	68.46	0.67
EN+HI-Data_Tagged_Proposed_System	26.82	0.17	22.51	0.165	63.65	0.632
HI-Data_Tagged_Proposed_System	23.00	0.15	21.32	0.146	58.54	0.582
HI-Data_Tagged_SNLTR_System	29.33	0.184	27.01	0.23	67.58	0.672

*Correctly Classified Instances (CCI)

Table:5 ML Methods with 5-Fold Cross Validation on Code-Mixed Data

References

Barman, U., Das, A., Wagner, J., and Foster, J. 2014 *Code-Mixing: A Challenge for Language Identification in the Language of Social Media.* The 1st Workshop on Computational Approaches to Code Switching, EMNLP 2014, October, 2014, Doha, Qatar.

Snyder, B., Naseem, T., Eisenstein, J. and Barzilay, R. 2008 *Unsupervised Multilingual Learning for POS Tagging.* In the proceedings of Conference on Empirical Methods in Natural Language Processing, ACL, Pages 1041–1050. Honolulu, Hawaii, USA.

Naseem, T., Snyder, B., Eisenstein, J., and Barzilay, R. 2009. *Multilingual Part-of-speech Tagging: Two Unsupervised Approaches.* In the Journal of Artificial Intelligence Research, Volume 36 Issue 1, ACM. Pages: 341–385, USA.

Yarowsky, D., and Ngai, G. 2001. *Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora* in the Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL), Pages 1-8, PA, USA

Xi, C., and Hwa, R. 2005. *A Backoff Model for*

Bootstrapping Resources for Non English Languages in the proceedings of the conference Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), ACL, Pages 851-858, Stroudsburg, PA, USA.

Das, D., and Petrov, S., 2011. *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections* in the proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Pages 600–609, Portland, Oregon, USA.

Gimpel, K., Schneider, N., O'Connor, B. Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments, in the proceedings of Association for Computational Linguistics, Pages 42--47, Portland, Oregon, USA.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N. and Smith, N.A. 2013 *Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters.* The Association for Computational Linguistics in the proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Pages 380–390 Atlanta, Georgia, USA.

Foster, J., Cetinoglu, ¨O., Wagner, J., Roux, J.L., Hogan, S., Nivre, J., Hogan, D. and Genabith, J. 2011 *#hardtoparse: POS Tagging and Parsing the Twittervers.* In the Proceedings of AAAI-11, Workshop on Analysing Microtext, San Francisco, CA. USA.

Yoonus, M. M., and Sinha, S., 2011. *A Hybrid POS Tagger for Indian Languages.* In Language in India, Pages 317 –330.

Singh, S., Gupta, K., Shrivastava,M. and Bhattacharyya, P. 2006 *Morphological richness offsets resource demand – experiences in constructing a POS tagger for Hindi*, In the Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL),Pages: 779-786, Sydney, Australia.

Ekbal, A. and Bandyopadhyay, S. 2012 *Part of Speech Tagging in Bengali using Support Vector Machine.* , In the Proceedings of the International Conference on Information Technology, IEEE CS Press, Pages 106–111. Bhubaneswar, Orissa, India.

Mukherjee, S., and Das Mandal, S.K. 2013. *Bengali Parts-of-Speech Tagging using Global Linear Model* in India Conference (INDICON-2013), IEEE, Pages 1 – 4, Mumbai, India.

Baskaran, S. 2006. *Hindi POS Tagging and Chunking.* In Proceeding of the NLPAI Machine Learning Competition, Hyderabad, India.

Dandapat, S., Sarkar, S., Basu, A. 2004. *A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali* In International conference on Computational Intelligence and Transactions on Engineering, Computing and Technology. Pages 169–172.

Dandapat, S., Sarkar, S. 2006. *Part of speech tagging for Bengali with hidden markov model* in proceeding of the NLPAI Machine Learning Competition. Hyderabad, India.

Dandapat, S. 2007. *Part of speech tagging and chunking with maximum entropy mode*l .In proceedings of the IJCAI Workshop on Shallow Parsing for South Asian Languages, Pages 29–32. Hyderabad, India.

Dandapat, S., Sarkar, S., Basu, A. 2007. *Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario*. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Pages 221–224

Dalal, A., Kumar N., Sawant, U., Shelke, S. and Bhattacharyya, P., 2007. *Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi.* In proceedings of ICON 2007 IIIT, Hyderabad.

Nooralahzadeh, F., Brun, C and Roux, C., 2014, *Part of Speech Tagging for French Social Media Data*. In proceedings of the 25th International Conference on Computational Linguistics (COLING-2014). Pages: 1764--1772. Dublin, Ireland.

Sarkar, S., and Bandyopadhyay, S., 2008 *Design of a Rule-based Stemmer for Natural Language Text in Bengali*, In the proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 65–72, Hyderabad, India.

Vyas, Y., Gella, S., Sharma, J., Bali, K.,Choudhury, M. 2014. *POS Tagging of English-Hindi Code-Mixed Social Media Content.* In proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, Pages 974–979, Doha, Qatar.